SIAM J. CONTROL OPTIM.
Vol. 40, No. 1, pp. 1–38

SOLUTION SENSITIVITY FROM GENERAL PRINCIPLES*

ADAM B. LEVY†

Abstract. We present a generic approach for the sensitivity analysis of solutions to parameterized finite-dimensional optimization problems. We study differentiability and continuity properties of quasi-solutions (stationary points or stationary point-multiplier pairs), as well as their existence and uniqueness, and the issue of when quasi-solutions are actually optimal solutions. Our approach is founded on a few general rules that can all be viewed as generalizations of the classical inverse mapping theorem, and sensitivity analyses of particular optimization models can be made by computing certain generalized derivatives in order to translate the general rules into the terminology of the particular model. The useful application of this approach hinges on an inverse mapping theorem that allows us to compute derivatives of solution mappings without computing solutions, which is crucial since numerical solutions to sensitive problems are fundamentally unreliable. We illustrate how this process works for parameterized nonlinear programs, but the generality of the rules on which our approach is based means that a similar sensitivity analysis is possible for practically any finite-dimensional optimization problem. Our approach is distinguished not only by its broad applicability but by its separate treatment of different issues that are frequently treated in tandem. In particular, meaningful generalized derivatives can be computed and continuity properties can be established even in cases of multiple or no quasi-solutions (or optimal solutions) for some parameters. This approach has not only produced unprecedented and computable conditions for traditional properties in well-studied situations, but has also characterized interesting new properties that might otherwise have remained unexplored.

Key words. sensitivity analysis, parameterized optimization, outer graphical derivative, calmness, protodifferentiability

AMS subject classification. 90C31

PII. S036301299935211X

1. Introduction.

Even the most basic parameterized optimization problems can exhibit poor solution behavior; there can be more than one solution for some parameters, or no solutions for some parameters, or the solutions can be sensitive to perturbations of the parameter. These same difficulties persist for quasi-solutions (stationary points or stationary point-multiplier pairs) both in the case when quasi-solutions and optimal solutions coincide, and in the case when they do not. To illustrate this, we consider the minimization over $x \in \mathbb{R}$ of three different parameterized objective functions: (A) $x^2/2 - wx$, (B) $x^3/3 - wx$, and (C) $x^4/4 - wx$. The stationary point mappings associated with each of these three problems are plotted against the parameter $w \in \mathbb{R}$ in Figure 1.1. The mapping in (A) behaves about as well as can be hoped; there exists a unique stationary point for each parameter, and the dependence of the stationary points on the parameters is linear. On the other hand, the mapping in (B) is very poorly behaved; there are either two stationary points or none for all parameters but $w = 0$, and moreover the vertical slope at zero indicates that very small perturbations from $w = 0$ cause dramatic changes in the stationary points. Finally, the mapping in (C) has mixed behavior; there exist unique stationary points for all parameters, however, there is again the sensitivity indicated by the vertical slope at zero. Notice also that the optimal solutions coincide with the stationary points in both cases (A) and (C), but not in (B), where the optimal solution mapping consists

*Received by the editors February 16, 1999; accepted for publication (in revised form) November 3, 2000; published electronically May 3, 2001.
http://www.siam.org/journals/sicon/40-1/35211.html
†Department of Mathematics, Bowdoin College, Brunswick, ME 04011 (alevy@bowdoin.edu).

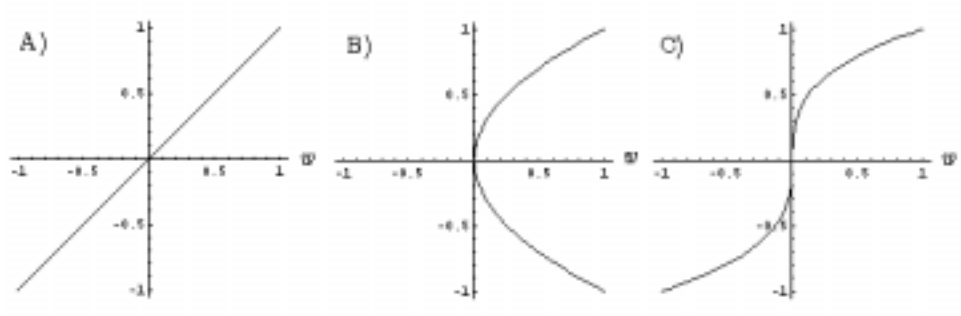


FIG. 1.1. *Simple stationary point mappings.*

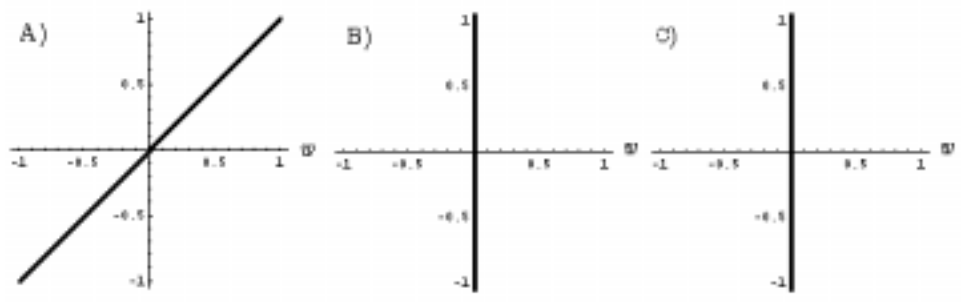


FIG. 1.2. *Derivatives of stationary point mappings.*

of only the upper branch of the square-root.

Simple examples like the three above are unusual among optimization problems generally, since the optimal solutions above can be calculated explicitly. It is more typical that numerical methods are used to identify solutions, and this fact provides motivation for the kind of sensitivity analysis that we present in this paper. In sensitive problems, small perturbations of parameters produce relatively large changes in solutions, so numerically computed solutions are fundamentally unreliable. It follows that any broadly useful sensitivity analysis of solutions should rest on tests that are independent of the computation of solutions. For the approach outlined in this paper, this paradox is handled by a kind of inverse mapping theorem that relates derivatives of the quasi-solution mapping to derivatives of the original data. For the simple examples above, the inverse mapping theorem gives a generalized derivative of the stationary point mapping at $w = 0$, as the inverse mapping associated with the second derivative in x of the objective function at $(x, w) = (0, 0)$. The second derivatives above are easy to compute and the inverse mapping theorem in these cases gives the generalized derivatives of the stationary point mappings displayed in Figure 1.2. In all of these cases, the graph of the generalized derivative is the tangent line to the stationary point mapping at $w = 0$, and any other reasonable generalized first derivative would be expected to produce the same thing for these examples.

The Fermat rule of differential calculus (the gradient is zero at a stationary point) is a very familiar example of a general rule which can be applied in different particular situations to obtain results about optimality, and it served to unify centuries of work on particular optimization problems. This kind of result is extraordinary in optimiza-

tion, the more typical generalization being an extension from some well-understood model. Our approach in this paper relies on general principles from variational analysis that are very much in the spirit of the Fermat rule. In fact, all of the general principles we use are different kinds of generalizations of the classical inverse mapping theorem.

CLASSICAL INVERSE MAPPING THEOREM. *For a single-valued $\mathcal{C}^1$ mapping $x : \mathbb{R}^d \to \mathbb{R}^n$ and any points $\bar{w} \in \mathbb{R}^d$ and $\bar{x} := x(\bar{w})$ the following are equivalent:*

(i) *$w' = 0$ is the only vector in $\mathbb{R}^d$ for which the Jacobian image $\nabla x(\bar{w}) \cdot w'$ equals zero.*

(ii) *There exists a neighborhood W of $\bar{w} \in \mathbb{R}^d$ and a neighborhood X of $\bar{x} \in \mathbb{R}^n$ such that the restriction of x to W is a bijection onto X with $\mathcal{C}^1$ inverse $w : X \to \mathbb{R}^d$, and moreover the Jacobians satisfy*

$$\nabla w(\bar{x}) = \big(\nabla x(\bar{w})\big)^{-1}. \tag{1.1}$$

We already alluded to one version of a generalization of this result, where an analogue to the identity (1.1) is obtained in a much less restrictive setting. The other generalizations that we use in this paper all characterize weaker properties than (ii) using generalized versions of (i).

We focus first on the most abstract and general objects, so that the specific applications become a second-stage matter of translating the abstract results into the language of the particular optimization problem at hand. We sometimes need to impose additional structure when we translate our results into verifiable conditions involving familiar terms of particular models, but these compromises are made after the fundamental rules have already been established, so their consequences and degree of necessity can be clarified. In this paper, we do not attempt to minimize the compromises we make to translate the general results, but instead we adopt reasonably standard assumptions so that the reader can see how the entire process works in a familiar setting. Of course, our approach is amenable to any desired refinement of the assumptions (e.g., reduced differentiability conditions on the original data) since it is based on completely general characterizations. Moreover, even under the standard assumptions, some of the sufficient conditions generated by our approach are unprecedented (even in cases of well-studied models like nonlinear programs), and our approach has identified interesting new properties that might otherwise have remained unexplored. We do not attempt to give a complete survey of the general principles underlying the sensitivity analysis of solutions, but rather we focus on a few principles and sensitivity properties that provide a basis for a reasonably complete sensitivity analysis.

One important aspect of our approach is that the general principles on which it is based are *characterizations* of certain desirable sensitivity properties. It follows that anyone who seeks derivative conditions for these sensitivity properties in connection with a particular optimization model must essentially reprove the general principles in their special case. Thus, anyone desiring to carry out such a sensitivity analysis need not create a new strategy from the ground up, and would be well served to translate the general principles for their particular case. In fact, one of the goals of this paper is to encourage others to work with the general principles presented here in order to carry out sensitivity analyses in more cases of interest.

To develop our general results, we study the variational properties of multifunctions (set-valued mappings) $S : \mathbb{R}^d \rightrightarrows \mathbb{R}^n$, and then specialize these results to the cases

where S represents stationary points or stationary point-multiplier pairs associated with the parameterized optimization problem

$$\min\{f(x,w)\} \text{ over all } x \in \mathbb{R}^n, \tag{1.2}$$

where f is a proper (so it is not identically equal to ∞ and nowhere equal to $-\infty$) extended real-valued function of $x \in \mathbb{R}^n$ and the parameter $w \in \mathbb{R}^d$. This is a completely generic finite-dimensional model which even includes constrained optimization problems since constraints can be incorporated through penalties into the objective function f. For instance, nonlinear programs like

$$\min\{g_0(x,w)\} \text{ over all } x \in C(w) := \left\{ x \;\middle|\; \begin{array}{ll} g_i(x,w) \le 0 & \text{for } i = 1,\dots,s, \\ g_i(x,w) = 0 & \text{for } i = s+1,\dots,m \end{array} \right\} \tag{1.3}$$

fit this model when the objective f is defined by

$$f(x,w) = \begin{cases} g_0(x,w) & \text{if } x \in C(w), \\ \infty & \text{otherwise.} \end{cases} \tag{1.4}$$

Throughout the paper, we illustrate how our approach applies to the analysis of the parametric sensitivity of general nonlinear programs.

As we already saw in the three simple examples at the beginning of this section, different issues are involved in solution sensitivity, including existence and uniqueness of solutions, as well as their continuity and differentiability properties. Another important issue is whether or not all stationary points are optimal solutions (every optimal solution is of course a stationary point, since the latter are defined by necessary conditions for optimality). Much of the sensitivity analysis carried out so far has followed the traditional hierarchy that has existence and uniqueness as the initial properties to be established (or assumed), followed by continuity properties, which in turn are followed by differentiability properties. In our approach, this hierarchy is inverted since these various properties depend on one another in new ways. For instance, the differentiability property that we use provides important sensitivity information even without uniqueness or existence for some parameters. Interestingly, existence and uniqueness properties cannot be characterized via first derivative objects in the absence of stability; this phenomenon is illustrated by the fact that the derivatives of the stationary point mappings above in cases (B) and (C) are the same even though the existence/uniqueness properties of the corresponding stationary point mappings are different. The unorthodox organization of the paper reflects the new hierarchy demanded by this situation; we begin with a discussion of differentiability properties, then move to continuity properties, and finally to questions of existence and uniqueness.

Sensitivity analysis has a long history, and many important contributions have been made over the years, both via the approach espoused here as well as by other approaches. The references at the end of this paper give some of the names of the major contributors; however, the purpose of this paper is not to provide a complete survey of the results in this area. Nor is its primary purpose to add particular results to the vast accumulation, but instead to frame certain kinds of sensitivity analyses generally in terms of a few basic principles that have been identified through this approach. One point of this paper then is that all analyses of certain sensitivity properties have at their core some general principle of variation. This means that any past or future attempt to characterize certain continuity or differentiability properties

of solutions in terms of derivatives of the original data of the problem must essentially involve the same general principles. It follows that translating these general principles to obtain particular results of interest is perhaps the most direct path to these results; and in any case it is an approach to results for new problems that has a head start, since it rests on already established general principles of variation.

Nonlinear programs have been studied so extensively that many of their important sensitivity properties have been well-understood long before this paper. In light of this extensive history as well as the fact that nonlinear programs are not the main focus of this paper, it is not feasible to give here an exhaustive description of past work in this area. When results presented here are apparently genuinely new, we do give some indication of that, but the default understanding in the applications sections is that similar results likely have been obtained before by other means. Nor are most aspects of the basic approach presented here new to this paper, and we try to indicate particular contributions along these lines where appropriate. Of particular note, the continuity and differentiability properties used here have been studied previously, and we refer the interested reader to [52] and [3] for a more complete treatment of these and related properties and their calculus. We also try to give some credit where it is due in the concluding section of the paper, where we describe some related solution sensitivity results obtained through the same basic approach, and compare our results to some obtained by fundamentally different approaches to solution sensitivity analysis.

2. Differentiability. One differentiability property frequently sought in connection with the sensitivity analysis of solutions is called B-differentiability: A single-valued mapping $f : \mathbb{R}^d \to \mathbb{R}^n$ is B-differentiable at $\bar{x} \in \mathbb{R}^d$ if it is directionally differentiable at $\bar{x}$ and the mapping $x \mapsto f(\bar{x}) + Df(\bar{x})(x - \bar{x})$ is a first-order approximation of f near $\bar{x}$. The differentiability property that we will use can be viewed as a direct generalization to multifunctions of B-differentiability for locally Lipschitz single-valued mappings, and it provides important sensitivity information even in the case of a true multifunction when the B-differentiability is not appropriate. This is important because there are many situations where there exist multiple or no solutions to a parameterized optimization problem. In addition, the property we use gives meaningful sensitivity information for single-valued mappings that are not B-differentiable.

We wish to study the differential properties of solutions associated with optimization problems (1.2), without making a priori assumptions about existence or uniqueness. This leads us naturally to generalized derivatives of multifunctions $S : \mathbb{R}^d \rightrightarrows \mathbb{R}^n$, since, for example, we can define a multifunction by

$$S(w) := \{x | x \text{ gives an optimal solution to (1.2) with parameter } w\}.$$

The primary generalized derivative that we use here is the *outer graphical derivative of S at $\bar{w}$ for $\bar{x}$* denoted $DS(\bar{w}|\bar{x}) : \mathbb{R}^d \rightrightarrows \mathbb{R}^n$ and defined as follows:

$$DS(\bar{w}|\bar{x})(w') = \left\{ x' \;\middle|\; \begin{array}{l} \exists w'_\nu \to w', \tau_\nu \downarrow 0 \text{ with } (\tilde{x}_\nu - \bar{x})/\tau_\nu \to x' \\ \text{for some } \tilde{x}_\nu \in S(\bar{w} + \tau_\nu w'_\nu) \end{array} \right\}.$$

The outer graphical derivative gets its name from the fact that it is the outer graphical limit as $t \downarrow 0$ of the sequence of difference quotient multifunctions defined for $t > 0$ by

$$w' \mapsto \frac{S(\bar{w} + tw') - \bar{x}}{t}.$$

The *outer graphical limit* of a sequence of multifunctions $\{G_t\}$ is the multifunction whose graph equals the set of all points obtained as limits of points in the sets $\operatorname{gph} G_{t_n}$

for some sequence $t_n \downarrow 0$. The outer graphical derivative always exists, though it may have empty values for some points.

A property called *protodifferentiability at $\bar{w}$ for $\bar{x}$* occurs when every element (w', x') in the graph of the outer graphical derivative $DS(\bar{x}, \bar{w})$ can actually be obtained as a limit

$$(w', x') = \lim_{t \downarrow 0} \frac{(w(t), x(t)) - (\bar{w}, \bar{x})}{t} \tag{2.1}$$

for some selection mapping $t \mapsto \big(w(t), x(t)\big) : [0, \epsilon] \to \operatorname{gph} S$ with $\epsilon > 0$. The concept of protodifferentiability is a generalization of B-differentiability [49] for single-valued mappings: A (single-valued) continuous mapping $S : \mathbb{R}^d \to \mathbb{R}^n$ is *B-differentiable at* $\bar{w}$ if the difference quotient mappings

$$w \mapsto \frac{S(\bar{w} + tw) - S(\bar{w})}{t} \quad \text{for} \quad t > 0$$

converge pointwise as $t \downarrow 0$ to a continuous (B-derivative) mapping $H : \mathbb{R}^d \to \mathbb{R}^n$ and do so uniformly on bounded sets.

Proposition 2.0.1 (see Proposition 2.2 of [27] and Proposition 3.5 of [29]). *Let O be an open neighborhood of a point $\bar{w} \in \mathbb{R}^d$ and consider a continuous single-valued mapping $S : O \to \mathbb{R}^n$. Then S is B-differentiable at $\bar{w}$ for $\bar{x} = S(\bar{w})$ if and only if S is protodifferentiable at $\bar{w}$ with $DS(\bar{w}|\bar{x})$ single-valued.*

When S happens to be Lipschitz continuous around $\bar{w}$, then S is B-differentiable at $\bar{w}$ if and only if S is protodifferentiable at $\bar{w}$ (and the single-valuedness of $DS(\bar{w}|\bar{x})$ is automatic under either of these equivalent conditions).

In either of these situations, the outer graphical derivative $DS(\bar{w}|\bar{x})$ is the same as the B-derivative, and one has the local expansion

$$S(\bar{w} + tw) = S(\bar{w}) + tDS(\bar{w}|\bar{x})(w) + o(t|w|) \quad \textit{for} \quad t > 0.$$

Remark. Since local Lipschitz continuity of solution mappings is traditionally established or assumed before studying differentiability properties, our approach using the outer graphical derivative and the concept of protodifferentiability covers at least as much ground as the traditional approaches to B-differentiability. However, the outer graphical derivative provides important sensitivity information even when S is genuinely set-valued, so our approach covers new ground too. From an analytical point of view, the image of the outer graphical derivative $DS(\bar{w}|\bar{x})(w')$ contains all the cluster points of sequences considered in the (uniform) limit defining the B-derivative. The situation when a single-valued mapping S fails to be B-differentiable at $\bar{w}$ thus corresponds to the case of the image set $DS(\bar{w}|\bar{x})(w')$ containing either multiple or no points. So the outer graphical derivative captures exactly the kind of derivative information that is usually desired, but which is often unnecessarily lost when attention is restricted to more traditional notions of differentiability. Protodifferentiability also signals an interesting geometric property where the graph of the multifunction is the image under an invertible nonlinear transformation of the graph of a B-differentiable single-valued mapping (see [29] for more details on this).

The following inverse mapping theorems for outer graphical derivatives form the theoretical basis for many of the computations in what follows.

Theorem 2.1 (see Theorem 4.1 of [26] and Theorem 3.1 of [23]). *For any multifunction $M : \mathbb{R}^n \times \mathbb{R}^d \rightrightarrows \mathbb{R}^n$ and any triple of points $(\bar{v}, \bar{w}, \bar{x})$ satisfying $\bar{v} \in$*

$M(\bar{x},\bar{w})$, consider the partial inverse multifunction $S:\mathbb{R}^n\times\mathbb{R}^d\rightrightarrows\mathbb{R}^n$ defined by

$$S(v,w):=\{x|v\in M(x,w)\} \tag{2.2}$$

and its restriction $S_0:\mathbb{R}^d\rightrightarrows\mathbb{R}^n$ defined by $S_0(w):=S(0,w)$. The outer graphical derivative of S at $(\bar{v},\bar{w})$ for $\bar{x}$ satisfies

$$DS(\bar{v},\bar{w}|\bar{x})(v',w')=\{x'|v'\in DM(\bar{x},\bar{w}|\bar{v})(x',w')\}, \tag{2.3}$$

and the outer graphical derivative of S_0 at $\bar{w}$ for $\bar{x}$ satisfies

$$DS_0(\bar{w}|\bar{x})(w')\subseteq\{x'|0\in DM(\bar{x},\bar{w}|0)(x',w')\}.$$

If the multifunction M is protodifferentiable at $(\bar{x},\bar{w})$ for $\bar{v}\in M(\bar{x},\bar{w})$, then the partial inverse multifunction S is protodifferentiable at $(\bar{v},\bar{w})$ for $\bar{x}$.

Remark. The explicit presence of the parameter v in the partial inverse multifunction S is essential for obtaining either the identity (2.3) or the protodifferentiability result. In the case when the mapping M satisfies the conditions of the classical inverse mapping theorem, Theorem 2.1 essentially reproduces the classical result (though obviously without the classical differentiability conclusions).

2.1. Differentiability of stationary points. A necessary condition for optimality in (1.2) that generalizes the Fermat rule is the inclusion $0\in\partial_x f(x,w)$ in terms of the partial subgradient multifunction $\partial_x f:\mathbb{R}^{n+d}\rightrightarrows\mathbb{R}^n$ (see [52, Theorem 10.1]). Recall from [52] that a vector $v\in\mathbb{R}^n$ is an element of the image set $\partial_x f(x,w)$ if and only if there exist sequences $v_n\to v$ and $x_n\to x$ such that $f(x_n,w)\to f(x,w)$, and for each fixed x'

$$f(x',w)\geq f(x_n,w)+\langle v_n,x'-x_n\rangle+o(|x'-x_n|).$$

Using the necessary condition for optimality $0\in\partial_x f(x,w)$ as a basis, the set of stationary points associated with (1.2) for with the parameter $w\in\mathbb{R}^d$ is the image set of the *stationary point multifunction* defined by

$$SP(w):=\{x|0\in\partial_x f(x,w)\}. \tag{2.4}$$

Notice that this multifunction is of the type S_0 covered by Theorem 2.1, so we know that its outer graphical derivative satisfies the estimate

$$D\big(SP\big)(\bar{w}|\bar{x})(w')\subseteq\{x'|0\in D\big(\partial_x f\big)(\bar{x},\bar{w}|0)(x',w')\} \tag{2.5}$$

in terms of the outer graphical derivative of the partial subgradient mapping. According to Theorem 2.1, if the partial subgradient multifunction is protodifferentiable, then the stationary point multifunction associated with a slightly modified problem is protodifferentiable too, and, moreover, in this case, the inclusion corresponding to (2.5) is an equality. The necessary modification involves a new parameter $v\in\mathbb{R}^n$ which serves to tilt the graph of the objective function:

$$\min\{f(x,w)-\langle v,x\rangle\}\text{ over }x\in\mathbb{R}^n. \tag{2.6}$$

Note that the new tilt parameter is precisely the kind of parameterization we saw for the simple examples considered in the introduction (though there we used the general parameter label w since there were no other parameters).

COROLLARY 2.1.1. *If the multifunction $\partial_x f$ is protodifferentiable at $(\bar{x}, \bar{w})$ for $\bar{v} \in \partial_x f(\bar{x}, \bar{w})$, then the stationary point multifunction associated with the problem* (2.6)

$$SP_{tilt}(v, w) := \{x | v \in \partial_x f(x, w)\} \tag{2.7}$$

is protodifferentiable at $(\bar{v}, \bar{w})$ for $\bar{x}$ and the identity holds that

$$D\big(SP_{tilt}\big)(\bar{v}, \bar{w}|\bar{x})(v', w') = \{x' | v' \in D\big(\partial_x f\big)(\bar{x}, \bar{w}|\bar{v})(x', w')\}.$$

Notice that Corollary 2.1.1 does not say anything directly about the protodifferentiability of the stationary point multifunction (2.4) associated with the original "untilted" optimization problem. However, indirect information about the differential properties of SP is available from the protodifferentiability of SP_{tilt} recorded in Corollary 2.1.1, and it can be extracted from $D\big(SP_{tilt}\big)$ according to the particular situation. In traditional situations when SP_{tilt} is B-differentiable, this procedure is trivial and the B-differentiability of SP follows immediately.

PROPOSITION 2.1.1. *If the stationary point multifunction SP_{tilt}* (2.7) *is (single-valued) continuous near $(0, \bar{w})$, the partial subgradient multifunction $\partial_x f$ is protodifferentiable at $(\bar{x}, \bar{w})$ for $0 \in \partial_x f(\bar{x}, \bar{w})$ and the outer graphical derivative mapping*

$$D\big(SP_{tilt}\big)(0, \bar{w}|\bar{x})(v', w') = \{x' | v' \in D\big(\partial_x f\big)(\bar{x}, \bar{w}|0)(x', w')\} \tag{2.8}$$

is single-valued, then SP_{tilt} is actually B-differentiable at $(0, \bar{w})$ with B-derivative given by (2.8).

If SP_{tilt} is Lipschitz continuous near $(0, \bar{w})$, then the same result holds without the assumption about the single-valuedness of (2.8) *(which single-valuedness is in this event assured by the B-differentiability).*

Moreover, in either of these situations, the stationary point mapping SP (2.4) *associated with the untilted problem is also B-differentiable at $\bar{w}$ with B-derivative given by*

$$D\big(SP\big)(\bar{w}|\bar{x})(w') = \{x' | 0 \in D\big(\partial_x f\big)(\bar{x}, \bar{w}|0)(x', w')\}.$$

Proof. This follows from Proposition 2.0.1 and Corollary 2.1.1 since $SP(w)$ is the same as $SP_{tilt}(0, w)$, so the joint B-differentiability of SP_{tilt} implies B-differentiability of the stationary point mapping SP. □

Remark. In the classical case when f is a twice differentiable function of x alone, both results in this section say that the stationary point mapping associated with the tilted minimization of f is protodifferentiable, and that its protoderivative is given by the inverse mapping associated with the Hessian $\nabla^2 f$.

2.2. Differentiability in the case of fully amenable functions. According to Corollary 2.1.1 and Proposition 2.1.1, the protodifferentiability of the partial subgradient multifunction $\partial_x f$ is the key to establishing differentiability properties for stationary point multifunctions. One particularly important class of composite functions in optimization has been shown to have protodifferentiable partial subgradient multifunctions.

Definition. A proper extended real-valued function g on $\mathbb{R}^m$ is called *piecewise linear-quadratic* if its effective domain $\operatorname{dom} g := \{y \in \mathbb{R}^m | g(y) < \infty\}$ can be represented as the union of finitely many polyhedral sets, relative to each of which $g(y)$ is

given by an expression of the form $\langle y, Ay\rangle/2 + \langle a, y\rangle + \alpha$ for some scalar $\alpha \in \mathbb{R}$, vector $a \in \mathbb{R}^m$, and symmetric matrix $A \in \mathbb{R}^{m\times m}$.

A proper extended real-valued function f on $\mathbb{R}^{n+d}$ is called *amenable in x at $\bar{x}$ with compatible parameterization in w at $\bar{w}$* if $f(\bar{x}, \bar{w})$ is finite and on some neighborhood $X \times W \subseteq \mathbb{R}^{n+d}$ of $(\bar{x}, \bar{w})$ there is a representation $f(x, w) = g\big(G(x, w)\big)$ in which G is a $\mathcal{C}^1$ mapping from $X \times W$ into $\mathbb{R}^m$, while g is a proper, lower semicontinuous, convex function on $\mathbb{R}^m$, such that the following constraint qualification is fulfilled:

$$0 \in \operatorname{int}\Big(\operatorname{dom} g - \big[G(\bar{x}, \bar{w}) + \nabla_x G(\bar{x}, \bar{w})\mathbb{R}^n\big]\Big). \tag{2.9}$$

If in addition such a representation exists with G not just $\mathcal{C}^1$ but $\mathcal{C}^2$ and with g piecewise linear-quadratic, then f is said to be *fully amenable in x at $\bar{x}$ with compatible parameterization in w at $\bar{w}$*.

The class of fully amenable functions covers many of the functions commonly involved in finite-dimensional optimization (see [52]), including maxima of finitely many $\mathcal{C}^2$ functions, as well as the essential objective associated with a nonlinear program (1.4) under the Mangasarian–Fromovitz constraint qualification. In [28], fully amenable functions with compatible parameterization were shown to have protodifferentiable partial subgradient multifunctions.

PROPOSITION 2.2.1 (see Theorem 1.1 of [28]). *If f is fully amenable in x at $\bar{x}$ with compatible parameterization in w at $\bar{w}$, then for all $(x, w) \in \mathbb{R}^{n+d}$ sufficiently close to $(\bar{x}, \bar{w})$ the partial subgradient multifunction $\partial_x f$ is protodifferentiable at (x, w) for $v \in \partial_x f(x, w)$. Moreover, in this case, the image sets of the outer graphical derivative $D\big(\partial_x f\big)(x, w|v)$ can be computed by taking the union over all $u \in \mathbb{R}^d$ satisfying $(v, u) \in \partial f(x, w)$ (for the full subgradient multifunction) of the projections onto $\mathbb{R}^n$ of the image sets for the outer graphical derivatives $D\big(\partial f\big)(x, w|v, u)$.*

Remark. In [43], this result was extended to a more general class of amenable functions where the function g is allowed to have $\mathcal{C}^2$ pieces in place of quadratic ones.

According to Proposition 2.2.1 fully amenable functions have protodifferentiable partial subgradient mappings, so the results of the previous section apply to give sensitivity information about the stationary points associated with the minimization of fully amenable functions. In the following section, we explore this further for one important particular case of fully amenable minimization.

2.3. Differentiability of stationary points for nonlinear programs. In [41] and [42], chain rules for the outer graphical derivatives of the full subgradient multifunction are worked out, and these can be combined with Proposition 2.2.1 to obtain more specific formulas for the outer graphical derivatives of the partial subgradient multifunction in particular cases that fit the model covered by Proposition 2.2.1. For example, this process was carried out in [27] for the case of the nonlinear program (1.3) with $\mathcal{C}^2$ functions g_i for $i = 0, \ldots, m$, and we will outline these results here under the same assumptions. The formulas obtained in [27] involved the polyhedral cone $Q(x, w) \subset \mathbb{R}^{n+d}$ defined by

$$Q(x, w) := \left\{(x', w') : \begin{array}{l} \langle \nabla g_i(x, w), (x', w')\rangle \le 0 \text{ for } i \in [1, s] \text{ with } g_i(x, w) = 0, \\ \langle \nabla g_i(x, w), (x', w')\rangle = 0 \text{ for } i \in [s+1, m] \end{array}\right\}, \tag{2.10}$$

the mapping

$$G(x, w) := \big(g_1(x, w), \ldots, g_m(x, w)\big), \tag{2.11}$$

the set

$$K := \{z \in \mathbb{R}^m | z_i \leq 0 \text{ for } i = 1, \ldots, s \text{ and } z_i = 0 \text{ for } i = s+1, \ldots, m\}, \tag{2.12}$$

its associated convex normal cone mapping $N_K : \mathbb{R}^m \rightrightarrows \mathbb{R}^m$, which is empty-valued at $z \notin K$ and is defined at $z \in K$ by

$$N_K(z) = \{y \in \mathbb{R}^m | \langle y, k - z\rangle \leq 0 \text{ for all } k \in K\}, \tag{2.13}$$

and the usual Lagrangian mapping $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^d$ defined by

$$L(x, y, w) := g_0(x, w) + y_1 g_1(x, w) + \cdots + y_m g_m(x, w).$$

To state the result from [27] we also need certain sets of multiplier vectors, first the bounded, polyhedral set

$$Y(x, w, v, u) := \big\{y = (y_1, \ldots, y_m) \in N_K\big(G(x, w)\big) : \nabla_{(x,w)} L(x, y, w) = (v, u)\big\} \tag{2.14}$$

and its face

$$Y_{\max}(x, w, v, u; x', w') = \underset{y \in Y(x,w,v,u)}{\arg\max} \left\langle (x', w'), \nabla^2_{(x,w)(x,w)} L(x, y, w) \cdot (x', w') \right\rangle \tag{2.15}$$

and then the polyhedral cone

$$Y'(x, w; x', w') = \left\{ \begin{array}{r} y' = (y'_1, \ldots, y'_m) \in N_K\big(G(x, w)\big) : \\ y'_i = 0 \text{ for } i \text{ with } \left\langle \nabla g_i(x, w), (x', w') \right\rangle \neq 0 \end{array} \right\}. \tag{2.16}$$

Remark. Notice that in the case when the constraint functions are affine, the face (2.15) is the entire set of multipliers $Y(x, w, v, u)$.

Another special case is when the parameterization $w = (w_1, w_2)$ includes the "canonical" parameter $w_2 \in \mathbb{R}^m$ which perturbs the constraint functions as follows: $g_i(x, w_1) + \big[w_2\big]_i$. In this case, the multiplier set $Y(x, w, v, u)$ for $u = (u_1, u_2)$ is empty unless u_2 is in the normal cone $N_K\big(G(x, w)\big)$ and $u_1 = \nabla_{w_1} g_0(x, w_1) - u_2 \cdot \nabla_{w_1} G(x, w)$, in which case $Y(x, w, v, u)$ reduces to the singleton $\{u_2\}$. It follows that in this case the face $Y_{\max}(x, w, v, u; x', w')$ is the same as the set $Y(x, w, v, u)$ regardless of the choice of x' and w'. However, the set $Y_{\max}(x, w, v, u; x', w')$ is frequently a proper subset of the set of all multipliers as in the case of the following example.

Example program. Consider the minimization problem

$$\min\{g_0(x_1, x_2, w) := (x_1)^2 - x_1 + (x_2)^2\} \text{ over all } x \in C(w),$$

where the constraint set is defined as follows:

$$C(w) := \{x \in \mathbb{R}^2 : g_1(x_1, x_2, w) := x_1 - x_2^2 + w^2 \leq 0 \text{ and } g_2(x_1, x_2, w) := x_1 - w^2 \leq 0\}.$$

For the base values $x = (0, 0)$, $w = 0$, and $v = (0, 0)$, the set of Lagrange multipliers is empty unless $u = 0$ in which case it consists of the $y \in \mathbb{R}^2_+$ satisfying $y_1 + y_2 = 1$, and the Hessian $\nabla^2_{(x,w)(x,w)} L\left((0, 0), y, 0\right)$ is the matrix

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 2(1 - y_1) & 0 \\ 0 & 0 & 2(y_1 - y_2) \end{bmatrix}.$$

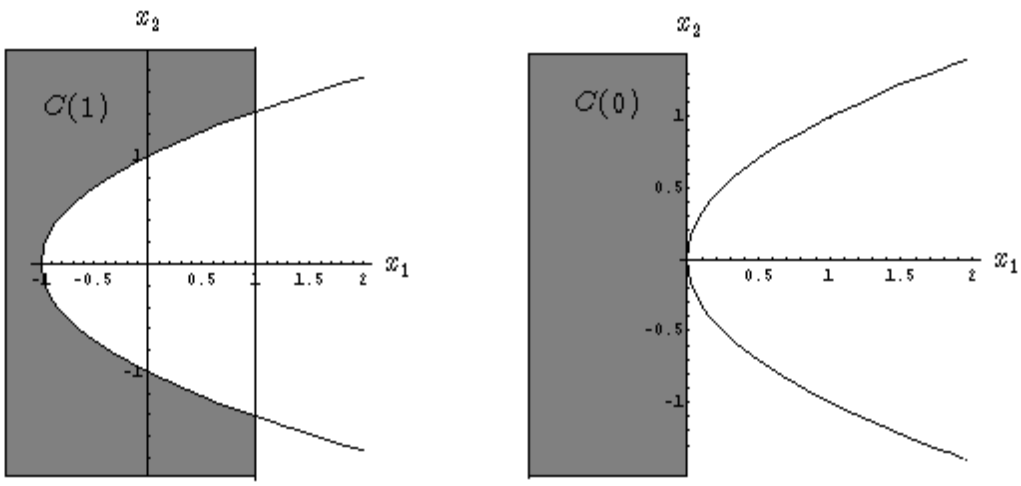


FIG. 2.1. *Constraint set for example program.*

It follows that the inner product defining the face $Y_{\max}\left((0,0),0,(0,0),0;x',w'\right)$ reduces to $2(x_1')^2 + 2(1-y_1)(x_2')^2 + 2(y_1-y_2)(w')^2$ which is maximized at the unique multiplier pair $(1,0)$ for any choice of $x' = (x_1', x_2')$ and w' with $x_2' = 0$ and $w' \neq 0$ or $x_2' \leq w' \neq 0$; thus in these cases, the face $Y_{\max}\left((0,0),0,(0,0),0;x',w'\right)$ contains only the pair $(1,0)$. On the other hand, if $w' = 0$ and $x_2' \neq 0$ or $0 \neq x_2' > w'$, then the face $Y_{\max}\left((0,0),0,(0,0),0;x',w'\right)$ consists solely of the multiplier pair $(0,1)$. Finally, if both $w' = 0$ and $x_2' = 0$, then the face $Y_{\max}\left((0,0),0,(0,0),0;x',w'\right)$ is the entire set of multipliers.

PROPOSITION 2.3.1 (see Theorem 3.2 of [27]). *For the nonlinear program* (1.3) *and its associated essential objective function* f (1.4)*, if the Mangasarian–Fromovitz constraint qualification holds at* $(\bar{x},\bar{w})$*,*

$$\nexists y = (y_1,\ldots,y_m) \in N_K\big(G(\bar{x},\bar{w})\big) \text{ with} \quad y_1\nabla_x g_1(\bar{x},\bar{w}) + \cdots + y_m\nabla_x g_m(\bar{x},\bar{w}) = 0, \text{ except } y = 0, \tag{2.17}$$

then for all pairs (x,w) *sufficiently close to* $(\bar{x},\bar{w})$ *and all* $v \in \partial_x f(x,w)$*, the partial subgradient multifunction for the essential objective function* (1.4) *is protodifferentiable at* (x,w) *for* v *and its outer graphical derivative can be computed as follows: For* $(x',w') \notin Q(x,w)$*, the set* $D\big(\partial_x f\big)(x,w|v)(x',w')$ *is empty. But for* $(x',w') \in Q(x,w)$*, the outer graphical derivative image set* $D\big(\partial_x f\big)(x,w|v)(x',w')$ *consists of all vectors*

$$\nabla^2_{x(x,w)} L(x,y,w)(x',w') + \sum_{i=1}^{m} y_i'\nabla_x g_i(x,w) - y_0'\big(v - \nabla_x g_0(x,w)\big)$$

generated by $y' \in Y'(x,w;x',w')$ *and* $y_0' \in \mathbb{R}$ *along with choices of* y *for which there exists* $u \in \mathbb{R}^d$ *with* $\langle (v,u),(x',w')\rangle = \langle \nabla g_0(x,w),(x',w')\rangle$ *such that* y *is an element of* $Y_{\max}(x,w,v,u;x',w')$*. (If no such choice of* y *is possible, then again* $D\big(\partial_x f\big)(x,w|v)(x',w')$ *is empty.)*

The characterization in Proposition 2.3.1 of the outer graphical derivative can be combined with the results in Corollary 2.1.1 and Proposition 2.1.1 to obtain the following formulas for the outer graphical derivatives of the stationary point multifunctions

associated with the nonlinear program (1.3):

$$SP(w) := \big\{x | \exists y \in N_K\big(G(x,w)\big) \text{ with } 0 = \nabla_x L(x,y,w)\big\} \tag{2.18}$$

and

$$SP_{tilt}(v,w) := \big\{x | \exists y \in N_K\big(G(x,w)\big) \text{ with } v = \nabla_x L(x,y,w)\big\}. \tag{2.19}$$

PROPOSITION 2.3.2 (see Theorem 3.1 of [27]). *For the nonlinear program* (1.3) *and its associated stationary point multifunctions SP* (2.18) *and* SP_{tilt} (2.19)*, if the Mangasarian–Fromovitz constraint qualification* (2.17) *holds at* $(\bar{x},\bar{w})$ *with* $\bar{x} \in SP(\bar{w})$*, then* SP_{tilt} *is protodifferentiable at* $(0,\bar{w})$ *for* $\bar{x}$ *and the image set of the outer graphical derivative* $D\big(SP_{tilt}\big)(0,\bar{w}|\bar{x})(v',w')$ *is equal to the set of* x' *such that* $(x',w') \in Q(\bar{x},\bar{w})$ *and there exist* $y' \in Y'(\bar{x},\bar{w};x',w')$*,* $y'_0 \in \mathbb{R}$*,* $\bar{u} \in \mathbb{R}^d$ *with* $\langle \bar{u}, w'\rangle = \langle \nabla g_0(\bar{x},\bar{w}), (x',w')\rangle$*, and* $\bar{y} \in Y_{\max}(\bar{x},\bar{w},0,\bar{u};x',w')$ *which together satisfy*

$$v' = \nabla^2_{x(x,w)} L(\bar{x},\bar{y},\bar{w}) \cdot (x',w') + y'_0 \nabla_x g_0(\bar{x},\bar{w}) + \sum_{i=1}^{m} y'_i \nabla_x g_i(\bar{x},\bar{w}). \tag{2.20}$$

Moreover, the image set of the outer graphical derivative $D\big(SP\big)(\bar{w}|\bar{x})(w')$ *is included in the set* $D\big(SP_{tilt}\big)(0,\bar{w}|\bar{x})(0,w')$ *from above.*

Proof. This follows from Corollary 2.1.1 and the inclusion (2.5) using the formula for the outer graphical derivative of $\partial_x f$ given in Proposition 2.3.1. □

Example program revisited. For the example program given at the beginning of this section and for $\bar{x} = (0,0)$ and $\bar{w} = 0$, both constraints are active and their gradients in x are both equal to $(1,0)$ so the Mangasarian–Fromovitz constraint qualification is satisfied, and Proposition 2.3.2 applies. In this case, the cone $Q(\bar{x},\bar{w})$ is the set of triples (x'_1,x'_2,w') with $x'_1 \le 0$, and $Y_{\max}(\bar{x},\bar{w},0,\bar{u};x',w')$ is empty unless $\bar{u} = 0$. When $\bar{u} = 0$, the condition

$$\langle \bar{u}, w'\rangle = \langle \nabla g_0(\bar{x},\bar{w}), (x',w')\rangle$$

reduces to $x'_1 = 0$, which implies that $Y'(\bar{x},\bar{w};x',w') = \mathbb{R}^2_+$. The identity (2.20) then reduces to

$$\begin{aligned} v'_1 &= -y'_0 + y'_1 + y'_2, \\ v'_2 &= 2(1-\bar{y}_1)x'_2 \end{aligned} \tag{2.21}$$

for some choices of $\bar{y} \in Y_{\max}(\bar{x},\bar{w},0,\bar{u};x',w')$, $y'_0 \in \mathbb{R}$, and $y' \in \mathbb{R}^2_+$. No matter what $v'_1 \in \mathbb{R}$ is, the first identity can be satisfied by choosing $y'_0 = -v'_1$ and $y'_1 = y'_2 = 0$. The form of the outer graphical derivative $D\big(SP_{tilt}\big)(0,\bar{w}|\bar{x})(v',w')$ thus depends entirely on the values of v'_2 and w'. We already have established that this outer graphical derivative has zero first component $x'_1 = 0$, and it remains to compute the possible second components x'_2 in this situation.

If $v'_2 = 0$, then the second identity in (2.21) is satisfied only when $\bar{y}_1 = 1$ or when $x'_2 = 0$. From our earlier analysis of the face $Y_{\max}(\bar{x},\bar{w},0,\bar{u};x',w')$, we know that $\bar{y}_1 = 1$ only when $x'_2 = 0$ or $x'_2 \le w' \ne 0$.

If $v'_2 \ne 0$, then the second identity in (2.21) is satisfied only when $\bar{y}_1 \ne 1$ in which case $x'_2 = v'_2/2(1-\bar{y}_1)$. From our earlier analysis of the face $Y_{\max}(\bar{x},\bar{w},0,\bar{u};x',w')$, it is clear that $\bar{y}_1 = 0$ is the only feasible option, in which case $x'_2 = v'_2/2$. This only

occurs, however, if $w' = 0$ or $v_2' > 2w'$. It follows that the outer graphical derivative for the tilted stationary point mapping associated with this example satisfies

$$D\big(SP_{tilt}\big)(0,\bar{w}|\bar{x})(v_1',v_2',w') = \begin{cases} \{0\}\times(\{0\}\cup(-\infty,w']) & \text{if } w'\neq 0 \text{ and } v_2'=0,\\ \{0\}\times\{v_2'/2\} & \text{if } w'=0 \text{ or } 0\neq v_2'>2w',\\ \emptyset & \text{otherwise.}\end{cases}$$

Moreover, the outer graphical derivative for the untilted stationary point mapping associated with this example satisfies

$$D\big(SP\big)(\bar{w}|\bar{x})(w') \subseteq \begin{cases} \{0\}\times(\{0\}\cup(-\infty,w']) & \text{if } w'\neq 0,\\ \{0\}\times\{0\} & \text{if } w'=0.\end{cases}$$

Notice that the derivative formulas for this example were computed without first computing the stationary points, so sensitivity information about stationary points is available without the stationary points themselves.

The next result is just the translation of Proposition 2.1.1 using the formula in Proposition 2.3.1.

PROPOSITION 2.3.3 (see Theorem 3.3 of [27]). *If in addition to the assumptions of Proposition* 2.3.2*, the stationary point multifunction SP_{tilt} is (single-valued) continuous near $(0,\bar{w})$, then it is B-differentiable at $(0,\bar{w})$ if for every pair (v',w') there is only one point x' for which $(x',w')\in Q(\bar{x},\bar{w})$ and there exist $y'\in Y'(\bar{x},\bar{w};x',w')$, $y_0'\in\mathbb{R}$, $\bar{u}\in\mathbb{R}^d$ with $\langle\bar{u},w'\rangle = \langle\nabla g_0(\bar{x},\bar{w}),(x',w')\rangle$, and $\bar{y}\in Y_{\max}(\bar{x},\bar{w},0,\bar{u};x',w')$ which together satisfy* (2.20)*.*

If SP_{tilt} is Lipschitz continuous near $(0,\bar{w})$, then it is automatically B-differentiable without any assumptions about uniqueness of solutions to (2.20) *(which uniqueness is in this event assured by the B-differentiability).*

Moreover, under either of these assumptions on SP_{tilt}, the stationary point mapping SP (2.18) *is B-differentiable at $\bar{w}$ with B-derivative $D\big(SP\big)(\bar{w}|SP(\bar{w}))(w')$ equal to the (unique) point x' for which $(x',w')\in Q(\bar{x},\bar{w})$ and for which there exist $y'\in Y'(\bar{x},\bar{w};x',w')$, $y_0'\in\mathbb{R}$, $\bar{u}\in\mathbb{R}^d$ with $\langle\bar{u},w'\rangle = \langle\nabla g_0(\bar{x},\bar{w}),(x',w')\rangle$, and $\bar{y}$ and element of $Y_{\max}(\bar{x},\bar{w},0,\bar{u};x',w')$ which together satisfy*

$$0 = \nabla^2_{x(x,w)}L(\bar{x},\bar{y},\bar{w})\cdot(x',w') + y_0'\nabla_x g_0(\bar{x},\bar{w}) + \sum_{i=1}^m y_i'\nabla_x g_i(\bar{x},\bar{w}).$$

2.4. Differentiability of Karush–Kuhn–Tucker pairs. In this section, we return to the study of general tilted optimization problems (2.6) but now where the objective function f takes the special form of a sum of a $\mathcal{C}^1$ function g_0 and an amenable function $g\big(G(x,w)+v_2\big)$ perturbed inside the composition by $v_2\in\mathbb{R}^m$ (see section 2.2 for the definition of amenability):

$$\min\{g_0(x,w) + g\big(G(x,w)+v_2\big) - \langle v_1,x\rangle\} \text{ over } x\in\mathbb{R}^n. \tag{2.22}$$

This optimization problem is a generalization of a canonically perturbed (by v_1 and v_2) nonlinear program since that model is covered by the mapping G (2.11) and the function g equal to the "indicator function" δ_K associated with the set K (2.12):

$$g(z) = \delta_K(z) := \begin{cases} 0 & \text{if } z\in K,\\ \infty & \text{otherwise.}\end{cases}$$

Of course, we lose nothing by explicitly including canonical perturbations, since sensitivity information for the unperturbed problem

$$\min\{g_0(x,w) + g\big(G(x,w)\big)\} \text{ over } x \in \mathbb{R}^n \tag{2.23}$$

can be recovered from the canonically perturbed model with $(v_1, v_2) = (0,0)$.

We can define a generalized Lagrangian function on $\mathbb{R}^{n+m+d}$ by

$$L(x,y,w) := g_0(x,w) + \langle y, G(x,w)\rangle,$$

as well as the generalized Karush–Kuhn–Tucker (KKT) pairs associated with parameters $(v,w) = (v_1, v_2, w) \in \mathbb{R}^{n+m} \times \mathbb{R}^d$, which are the pairs $(x,y) \in \mathbb{R}^{n+m}$ satisfying

$$\begin{aligned} &\nabla_x L(x,y,w) - v_1 = 0 \text{ and} \\ &G(x,w) + v_2 \in \big(\partial g\big)^{-1}(y), \end{aligned}$$

where $\big(\partial g\big)^{-1}(y)$ denotes the set of vectors $z \in \mathbb{R}^m$ for which $y \in \partial g(z)$. Under the constraint qualification (2.9) stipulated by amenability, stationary points x for the problem (2.22) for parameters (v,w) close to $(0,\bar{w})$ can always be paired with a multiplier vector $y \in \mathbb{R}^m$ so that the pair (x,y) is a KKT pair for (v,w) (see [52, Exercise 10.26]). To study these stationary point-multiplier pairs, we define the KKT pair multifunction whose value at (v,w) is the set of KKT pairs for (v,w)

$$KKT(v,w) := \left\{(x,y) | v \in \Big(\nabla_x L(x,y,w), -G(x,w)\Big) + \Big(\{0\}^n \times \big(\partial g\big)^{-1}(y)\Big)\right\}. \tag{2.24}$$

From this formulation, the role of the canonical perturbations is clear; since the parameter v appears explicitly, the KKT pair multifunction belongs to the class of multifunctions S covered by our basic Theorem 2.1. According to Theorem 2.1 then, the protodifferentiability of the KKT pair multifunction is assured under the protodifferentiability of the multifunction

$$M(x,y,w) := \Big(\nabla_x L(x,y,w), -G(x,w)\Big) + \Big(\{0\}^n \times \big(\partial g\big)^{-1}(y)\Big). \tag{2.25}$$

Proposition 2.4.1. *If the mapping $\nabla_x L$ is B-differentiable at $(\bar{x},\bar{y},\bar{w})$ and the subgradient multifunction ∂g is protodifferentiable at $\bar{z} := \bar{v}_2 + G(\bar{x},\bar{w})$ for $\bar{y} \in \partial g(\bar{z})$, then the KKT pair multifunction* (2.24) *is protodifferentiable at $(\bar{v},\bar{w})$ for $(\bar{x},\bar{y})$ and the image set of its outer graphical derivative $D\big(KKT\big)(\bar{v},\bar{w}|\bar{x},\bar{y})(v_1', v_2', w')$ is given by*

$$\left\{(x',y') \;\middle|\; \begin{array}{l} v_1' = D\left(\nabla_x L\right)(\bar{x},\bar{y},\bar{w}|\nabla_x L(\bar{x},\bar{y},\bar{w}))\,(x',y',w') \\ v_2' + \nabla G(\bar{x},\bar{w})\cdot(x',w') \in D\left(\big(\partial g\big)^{-1}\right)\big(\bar{y}|\bar{z}\big)(y') \end{array}\right\}. \tag{2.26}$$

Moreover, the image set at $w' \in \mathbb{R}^d$ of the outer graphical derivative at $\bar{w}$ for $(\bar{x},\bar{y})$ of the KKT pair multifunction $w \mapsto KKT(0,w)$ associated with the unperturbed problem (2.23) *is contained in the set*

$$\left\{(x',y') \;\middle|\; \begin{array}{l} 0 = D\left(\nabla_x L\right)(\bar{x},\bar{y},\bar{w}|\nabla_x L(\bar{x},\bar{y},\bar{w}))\,(x',y',w') \\ \nabla G(\bar{x},\bar{w})\cdot(x',w') \in D\left(\big(\partial g\big)^{-1}\right)\big(\bar{y}|G(\bar{x},\bar{w})\big)(y') \end{array}\right\}. \tag{2.27}$$

Proof. This follows from Theorem 2.1 since according to [26, Propositions 2.2 and 3.4], the protodifferentiability of M (2.25) is ensured by the assumptions. To show the formula (2.26) for the outer graphical derivative, we define the multifunction

$$M_2(x, y, w) := -G(x, w) + \big(\partial g\big)^{-1}(y)$$

so that the outer graphical derivative $DM(\bar{x}, \bar{y}, \bar{w}|\bar{v})(x', y', w')$ is the set

$$\Big(D\left(\nabla_x L\right)(\bar{x}, \bar{y}, \bar{w}|\nabla_x L(\bar{x}, \bar{y}, \bar{w}))\,(x', y', w'), DM_2(\bar{x}, \bar{y}, \bar{w}|\bar{v}_2)(x', y', w')\Big) \tag{2.28}$$

and the outer graphical derivative of M_2 is given by

$$DM_2(\bar{x}, \bar{y}, \bar{w}|\bar{v})(x', y', w') = -\nabla G(\bar{x}, \bar{w}) \cdot (x', w') + D\Big(\big(\partial g\big)^{-1}\Big)(\bar{y}|\bar{z})(y'). \tag{2.29}$$

The formula for the outer graphical derivative then follows directly from Theorem 2.1. ∎

Remark. Proposition 2.4.1 is quite broad since many optimization problems satisfy its assumptions: The mapping $\nabla_x L$ is B-differentiable under very mild additional assumptions on the function g_0 and the mapping G, and there are many known protodifferentiable subgradient multifunctions associated with convex functions g, including all cases when g is piecewise linear-quadratic.

A result similar to Proposition 2.4.1 but covering a slightly different model was obtained in [51], where the formula for the outer graphical derivative was given in terms of the solutions to an auxiliary optimization problem.

Following the approach outlined in section 2.1, we can study the B-differentiability of KKT pairs associated with the unperturbed and untilted problem (where $v = 0$) from the outer graphical derivative formula in Proposition 2.4.1.

PROPOSITION 2.4.2. *Under the assumptions of Proposition* 2.4.1 *with* $\bar{v} = 0$*, if either the KKT pair multifunction* (2.24) *is Lipschitz continuous near* $(0, \bar{w})$ *or it is (single-valued) continuous near* $(0, \bar{w})$ *and the outer graphical derivative* $D\big(KKT\big)$ *at* $(0, \bar{w})$ *for* $(\bar{x}, \bar{y})$ *is single-valued, then the KKT pair multifunction* (2.24) *is B-differentiable at* $(0, \bar{w})$*.*

Moreover, under either of these conditions, the KKT pair mapping $w \mapsto KKT(0, w)$ *associated with the optimization problem* (2.23) *is B-differentiable at* $\bar{w}$ *with B-derivative evaluated at* $w' \in \mathbb{R}^d$ *equal to the unique pair* $(x', y') \in \mathbb{R}^{n+m}$ *in the set* (2.27)*.*

2.5. Differentiability of KKT pairs for nonlinear programs. In this section, we move again to our most specific level by translating the results in section 2.4 into the case of nonlinear programs with canonical perturbations v_1 and v_2

$$\min\{g_0(x, w) - \langle v_1, x\rangle + \delta_K\big(G(x, w) + v_2\big)\}, \tag{2.30}$$

where the set K given by (2.12) and $\mathcal{C}^2$ functions g_i for $i = 0, \ldots, m$ define the mapping G given by (2.11). The Lagrangian function associated with this program is automatically $\mathcal{C}^2$, so the key to accessing the differentiability results in section 2.4 is the protodifferentiability of the subgradient multifunction associated with the indicator function $g(z) = \delta_K(z)$. The subgradient multifunction associated with δ_K is just the convex normal cone mapping $N_K : \mathbb{R}^m \rightrightarrows \mathbb{R}^m$ (2.13), which according to Proposition (2.2.1) is protodifferentiable under the Mangasarian–Fromovitz constraint qualification (2.17). Moreover, at any $\bar{z} \in K$ and for $\bar{y} \in N_K(\bar{z})$ the outer graphical

derivative of N_K is empty-valued at $z' \in \mathbb{R}^m$ unless $z' \in Z(\bar{z}, \bar{y})$, where

$$Z(\bar{z}, \bar{y}) := \left\{ z' \in \mathbb{R}^m \;\middle|\; \begin{array}{ll} z'_i \leq 0 & \text{for } i \in [1,s] \text{ with } \bar{z}_i = 0 \text{ and } \bar{y}_i = 0 \\ z'_i = 0 & \text{for } i \in [1,s] \text{ with } \bar{z}_i = 0 \text{ and } \bar{y}_i > 0 \\ z'_i = 0 & \text{for } i \in [s+1,m] \end{array} \right\},$$

in which case it satisfies

$$D\left(N_K\right)(\bar{z}|\bar{y})(z') = \left\{ y' \in \mathbb{R}^m \;\middle|\; \begin{array}{ll} y'_i \geq 0 & \text{for } i \in [1,s] \text{ with } \bar{z}_i = 0, \bar{y}_i = 0, \text{ and } z'_i = 0 \\ y'_i = 0 & \text{for } i \in [1,s] \text{ with } \bar{z}_i = 0, \bar{y}_i = 0, \text{ and } z'_i < 0 \\ y'_i = 0 & \text{for } i \in [1,s] \text{ with } \bar{z}_i < 0 \end{array} \right\}. \tag{2.31}$$

The next result follows from these facts and uses the following sets of indices:

$$\begin{aligned} I_1 &:= \{i \in [s+1,m] \text{ and } i \in [1,s] \text{ with } \bar{y}_i > 0 = g_i(\bar{x}, \bar{w})\}, \\ I_2 &:= \{i \in [1,s] \text{ with } \bar{y}_i = 0 = g_i(\bar{x}, \bar{w})\}, \\ I_3 &:= \{i \in [1,s] \text{ with } \bar{y}_i = 0 > g_i(\bar{x}, \bar{w})\}. \end{aligned} \tag{2.32}$$

Proposition 2.5.1. *For the canonically perturbed nonlinear program* (2.30) *and its associated KKT pair multifunction*

$$KKT(v,w) := \left\{ (x,y) | v \in \Big(\nabla_x L(x,y,w), -G(x,w) \Big) + \Big(\{0\}^n \times \left(N_K\right)^{-1}(y) \Big) \right\}, \tag{2.33}$$

if the Mangasarian–Fromovitz constraint qualification (2.17) *holds at* $(\bar{x}, \bar{w})$ *with* $(\bar{x}, \bar{y}) \in KKT(0, \bar{w})$*, then the KKT pair multifunction is protodifferentiable at* $(0, \bar{w})$ *for* $(\bar{x}, \bar{y}) \in KKT(0, \bar{w})$*, and the image set of outer graphical derivative*

$$D\big(KKT\big)(0, \bar{w} | \bar{x}, \bar{y})(v'_1, v'_2, w')$$

is the set of pairs $(x', y') \in \mathbb{R}^{n+m}$ *that satisfy*

$$\begin{aligned} & v'_1 = \nabla^2_{x(x,y,w)} L(\bar{x}, \bar{y}, \bar{w}) \cdot (x', y', w'), \\ & y'_i \geq 0 \text{ for } i \in [1,s] \text{ with } \bar{y}_i = 0, \\ & y'_i = 0 \text{ for } i \in I_3, \\ & \langle \nabla g_i(\bar{x}, \bar{w}), (x', w') \rangle + \Big[v'_2 \Big]_i \leq 0 \text{ for } i \in I_2 \text{ with } y'_i = 0, \\ & \langle \nabla g_i(\bar{x}, \bar{w}), (x', w') \rangle + \Big[v'_2 \Big]_i = 0 \text{ for } i \in I_2 \text{ with } y'_i > 0, \\ & \langle \nabla g_i(\bar{x}, \bar{w}), (x', w') \rangle + \Big[v'_2 \Big]_i = 0 \text{ for } i \in I_1. \end{aligned} \tag{2.34}$$

Moreover, the image set at $w' \in \mathbb{R}^d$ *of the outer graphical derivative at* $\bar{w}$ *for* $(\bar{x}, \bar{y})$ *of the KKT pair multifunction* $w \mapsto KKT(0, w)$ *associated with the nonlinear program* (1.3) *without canonical perturbations is contained in the set of pairs* $(x', y') \in \mathbb{R}^{n+m}$ *that satisfy the same six conditions* (2.34) *but with* $v'_1 = 0$ *and* $v'_2 = 0$.

Proof. According to the preceding discussion, the protodifferentiability of the KKT pair mapping follows from Proposition 2.4.1, as do the formulas for the outer graphical derivatives once we develop a formula for the outer graphical derivative of the inverse of the normal cone multifunction N_K. Since the outer graphical derivative of any multifunction S is defined in terms of the graph of S, the outer graphical derivative of the inverse multifunction S^{-1} is just the inverse of the outer graphical

derivative of S. It follows from the formula (2.31) for the outer graphical derivative of N_K then that the outer graphical derivative of $\big(N_k\big)^{-1}$ at $\bar{y}$ for $G(\bar{x},\bar{w})$ is empty-valued at $y' \in \mathbb{R}^m$ unless y' satisfies $y'_i \geq 0$ for $i \in [1,s]$ with $\bar{y}_i = 0$ and $y'_i = 0$ for $i \in I_3$, in which case the image set of the outer graphical derivative is equal to

$$\left\{ z' \in \mathbb{R}^m \;\middle|\; \begin{array}{ll} z'_i \leq 0 & \text{for } i \in I_2 \text{ with } y'_i = 0 \\ z'_i = 0 & \text{for } i \in I_2 \text{ with } y'_i > 0 \\ z'_i = 0 & \text{for } i \in I_1 \end{array} \right\}. \tag{2.35}$$

The formula claimed then follows from Proposition 2.4.1 after substituting $z' = v' + \nabla G(\bar{x},\bar{w}) \cdot (x', w')$. □

Example program revisited. We consider again the example program from section 2.3, but this time with canonical perturbations included.

$$\min\{(x_1)^2 - x_1 + (x_2)^2 - [v_1]_1 x_1 - [v_1]_2 x_2\} \text{ over all } x \in C(w, v_2),$$

where the constraint set is defined as follows:

$$C(w, v_2) := \{x \in \mathbb{R}^2 : x_1 - x_2^2 + w^2 + [v_2]_1 \leq 0 \text{ and } x_1 - w^2 + [v_2]_2 \leq 0\}.$$

The functions g_0, g_1, and g_2 and the Lagrangian are the same as in section 2.3, and we use the same base points $\bar{x} = (0,0)$ and $\bar{w} = 0$. There are a variety of $\bar{y}$ which can be paired with $\bar{x}$ to form a KKT pair in this case; namely, any $\bar{y} \in \mathbb{R}^2_+$ satisfying $\bar{y}_1 + \bar{y}_2 = 1$. The matrix $\nabla^2_{x(x,y,w)} L(\bar{x},\bar{y},\bar{x})$ is given by

$$\begin{bmatrix} 2 & 0 & 1 & 1 & 0 \\ 0 & 2(1-\bar{y}_1) & 0 & 0 & 0 \end{bmatrix}$$

and the index set I_3 is always empty for our choice of $(\bar{x},\bar{w})$. If we choose $\bar{y} = (1,0)$, the other index sets satisfy $I_1 = \{1\}$ and $I_2 = \{2\}$, and the six conditions in Proposition 2.5.1 become

$$\begin{aligned} &[v'_1]_1 = 2x'_1 + y'_1 + y'_2, \\ &[v'_1]_2 = 0, \\ &y'_2 \geq 0, \\ &x'_1 + [v'_2]_2 \leq 0 \text{ if } y'_2 = 0, \\ &x'_1 + [v'_2]_2 = 0 \text{ if } y'_2 > 0, \\ &x'_1 + [v'_2]_1 = 0. \end{aligned}$$

Thus Proposition 2.5.1 says that the outer graphical derivative

$$D\big(KKT\big)(0,\bar{w}|\bar{x},\bar{y})(v'_1, v'_2, w')$$

for $\bar{y} = (1,0)$ is empty unless $[v'_1]_2 = 0$ and $[v'_2]_2 \leq [v'_2]_1$, in which case it equals the set

$$\big\{(x', y') \in \mathbb{R}^2 \times \mathbb{R}^2 \,|\quad x'_1 = -[v'_2]_1, \quad y' = ([v'_1]_1 + 2[v'_2]_1, 0)\big\}$$

if $[v'_2]_2 < [v'_2]_1$, but instead equals the set

$$\big\{(x', y') \in \mathbb{R}^2 \times \mathbb{R}^2 \,|\quad x'_1 = -[v'_2]_1, \quad y'_2 \geq 0, \quad y'_1 = [v'_1]_1 + 2[v'_2]_1 - y'_2\big\}$$

in the case that $[v_2']_2 = [v_2']_1$.

If we consider instead the multiplier $\bar{y} = (0, 1)$, the index sets satisfy $I_1 = \{2\}$ and $I_2 = \{1\}$, and the same set of conditions work out as

$$\begin{aligned}
&[v_1']_1 = 2x_1' + y_1' + y_2',\\
&[v_1']_2 = 2x_2',\\
&y_1' \geq 0,\\
&x_1' + [v_2']_1 \leq 0 \text{ if } y_1' = 0,\\
&x_1' + [v_2']_1 = 0 \text{ if } y_1' > 0,\\
&x_1' + [v_2']_2 = 0.
\end{aligned}$$

Thus Proposition 2.5.1 says that the outer graphical derivative

$$D\big(KKT\big)(0, \bar{w}|\bar{x}, \bar{y})(v_1', v_2', w')$$

for $\bar{y} = (0, 1)$ equals the set

$$\left\{(x', y') \in \mathbb{R}^2 \times \mathbb{R}^2 \,|\quad x' = (-[v_2']_2, [v_1']_2/2)\,,\quad y' = (0, [v_1']_1 + 2[v_2']_2)\right\}$$

if $[v_2']_1 < [v_2']_2$ but instead equals the set

$$\left\{(x', y') \in \mathbb{R}^2 \times \mathbb{R}^2 \,|\quad x' = (-[v_2']_2, [v_1']_2/2)\,,\quad y_1' \geq 0,\quad y_2' = [v_1']_1 + 2[v_2']_2 - y_1'\right\}$$

in the case that $[v_2']_1 = [v_2']_2$ and is empty if $[v_2']_1 > [v_2']_2$.

For any other choice of multiplier $\bar{y}$, the index set I_2 is empty and $I_1 = \{1, 2\}$, so the conditions work out as

$$\begin{aligned}
&[v_1']_1 = 2x_1' + y_1' + y_2',\\
&[v_1']_2 = 2(1 - \bar{y}_1)x_2',\\
&x_1' + [v_2']_1 = 0,\\
&x_1' + [v_2']_2 = 0.
\end{aligned}$$

Thus Proposition 2.5.1 says that the outer graphical derivative

$$D\big(KKT\big)(0, \bar{w}|\bar{x}, \bar{y})(v_1', v_2', w')$$

in this case is empty unless $[v_2']_1 = [v_2']_2$, in which case it equals the set

$$\left\{(x', y') \in \mathbb{R}^2 \times \mathbb{R}^2 \,|\quad x' = (-[v_2']_1, [v_1']_2/\left(2(1 - \bar{y}_1)\right))\,,\ y_1' = [v_1']_1 + 2[v_2']_1 - y_2',\ y_2' \in \mathbb{R}\right\}.$$

The result in Proposition 2.5.1 for the case without the canonical perturbations is simpler for this example, and implies that the outer graphical derivative at $\bar{w}$ for $(\bar{x}, \bar{y})$ of the KKT pair multifunction $w \mapsto KKT(0, w)$ for the different possible choices of $\bar{y}$ is contained in the sets

$$\begin{aligned}
&\{(x', y') \in \mathbb{R}^2 \times \mathbb{R}^2 \,|\quad x_1' = 0,\quad y_1' \leq 0,\quad y_2' = -y_1'\} \text{ when } \bar{y} = (1, 0),\\
&\{(x', y') \in \mathbb{R}^2 \times \mathbb{R}^2 \,|\quad x_1' = 0,\quad x_2' = 0,\quad y_1' \geq 0,\quad y_2' = -y_1'\} \text{ when } \bar{y} = (0, 1),\\
&\{(x', y') \in \mathbb{R}^2 \times \mathbb{R}^2 \,|\quad x_1' = 0,\quad x_2' = 0,\quad y_1' \in \mathbb{R},\quad y_2' = -y_1'\} \text{ for other } \bar{y}.
\end{aligned}$$

Notice again that the derivative formulas for this example were computed without first computing the KKT pairs, so sensitivity information is available without the KKT pairs themselves.

Finally, we apply Proposition 2.4.2 to the canonically perturbed nonlinear program (2.30) to get the following sufficient conditions for B-differentiability of KKT pairs associated with the nonlinear program without the canonical perturbations.

PROPOSITION 2.5.2. *Under the Mangasarian–Fromovitz constraint qualification* (2.17)*, if either the KKT pair multifunction* (2.33) *is Lipschitz continuous near* $(0, \bar{w})$ *or it is (single-valued) continuous near* $(0, \bar{w})$ *and its outer graphical derivative* $D\big(KKT\big)(0, \bar{w}|\bar{x}, \bar{y})$ *is single-valued, then the KKT pair multifunction is B-differentiable at* $(0, \bar{w})$.

Moreover, the KKT pair multifunction $w \mapsto KKT(0, w)$ *associated with the nonlinear program* (1.3) *is B-differentiable at* $\bar{w}$ *with B-derivative evaluated at* $w' \in \mathbb{R}^d$ *equal to the unique pair* $(x', y') \in \mathbb{R}^{n+m}$ *satisfying the six conditions* (2.34) *with* $v_1' = v_2' = 0$.

3. Continuity. The continuity property on which we focus is a slightly restricted form of local Lipschitz continuity called "calmness." A function $x(w)$ on $\mathbb{R}^d$ is *calm at* $\bar{w}$ if there is a constant $L > 0$ such that for all w near $\bar{w}$, the function satisfies $|x(w) - x(\bar{w})| \leq L|w - \bar{w}|$. Notice that this is the same as the usual local Lipschitz continuity except that the base point $\bar{w}$ is always one of the points of comparison. A multifunction $S : \mathbb{R}^d \rightrightarrows \mathbb{R}^n$ is said to have *calm selections near* $(\bar{w}, \bar{x})$ if there exist neighborhoods $W \subseteq \mathbb{R}^d$ of $\bar{w}$ and $X \subseteq \mathbb{R}^n$ of $\bar{x}$ together with a constant $L > 0$ such that any local selection $x(w) \in S(w) \cap X$ for $w \in W$ satisfies

$$|x(w) - \bar{x}| \leq L|w - \bar{w}|. \tag{3.1}$$

Calmness is an important property to study when analyzing solutions to parameterized optimization problems since it gives a Lipschitz bound on the distance between perturbed and unperturbed solutions. The term "calmness" was used by Clarke [6] to describe an optimization problem whose optimal value function obeyed a Lipschitz bound with a fixed base point as in (3.1), and our use of this terminology was inspired by the earlier notion. Note also that property we are calling local selection calmness has been widely studied under various labels. In [23] it was called "local upper Lipschitz continuity," in [7] it was called "upper Lipschitz continuity at a point," and in [4] it was called "semistability" in the context of variational inequalities. Finally, in [40], [39], and [12] the term "stability" was used to indicate when certain multifunctions associated with nonlinear equations satisfied condition (ii) and at the same time had nonempty local image sets $S(w) \cap X$.

The results in this section are based on the following generalization of the classical inverse mapping theorem that gives a characterization of the calmness of selections in terms of the outer graphical derivative.

THEOREM 3.1 (see Proposition 4.1 of [23]). *For any multifunction* $S : \mathbb{R}^d \rightrightarrows \mathbb{R}^n$ *and any pair* $(\bar{w}, \bar{x}) \in \operatorname{gph} S$*, the following are equivalent:*

(i) *The outer graphical derivative image set* $DS(\bar{w}|\bar{x})(0)$ *equals the trivial set* $\{0\}$.

(ii) S *has calm selections near* $(\bar{w}, \bar{x})$.

Remark. Note that the implication (i) $\Rightarrow$ (ii) was shown first in [13, Proposition 2.1]. Theorem 3.1 has the advantage over the classical inverse mapping theorem that the outer graphical derivative exists for all multifunctions. This means that we can rule in or out useful Lipschitz bounds in practically any situation we might encounter simply by computing the outer graphical derivative in the nonsingularity condition (i) above. This is very important in terms of sensitivity analysis of quasi-solutions to optimization problems, since verifying calmness directly demands that quasi-solutions are reliably known (which is not typical if the problems are sensitive),

whereas the nonsingularity condition (i) demands only that we be able to compute a generalized derivative (which we can do as in the preceding chapter). Moreover, since the nonsingularity condition (i) characterizes the local selection calmness property (ii), any other sufficient condition for local selection calmness must essentially also involve the nonsingularity condition (i), so it makes sense to focus our attention on (i).

Example program revisited. In section 2.3, we computed outer graphical derivative estimates for stationary point mappings associated with both tilted and untilted versions of a parameterized nonlinear program. We can use these estimates to apply Theorem 3.1 to deduce whether those stationary point mappings have calm selections. From the formulas worked out in section 2.3, it is clear that the outer graphical derivative of the tilted stationary point mapping $D\left(SP_{tilt}\right)(0,\bar{w}|\bar{x})(0,0,0)$ equals the trivial set $\{0\}\times\{0\}$, so Theorem 3.1 guarantees calm selections near $(0,\bar{w},\bar{x})$. Likewise, the estimate for the outer graphical derivative of the untilted stationary point mapping ensures that $D\left(SP\right)(\bar{w}|\bar{x})(0)$ is contained in the trivial set $\{0\}\times\{0\}$, and since the trivial set is included in the image of any outer graphical derivative evaluated at 0, we can again apply Theorem 3.1 to deduce that the untilted stationary point mapping has calm selections near $(\bar{w},\bar{x})$. Of course, the untilted stationary point mapping is a restriction of the tilted stationary point mapping, so calm selections for the former are expected in the presence of calm selections for the latter.

In general, it is not necessary to compute the full outer graphical derivative mapping in order to decide whether a certain mapping has calm selections, since only the image at 0 is required by Theorem 3.1. In the following sections, we use the general formulas for outer graphical derivatives computed in the previous sections to translate Theorem 3.1 in some cases of particularly important mappings S. As in the previous chapter then, we use a fundamental general result (in this case Theorem 3.1) as the starting point for progressively more specific results. Along the way, layers of complication are added and restrictions are made as we pass from the general Theorem 3.1 to the language of the particular applications.

3.1. Continuity of stationary points. According to the estimate (2.5) for the outer graphical derivative of the stationary point mapping, the following sufficient condition for the local selection calmness property is immediate from Theorem 3.1.

PROPOSITION 3.1.1. *For the stationary point multifunction* SP (2.4) *and any pair* $(\bar{x},\bar{w})\in\mathbb{R}^{n+d}$ *with* $0\in\partial_x f(\bar{x},\bar{w})$*, if* $x'=0$ *is the only vector in* $\mathbb{R}^n$ *satisfying*

$$0\in D\big(\partial_x f\big)(\bar{x},\bar{w}|0)(x',0),$$

then SP *has calm selections near* $(\bar{w},\bar{x})$*.*

The gap between the sufficient condition in Proposition 3.1.1 and the local selection calmness property for the stationary point mapping SP is evident from the following proposition characterizing this same property for SP_{tilt} (2.7).

PROPOSITION 3.1.2. *For the stationary point multifunction* SP_{tilt} (2.7) *and any pair* $(\bar{x},\bar{w})\in\mathbb{R}^{n+d}$ *with* $0\in\partial_x f(\bar{x},\bar{w})$*, the following are equivalent:*

(i) $x'=0$ *is the only vector in* $\mathbb{R}^n$ *satisfying* $0\in D\big(\partial_x f\big)(\bar{x},\bar{w}|0)(x',0)$.

(ii) SP_{tilt} *has calm selections near* $(0,\bar{w},\bar{x})$*.*

Proof. This follows directly from Theorem 3.1 and the remark following Theorem 2.1. □

Remark. According to Proposition 3.1.2, the sufficient condition in Proposition 3.1.1 for local selection calmness for the stationary point mapping SP is equivalent to the local selection calmness of the tilted stationary point mapping SP_{tilt}. Thus

the gap between the sufficient condition (i) and the local selection calmness property for SP is filled precisely by the stationary point mappings SP that have calm local selections even though SP_{tilt} does not. An easy example that belongs in this category involves the unconstrained minimization of the smooth function $f(x,w) = x^4/4$. The lone stationary point for this function (for all parameters w) is $x = 0$, but as we saw in the introduction, the stationary point for the tilted minimization $\min\{x^4/4 - v\cdot x\}$ is $x = v^{1/3}$. In this case, the stationary point multifunction $SP(w) = 0$ is trivially calm at 0 while the mapping $SP_{tilt}(v,w) = v^{1/3}$ is not. Notice that the outer graphical derivative in this case satisfies $D\big(\partial_x f\big)(0,\bar{w}|0) = \nabla^2_{x(x,w)} f(0) = (0,0)$, so condition (i) above is certainly not satisfied.

3.2. Continuity of stationary points for nonlinear programs. In the case of the nonlinear program (1.3), the results in the preceding section can be translated into more familiar terms using the formulas for the outer graphical derivative of $\partial_x f$ from Proposition 2.3.1.

PROPOSITION 3.2.1 (see Theorem 5.1 of [24]). *For the nonlinear program* (1.3) *and its associated stationary point multifunction SP* (2.18)*, if the Mangasarian–Fromovitz constraint qualification* (2.17) *holds at* $(\bar{x},\bar{w})$ *with* $\bar{x} \in SP(\bar{w})$*, then SP has calm selections near* $(\bar{w},\bar{x})$ *if the condition holds that*

(i) $x' = 0$ *is the only vector satisfying* $(x',0) \in Q(\bar{x},\bar{w})$ (2.10)*,* $\nabla_x g_0(\bar{x},\bar{w})\cdot x' = 0$*, and*

$$0 = \nabla^2_{xx} L(\bar{x},\bar{y},\bar{w})\cdot x' + \sum_{i=1}^{m} y_i' \nabla_x g_i(\bar{x},\bar{w}) + y_0' \nabla_x g_0(\bar{x},\bar{w}) \tag{3.2}$$

for some $y' \in Y'(\bar{x},\bar{w};x',0)$ (2.16) *and* $y_0' \in \mathbb{R}$ *along with a vector* $\bar{y}$ *for which there exists* $\bar{u} \in \mathbb{R}^d$ *such that* $\bar{y} \in Y_{\max}(\bar{x},\bar{w},0,\bar{u};x',0)$ (2.15)*.*

Moreover, condition (i) *is equivalent to the stationary point multifunction* SP_{tilt} (2.19) *having calm selections near* $(0,\bar{w},\bar{x})$*.*

Remark. Notice that condition (i) is equivalent to $x' = 0$ being the only vector in $\mathbb{R}^n$ that is a stationary point (i.e., satisfies the KKT conditions) for any of the auxiliary problems with constraint system $(x',0) \in Q(\bar{x},\bar{w})$ with $\nabla_x g_0(\bar{x},\bar{w})\cdot x' = 0$ and with objective functions $x' \mapsto \langle x', \nabla^2_{xx} L(\bar{x},\bar{y},\bar{w})\cdot x'\rangle$, where $\bar{y}$ is any vector in the set $Y_{\max}(\bar{x},\bar{w},0,\bar{u};x',0)$ defined by any $\bar{u} \in \mathbb{R}^d$.

In the special case where the canonical constraint perturbation $w_2 \in \mathbb{R}^m$ is present $g_i(x,w_1) + [w_2]_i$, any Lagrange multiplier $\bar{y}$ (so $\bar{y} \in N_K\big(G(x,w)\big)$ with $\nabla_x L(x,y,w) = 0$) can be realized as an element of $Y_{\max}(\bar{x},\bar{w},0,\bar{u};x',0)$ for some $\bar{u} \in \mathbb{R}^d$, so the condition in Proposition 3.2.1 is equivalent to $x' = 0$ being the only vector in $\mathbb{R}^n$ that is a stationary point for any of the auxiliary problems with constraint system $(x',0) \in Q(\bar{x},\bar{w})$ with $\nabla_x g_0(\bar{x},\bar{w})\cdot x' = 0$ and with objective functions $x' \mapsto \langle x', \nabla^2_{xx} L(\bar{x},\bar{y},\bar{w})\cdot x'\rangle$ where $\bar{y}$ is any Lagrange multiplier. A flawed version of this result was stated in [10, Theorem 3.1] and a correct version was subsequently proved in [15] where the authors translated Theorem 3.1 for "generalized Kojima-functions" whose zeroes can be identified, for instance, with the stationary points of the nonlinear program (1.3).

If we take the product of x' with each side of (3.2), we can derive an even more familiar form of sufficient condition.

PROPOSITION 3.2.2 (see Corollary 5.2 of [24]). *Condition* (i) *from Proposition* 3.2.1 *holds if the condition*

$$\max_{y \in Y(\bar{x},\bar{w},0,\bar{u})} \langle x', \nabla^2_{xx} L(\bar{x},y,\bar{w})\cdot x'\rangle > 0$$

holds for all $\bar{u} \in \mathbb{R}^d$ *such that* $Y(\bar{x}, \bar{w}, 0, \bar{u}) \neq \emptyset$ (2.14) *and all nonzero* x' *satisfying* $(x', 0) \in Q(\bar{x}, \bar{w})$ (2.10) *and* $\nabla_x g_0(\bar{x}, \bar{w}) \cdot x' = 0$.

Remark. Notice that there is a larger gap between the condition in Proposition 3.2.2 and the local selection calmness property for SP than the one associated with condition (i) of Proposition 3.2.1. In particular, the condition in Proposition 3.2.2 demands the positive-definiteness of $\nabla_{xx} L(\bar{x}, y, \bar{w})$ while condition (i) also holds if $\nabla_{xx} L(\bar{x}, y, \bar{w})$ is negative-definite. Assuming a positive-definite $\nabla_{xx} L(\bar{x}, y, \bar{w})$ of course also addresses the issue of the stationary points being optimal solutions to the original minimization problem, which is one reason why conditions like the one in Proposition 3.2.2 are more typical in the literature. However, our approach yields a condition (i) that is closer to characterizing the desired property of local selection calmness, and so our approach more closely captures the essence of the stability property in which we are interested.

Example program revisited. At the beginning of section 3, we showed that the nonlinear program first introduced in section 2.3 has calm local selections from its stationary point mappings. We will show now that this example program, moreover, satisfies the conditions of Proposition 3.2.2 for the base point $\bar{x} = (0, 0)$ and the base parameter $\bar{w} = 0$. In this case, the Hessian matrix $\nabla_{xx} L(\bar{x}, y, \bar{w})$ is just

$$\begin{bmatrix} 2 & 0 \\ 0 & 2(1 - y_1) \end{bmatrix}$$

and the set $Y(\bar{x}, \bar{w}, 0, \bar{u})$ is only nonempty when $\bar{u} = 0$, in which case it equals the set of $y \in \mathbb{R}^2_+$ with $y_1 + y_2 = 1$. The cone $Q(\bar{x}, \bar{w})$ is just $\mathbb{R}_- \times \mathbb{R}^2$ and the condition $\nabla_x g_0(\bar{x}, \bar{w}) \cdot x' = 0$ is satisfied only by $x' = (x'_1, x'_2)$ with $x'_1 = 0$. It follows that the inner product in Proposition 3.2.2 reduces to $2(1 - y_1)(x'_2)^2$ which is maximized over $Y(\bar{x}, \bar{w}, 0, \bar{u})$ when $y_1 = 0$, with maximum value $2(x'_2)^2$. This maximum value is clearly nonzero as long as x'_2 is nonzero, so the condition in Proposition 3.2.2 is satisfied.

It is interesting to note that the Hessian of the Lagrangian for this example is not positive definite for all multipliers in $Y(\bar{x}, \bar{w}, 0, \bar{u})$ (when $y = (1, 0)$ the Hessian matrix is only positive semidefinite), so this program does not satisfy many of the standard second-order sufficient conditions for stability available prior to Proposition 3.2.2.

3.3. Continuity of KKT pairs. We return to the optimization model with canonical perturbations $v = (v_1, v_2)$

$$\min\{g_0(x, w) + g\big(G(x, w) + v_2\big) - \langle v_1, x\rangle\} \text{ over } x \in \mathbb{R}^n$$

from section 2.4, but now with an eye toward studying the continuity of its KKT pairs. As in the preceding two sections, Theorem 3.1 is the basis for the results here.

PROPOSITION 3.3.1. *If the mapping* $\nabla_x L$ *is B-differentiable at* $(\bar{x}, \bar{y}, \bar{w})$ *and the subgradient multifunction* ∂g *is protodifferentiable at* $G(\bar{x}, \bar{w})$ *for* $\bar{y} \in \partial g\big(G(\bar{x}, \bar{w})\big)$, *then for the KKT pair multifunction* (2.24), *the following are equivalent:*

(i) *The pair* $(x', y') = (0, 0)$ *is the only solution to the conditions*

$$0 = D\left(\nabla_x L\right)(\bar{x}, \bar{y}, \bar{w} | \nabla_x L(\bar{x}, \bar{y}, \bar{w}))(x') + \sum_{i=1}^{m} y'_i \nabla_x g_i(\bar{x}, \bar{w})$$

$$\nabla_x G(\bar{x}, \bar{w}) \cdot x' \in D\left(\left(\partial g\right)^{-1}\right)\left(\bar{y} | G(\bar{x}, \bar{w})\right)(y'). \tag{3.3}$$

(ii) *KKT has calm selections near* $(0, \bar{w}, \bar{x}, \bar{y})$.

Proof. This follows directly from Theorem 3.1 and Proposition 2.4.1. ∎

3.4. Continuity of KKT pairs for nonlinear programs. For nonlinear programs with canonical perturbations (2.30), the results of section 3.3 can be translated into even more specific terms. Again we use the index sets (2.32).

PROPOSITION 3.4.1. *For the canonically perturbed nonlinear program* (2.30) *and its associated KKT pair multifunction* (2.33)*, if the Mangasarian–Fromovitz constraint qualification* (2.17) *holds at* $(\bar{x}, \bar{w})$ *with* $(\bar{x}, \bar{y}) \in KKT(0, \bar{w})$*, then the following are equivalent:*

(i) *The pair* $(x', y') = (0, 0)$ *is the only solution to the conditions*

$$\begin{aligned}
&0 = \nabla^2_{xx} L(\bar{x}, \bar{y}, \bar{w}) \cdot x' + \sum_{i=1}^{m} y'_i \nabla_x g_i(\bar{x}, \bar{w}),\\
&y'_i \geq 0 \text{ for } i \in [1, s] \text{ with } \bar{y}_i = 0,\\
&y'_i = 0 \text{ for } i \in I_3,\\
&\langle \nabla_x g_i(\bar{x}, \bar{w}), x' \rangle \leq 0 \text{ for } i \in I_2 \text{ with } y'_i = 0,\\
&\langle \nabla_x g_i(\bar{x}, \bar{w}), x' \rangle = 0 \text{ for } i \in I_2 \text{ with } y'_i > 0,\\
&\langle \nabla_x g_i(\bar{x}, \bar{w}), x' \rangle = 0 \text{ for } i \in I_1.
\end{aligned}$$

(ii) KKT *has calm selections near* $(0, \bar{w}, \bar{x}, \bar{y})$.

Proof. This follows directly from Proposition 3.3.1 and the formula (2.35) for the outer graphical derivative of the inverse of N_K. □

The continuity property (ii) characterized in Proposition 3.4.1 is of course stronger than the property (ii) characterized in Proposition 3.2.1, since condition (ii) here stipulates joint calmness with respect to both stationary points and multipliers, whereas (ii) of Proposition 3.2.1 deals exclusively with stationary points. However, it is not as obvious that condition (i) of Proposition 3.2.1 is implied by (i) of Proposition 3.4.1 (which it must be if these characterizations are consistent), so we briefly show this fact here.

Proof of Proposition 3.4.1 (i) ⇒ *Proposition* 3.2.1 (i). Suppose x' satisfies $(x', 0) \in Q(\bar{x}, \bar{w}), \nabla_x g_0(\bar{x}, \bar{w}) \cdot x' = 0$, and

$$0 = \nabla^2_{xx} L(\bar{x}, \bar{y}, \bar{w}) \cdot x' + \sum_{i=1}^{m} y'_i \nabla_x g_i(\bar{x}, \bar{w}) + y'_0 \nabla_x g_0(\bar{x}, \bar{w})$$

as in Proposition 3.2.1 (i) for some $y' \in Y'(\bar{x}, \bar{w}; x', 0)$ and $y'_0 \in \mathbb{R}$ along with a vector $\bar{y}$ for which there exists $\bar{u}$ such that $\bar{y} \in Y_{\max}(\bar{x}, \bar{w}, 0, \bar{u}; x', 0)$. According to Proposition 3.4.1, under (i) (which is equivalent to (ii)) there is only one multiplier $\bar{y}$ such that $(\bar{x}, \bar{y}) \in KKT(0, \bar{w})$, so this is the only vector for which there exists $\bar{u}$ such that $\bar{y} \in Y_{\max}(\bar{x}, \bar{w}, 0, \bar{u}; x', 0)$. The conditions from Proposition 3.2.1 (i) then become

$$\begin{aligned}
&\langle \nabla_x g_i(\bar{x}, \bar{w}), x' \rangle \leq 0 \text{ for } i \in I_1 \cup I_2,\\
&\langle \nabla_x g_i(\bar{x}, \bar{w}), x' \rangle = 0 \text{ for } i \in [s+1, m] \cup \{0\},\\
&0 = \nabla^2_{xx} L(\bar{x}, \bar{y}, \bar{w}) \cdot x' + \sum_{i=1}^{m} y'_i \nabla_x g_i(\bar{x}, \bar{w}) + y'_0 \nabla_x g_0(\bar{x}, \bar{w})
\end{aligned}$$

for some $y'_0 \in \mathbb{R}$ and some $y' \in N_K\big(G(\bar{x}, \bar{w})\big)$ satisfying $y'_i \nabla_x g_i(\bar{x}, \bar{w}) \cdot x' = 0$ for $i \in [1, m]$. If we define $\tilde{y}' := y' - y_{0'} \bar{y}$ and use the fact that

$$0 = \nabla_x g_0(\bar{x}, \bar{w}) + \sum_{i=1}^{m} \bar{y}_i \nabla_x g_i(\bar{x}, \bar{w}),$$

we conclude that $\tilde{y}'$ satisfies

$$0 = \nabla^2_{xx} L(\bar{x}, \bar{y}, \bar{w}) \cdot x' + \sum_{i=1}^{m} \tilde{y}'_i \nabla_x g_i(\bar{x}, \bar{w}). \tag{3.4}$$

Since both $\bar{y}$ and y' are in the set $N_K\big(G(\bar{x}, \bar{w})\big)$, we know that $\tilde{y}'$ also satisfies

$$\tilde{y}'_i = 0 \text{ for } i \in I_3. \tag{3.5}$$

Finally, since $\tilde{y}'_i = y'_i$ if $\bar{y}_i = 0$, we also know that $\tilde{y}'$ satisfies

$$\tilde{y}'_i \geq 0 \text{ for } i \in [1, s] \text{ with } \bar{y}_i = 0 \tag{3.6}$$

and

$$\langle \nabla_x g_i(\bar{x}, \bar{w}), x' \rangle = 0 \text{ for } i \in I_2 \text{ with } \tilde{y}'_i > 0. \tag{3.7}$$

The conditions (3.4)–(3.7) show that the pair $(x', \tilde{y}')$ satisfies the system in condition (i) of Proposition 3.4.1. It follows from this condition that x' must equal zero, so condition (i) of Proposition 3.2.1 is verified. □

It follows directly from Proposition 3.4.1 that its condition (i) is sufficient to ensure the local selection calmness property for the KKT pairs associated with the nonlinear program (1.3) without the canonical perturbations.

PROPOSITION 3.4.2. *For the KKT pair multifunction* (2.33)*, if the Mangasarian–Fromovitz constraint qualification* (2.17) *holds at* $(\bar{x}, \bar{w})$ *with* $(\bar{x}, \bar{y}) \in KKT(0, \bar{w})$*, then either of the equivalent conditions in Proposition* 3.4.1 *is enough to ensure the existence of neighborhoods* $X \times Y \subseteq \mathbb{R}^{n+m}$ *of* $(\bar{x}, \bar{y})$ *and* $W \subseteq \mathbb{R}^d$ *of* $\bar{w}$ *as well as a constant* $L > 0$ *such that any local selection* $\big(x(w), y(w)\big) \in KKT(0, w) \cap \big(X \times Y\big)$ *of KKT pairs for the nonlinear program* (1.3) *for* $w \in W$ *satisfies the estimate*

$$|\big(x(w), y(w)\big) - (\bar{x}, \bar{y})| \leq L|w - \bar{w}|.$$

Example program revisited. In section 2.5, we added canonical perturbations to our example program and computed the outer graphical derivatives of the associated KKT pair multifunction at the different possible multipliers associated with $\bar{x} = (0, 0)$. The conditions in Proposition 3.4.1 are of course just the result of evaluating these outer graphical derivatives at zero, so they too work out differently for the different possible multipliers $\bar{y}$: When $\bar{y} = (1, 0)$, the conditions in Proposition 3.4.1 are satisfied by any element of the set

$$\left\{(x', y') \in \mathbb{R}^2 \times \mathbb{R}^2 | \quad x'_1 = 0, \quad y'_2 \geq 0, \quad y'_1 = -y'_2\right\},$$

but when $\bar{y} = (0, 1)$, the conditions in Proposition 3.4.1 are satisfied by any element of the set

$$\left\{(x', y') \in \mathbb{R}^2 \times \mathbb{R}^2 | \quad x' = (0, 0), \quad y'_1 \geq 0, \quad y'_2 = -y'_1\right\},$$

and when $\bar{y}$ is any other multiplier, the conditions in Proposition 3.4.1 are satisfied by any element of the set

$$\left\{(x', y') \in \mathbb{R}^2 \times \mathbb{R}^2 | \quad x' = (0, 0), \quad y'_1 = -y'_2\right\}.$$

Since none of these sets is trivial, it follows from Proposition 3.4.1 that the KKT pair multifunction does not have calm selections near $(0, \bar{w}, \bar{x}, \bar{y})$ at any of the possible KKT pairs $(\bar{x}, \bar{y})$ associated with the base point $\bar{x} = (0, 0)$ and the base parameter $\bar{w} = 0$. Notice that we cannot deduce anything from Proposition 3.4.2 about the calmness of selections from the KKT pair multifunction $w \mapsto KKT(0, w)$ for the program without the canonical perturbations, since we know only that a sufficient condition for this property is violated.

It is interesting to note that this example program did have calm selections from its associated stationary point mappings where multipliers were not involved directly, so this example highlights the gap between Propositions 3.2.1 and 3.4.1.

4. Existence and uniqueness. In contrast to our experience with continuity and differentiability properties, it has proven to be more difficult to characterize the properties of existence and uniqueness in terms of simple, verifiable conditions on generalized derivatives. In the introduction, we already discussed why it is not possible to use derivatives to characterize existence and uniqueness properties without involving stability properties at the same time. To discuss such combinations of properties, we introduce some additional generalized derivative concepts. One of these is the coderivative developed by Mordukhovich [31]. For a multifunction $S : \mathbb{R}^d \rightrightarrows \mathbb{R}^n$, the *coderivative of S at $\bar{w}$ for an element $\bar{x} \in S(\bar{w})$* is the multifunction $D^*S(\bar{w}|\bar{x}) : \mathbb{R}^n \rightrightarrows \mathbb{R}^d$ whose graph is obtained by the following transformation of the set $N_{\operatorname{gph} S}(\bar{w}, \bar{x})$ of normals to $\operatorname{gph} S := \{(w, x) : x \in S(w)\}$ at $(\bar{w}, \bar{x})$ (this normal cone $N_{\operatorname{gph} S}(\bar{w}, \bar{x})$ is equal to the subgradient of the indicator function $\delta_{\operatorname{gph} S}$ associated with the set $\operatorname{gph} S$; see [31] or [52]):

$$w' \in D^*S(\bar{w}|\bar{x})(x') \Leftrightarrow (w', -x') \in N_{\operatorname{gph} S}(\bar{w}, \bar{x}).$$

In the case of a single-valued mapping $x : \mathbb{R}^d \to \mathbb{R}^n$ that is $\mathcal{C}^1$ near $\bar{w}$, the coderivative coincides with the adjoint of the Jacobian $\nabla x(\bar{w})^*$. A calculus for the coderivative was developed by Mordukhovich in [37], and, moreover, the coderivative can be used to characterize a generalized Lipschitz property for multifunctions whose graphs are locally closed sets near a base point. A set K is *locally closed at c* if $K \cap C$ is closed for some closed neighborhood C of c (see [52]).

THEOREM 4.1 (see Theorem 5.4 of [32]).[1] *For any multifunction $S : \mathbb{R}^d \rightrightarrows \mathbb{R}^n$ and any pair $(\bar{w}, \bar{x})$ satisfying $\bar{x} \in S(\bar{w})$, if the set $\operatorname{gph} S$ is locally closed at $(\bar{w}, \bar{x})$, then the following are equivalent:*

(i) *The coderivative image set $D^*S(\bar{w}|\bar{x})(0)$ equals the trivial set $\{0\}$.*

(ii) *There exist neighborhoods $X \subseteq \mathbb{R}^n$ of $\bar{x}$ and $W \subseteq \mathbb{R}^d$ of $\bar{w}$ together with a constant $L > 0$ such that*

$$S(w) \cap X \subseteq S(w') + L|w - w'|\mathbb{B} \text{ for all } w, w' \in W, \tag{4.1}$$

where $\mathbb{B}$ denotes the unit ball in $\mathbb{R}^n$.

Remark. The continuity property in condition (ii) was originally called "pseudo Lipschitz continuity" in [1], and has also been referred to as the "Aubin property" [52] and "Aubin continuity" [21]. In [2], a sufficient condition for condition (ii) was given in terms of outer graphical derivatives, and Theorem 4.1 can be viewed as a kind of dual version of this result. Note that the existence of a local selection

[1] The first complete proof of this result appeared in [33], where there is also a thorough discussion of equivalent criteria and particular special cases.

$x(w) \in S(w)$ for $w \in W$ is guaranteed under this condition, and that in the case when S is actually a single-valued mapping $x : \mathbb{R}^d \to \mathbb{R}^n$, condition (ii) is the same as local Lipschitz continuity. The continuity property (ii) has been much studied for general multifunctions as well as for certain special cases. One example of the latter can be found in [8] where the multifunction S represents the solutions to certain parameterized generalized equations.

For a nice survey of coderivatives, their calculus, and applications, see [52] (for finite-dimensional spaces only) and [38] (for more general spaces).

To study uniqueness, yet another generalized derivative has proven to be useful. For any multifunction $S : \mathbb{R}^d \rightrightarrows \mathbb{R}^n$ and any pair $(\bar{w}, \bar{x})$ satisfying $\bar{x} \in S(\bar{w})$, the *strict derivative* $D_*S(\bar{w}|\bar{x}) : \mathbb{R}^d \rightrightarrows \mathbb{R}^n$ of S at $\bar{w}$ for $\bar{x}$ is defined for any $w' \in \mathbb{R}^d$ by

$$D_*S(\bar{w}|\bar{x})(w') = \left\{ x' \;\middle|\; \begin{array}{l} \exists w_\nu \to \bar{w}, x_\nu \in S(w_\nu), x_\nu \to \bar{x}, w'_\nu \to w', \tau_\nu \downarrow 0 \text{ with} \\ (\tilde{x}_\nu - x_\nu)/\tau_\nu \to x' \text{ for some } \tilde{x}_\nu \in S(w_\nu + \tau_\nu w'_\nu) \end{array} \right\}.$$

This derivative was used first in [53], but only for Lipschitz functions f. It was consequently called "Thibault's limit set" in [17], where it was used to prove an inverse function theorem for Lipschitz functions. The next theorem is a generalization of [17, Theorem 1.1] (use $S = f^{-1}$ below), and it shows that a nonsingular strict derivative characterizes the uniqueness of local selections in tandem with their Lipschitz continuity.

THEOREM 4.2 (see Theorem 1.3 of [19]). *For any multifunction $S : \mathbb{R}^d \rightrightarrows \mathbb{R}^n$ and any pair $(\bar{w}, \bar{x})$ satisfying $\bar{x} \in S(\bar{w})$, the following are equivalent:*

(i) *The strict derivative image set $D_*S(\bar{w}|\bar{x})(0)$ equals the trivial set $\{0\}$.*

(ii) *There exist neighborhoods $X \subseteq \mathbb{R}^n$ of $\bar{x}$ and $W \subseteq \mathbb{R}^d$ of $\bar{w}$ and a constant $L > 0$ such that there is at most one element $x(w)$ in the local image set $S(w) \cap X$ for $w \in W$ and it satisfies*

$$|x(w) - x(w')| \le L|w - w'| \quad \text{for} \quad w, w' \in W.$$

Remark. Notice that the gap between the nonsingularity condition (i) and the uniqueness property is identified precisely by the additional property of Lipschitz continuity in condition (ii) of Theorem 4.2.

For one important class of multifunctions, the existence of a Lipschitz local selection is also guaranteed by the nonsingularity condition (i) on the strict derivative. This class is studied in [21], where it is defined to consist of all multifunctions $S : \mathbb{R}^n \to \mathbb{R}^n$ whose graph near a base pair $(\bar{w}, \bar{x}) \in \operatorname{gph} S$ is the same set as $A \operatorname{gph} F$ for some Lipschitz continuous mapping $F : \mathbb{R}^n \to \mathbb{R}^n$ and a matrix $A \in \mathbb{R}^{2n \times 2n}$ of the form

$$\begin{bmatrix} 0 & A_2 \\ A_3 & A_4 \end{bmatrix}$$

for invertible $n \times n$ matrices A_2 and A_3. The graph of such a multifunction is said to be a *kernel inverting Lipschitzian manifold near* $(\bar{w}, \bar{x})$. For example, the graph of the inverse of any Lipschitzian mapping F is a kernel inverting Lipschitzian manifold under the matrix defined by $A_2 = A_3 = I$ and $A_4 = 0$. In particular, when this matrix is applied to the graph of the function $F(x) = 0$, the result is the graph of the multifunction F^{-1} which looks exactly like the derivative mapping (B) displayed in the introduction.

PROPOSITION 4.0.3 (see Theorem 2.1 of [21]). *For any multifunction $S : \mathbb{R}^n \to \mathbb{R}^n$ having $\operatorname{gph} S$ a kernel inverting Lipschitzian manifold near $(\bar{w}, \bar{x}) \in \operatorname{gph} S$, the following are equivalent:*

(i) *The strict derivative image set* $D_*S(\bar{w}|\bar{x})(0)$ *equals the trivial set* $\{0\}$.

(ii) *There exist neighborhoods* $X \subseteq \mathbb{R}^n$ *of* $\bar{x}$ *and* $W \subseteq \mathbb{R}^d$ *of* $\bar{w}$ *such that there exists a unique element* $x(w)$ *in* $S(w) \cap X$ *for* $w \in W$, *and in addition there is a constant* $L > 0$ *such that* $x(w)$ *is Lipschitz continuous with modulus* L *on* W.

Thus, for multifunctions satisfying the assumptions of Proposition 4.0.3, the nonsingularity condition on the strict derivative (i) guarantees not only the Lipschitz continuity and uniqueness of local selections (as in Theorem 4.2) but even guarantees their existence. We have already seen that for any Lipschitzian mapping, the multifunction F^{-1} is covered by Proposition 4.0.3, and we will see in a later section that Proposition 4.0.3 also covers KKT pair multifunctions. More generally, Proposition 4.0.3 applies to any solution mapping associated with a monotone generalized equation $0 \in F(x, w) + Q(x)$ for a monotone multifunction $Q : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$. The Lipschitz continuity (as in (ii) above) of this important particular solution mapping is also characterized by a coderivative nonsingularity condition [35].

As in the case of the outer graphical derivative, there are calculus rules for coderivatives and strict derivatives that help us apply the results in this section.

THEOREM 4.3 (see Corollary 4.4 of [37] and Exercise 10.43 of [52]). *For any multifunction* $M : \mathbb{R}^n \times \mathbb{R}^d \rightrightarrows \mathbb{R}^n$ *and any triple* $(\bar{x}, \bar{v}, \bar{w})$ *satisfying* $\bar{v} \in M(\bar{x}, \bar{w})$, *the following identities hold for the inverse multifunction* $S(v, w) := \{x | v \in M(x, w)\}$:

$$D^*S(\bar{v}, \bar{w}|\bar{x})(x') = \{(v', w') | (-x', w') \in D^*M(\bar{x}, \bar{w}|\bar{v})(-v')\},$$
$$D_*S(\bar{v}, \bar{w}|\bar{x})(v', w') = \{x' | v' \in D_*M(\bar{x}, \bar{w}|\bar{v})(x', w')\}.$$

For any multifunction $M : \mathbb{R}^d \rightrightarrows \mathbb{R}^n$, *any mapping* $F : \mathbb{R}^d \to \mathbb{R}^n$, *and any points* $\bar{w}$, $\bar{x}_0 \in M(\bar{w})$, *and* $\bar{x} := \bar{x}_0 + F(\bar{w})$, *if* F *is* $\mathcal{C}^1$ *near* $\bar{w}$, *then the sum* $M + F$ *satisfies*

$$D^*\big(M + F\big)(\bar{w}|\bar{x})(x) = D^*M(\bar{w}|\bar{x}_0)(x) + \nabla F(\bar{w})^* \cdot x,$$
$$D_*\big(M + F\big)(\bar{w}|\bar{x})(w) = D_*M(\bar{w}|\bar{x}_0)(w) + \nabla F(\bar{w}) \cdot w.$$

Proof. Both of the first identities follow easily from the fact that the graph of S is a permutation of the arguments of the graph of M. The third identity was established in [37, Corollary 4.4], and the last follows from [52, Exercise 10.43]. □

4.1. Existence and uniqueness of stationary points. In this section, we apply the results in the preceding section to the stationary point mapping to obtain the following sufficient conditions for the existence and uniqueness of stationary points.

PROPOSITION 4.1.1. *For the stationary point multifunction* SP_{tilt} (2.7) *and any pair* $(\bar{x}, \bar{w}) \in \mathbb{R}^{n+d}$ *satisfying* $0 \in \partial_x f(\bar{x}, \bar{w})$ *we have*

(Existence.) If the set $\operatorname{gph} SP_{tilt}$ *is locally closed at* $\big((0, \bar{w}), \bar{x}\big)$ *and* $(v', w') = 0$ *is the only vector in* $\mathbb{R}^{n+d}$ *satisfying*

$$(0, w') \in D^*\big(\partial_x f\big)(\bar{x}, \bar{w}|0)(-v'),$$

then there exist neighborhoods $X \subseteq \mathbb{R}^n$ *of* $\bar{x}$, $W \subseteq \mathbb{R}^d$ *of* $\bar{w}$, *and* $V \subseteq \mathbb{R}^n$ *of 0 such that the stationary point sets* $SP_{tilt}(v, w) \cap X$ *are nonempty for all pairs* $(v, w) \in V \times W$.

(Uniqueness.) If $x' = 0$ *is the only vector in* $\mathbb{R}^n$ *satisfying*

$$0 \in D_*\big(\partial_x f\big)(\bar{x}, \bar{w}|0)(x', 0),$$

then there exist neighborhoods $X \subseteq \mathbb{R}^n$ *of* $\bar{x}$, $W \subseteq \mathbb{R}^d$ *of* $\bar{w}$, *and* $V \subseteq \mathbb{R}^n$ *of 0 such that for all pairs* $(v, w) \in V \times W$, *there is at most one stationary point in the set* $SP_{tilt}(v, w) \cap X$.

Proof. This follows immediately from Theorems 4.1, 4.2, and 4.3. ∎

Remark. Notice that the conditions in Proposition 4.1.1 are also sufficient conditions for the existence and uniqueness of stationary points associated with the untilted problem (1.2), since these points are represented by the mapping SP (2.4) which satisfies $SP(w) = SP_{tilt}(0, w)$.

Recall from Theorem 4.2 that the gap created by the sufficient condition in Proposition 4.1.1 for the uniqueness of stationary points contains exactly the stationary point mappings that violate the Lipschitz continuity property.

When paired with the assumption that there exists a unique locally optimal solution to the unperturbed problem

$$\min\{f(x, \bar{w})\} \text{ over all } x \in \mathbb{R}^n,$$

the sufficient condition for uniqueness from Proposition 4.1.1 actually characterizes a much stronger property where, in particular, the stationary points are optimal solutions.

PROPOSITION 4.1.2 (see Theorem 5.2 of [19]). *For a lower semicontinuous function $f : \mathbb{R}^{n+d} \to \mathbb{R} \cup \{\infty\}$ and a pair $(\bar{x}, \bar{w}) \in \mathbb{R}^{n+w}$ satisfying $0 \in \partial_x f(\bar{x}, \bar{w})$ as well as the constraint qualification*

$$(y, 0) \in \partial^\infty f(\bar{w}, \bar{x}) \Rightarrow y = 0,$$

in terms of the set $\partial^\infty f(\bar{w}, \bar{x})$ of horizon subgradients of f at $(\bar{w}, \bar{x})$, the following are equivalent:

(i) *The pair of conditions hold that*

(a) *There exists a neighborhood $X \subseteq \mathbb{R}^n$ of $\bar{x}$ such that $\bar{x}$ is the only solution to the unperturbed problem*

$$\min\{f(x, \bar{w})\} \text{ over all } x \in X;$$

(b) $0 \in D_*\big(\partial_x f\big)(\bar{x}, \bar{w}|0)(x', 0)$ *only for* $x' = 0$,

(ii) *There exist neighborhoods $X \subseteq \mathbb{R}^n$ of $\bar{x}$, $V \subseteq \mathbb{R}^n$ of 0, and $W \subseteq \mathbb{R}^d$ of $\bar{w}$ such that for every parameter pair $(v, w) \in V \times W$, there is exactly one optimal solution $x(v, w)$ to*

$$\min\{f(x, w) - \langle v, x\rangle\} \text{ over all } x \in X$$

and, moreover, this function $x(v, w)$ is Lipschitz continuous on $V \times W$. In addition, the optimal solution $x(v, w)$ is also the unique stationary point $x(v, w) = SP_{tilt}(v, w) \cap X$ for $(v, w) \in V \times W$.

Remark. Insight into condition (i)(b) of Proposition 4.1.2 can be gained by considering the case when f is twice continuously differentiable. In this situation, condition (i)(b) reduces to the demand that the kernel of the Hessian $\nabla^2_{xx} f(\bar{x}, \bar{w})$ be trivial. It follows that the pair of conditions (i)(a) and (i)(b) in Proposition 4.1.2 is equivalent in this case to the positive-definiteness of the Hessian $\nabla^2_{xx} f(\bar{x}, \bar{w})$.

The constraint qualification used in Proposition 4.1.2 is a generalization of the constraint qualification (2.9) used in the definition of fully amenable functions with compatible parameterization (see [25]), so this proposition certainly covers those functions. We will see some similar characterizations in the next section, but they all cover smaller classes of functions than does Proposition 4.1.2.

4.2. Existence and uniqueness in the case of fully amenable functions. The calculus for the coderivative and the strict derivative is less developed than that for the outer graphical derivative, so it is at present more difficult to apply the results in section 4.1 directly to some particular cases, including the stationary point multifunction associated with a nonlinear program. Fortunately, for this important particular case and others, there is an alternative characterization of the existence and uniqueness of stationary points in terms of the outer graphical derivative at x for $v \in \partial_x f(x,w)$ of the multifunction $x \mapsto \partial_x f(x,w)$. The generalized derivative that results from this construction maps points in $\mathbb{R}^n$ to sets in $\mathbb{R}^n$ and is denoted by $D_{xx}f(x,w|v)$. If we "strengthen" this derivative by taking the outer graphical limit as $x \to \bar{x}$, $w \to \bar{w}$, and $v \to \bar{v}$ of the sequence of multifunctions $D_{xx}f(x,w|v)$, we obtain the *strong partial outer graphical derivative of f at $\bar{x}$ for $\bar{v} \in \partial_x f(\bar{x},\bar{w})$*, which we denote by $\tilde{D}^2_{xx}f(\bar{x},\bar{w}|\bar{v})$. Both $D_{xx}f(\bar{x},\bar{w}|\bar{v})$ and its strengthening $\tilde{D}^2_{xx}f(\bar{x},\bar{w}|\bar{v})$ are generalizations of the Hessian mapping $x' \mapsto \nabla_{xx}f(\bar{x},\bar{w}) \cdot x'$.

PROPOSITION 4.2.1 (see Proposition 3.4 of [20]). *If f is fully amenable in x at $\bar{x}$ with compatible parameterization in w at $\bar{w}$ and if $0 \in \partial_x f(\bar{x},\bar{w})$, then the following are equivalent:*

(i) *The pair of conditions holds that*

(a) *There exist neighborhoods $X \subseteq \mathbb{R}^n$ of $\bar{x}$, $V \subseteq \mathbb{R}^n$ of 0, and $W \subseteq \mathbb{R}^d$ of $\bar{w}$ such that for every parameter pair $(v,w) \in V \times W$, there is exactly one stationary point $x(v,w)$ in the set $SP_{tilt}(v,w) \cap X$;*

(b) *There is a constant $L > 0$ such that for any fixed $w \in W$, the (single-valued) mapping $v \mapsto SP_{tilt}(v,w) \cap X$ is monotone and Lipschitz continuous with modulus L on V.*

(ii) *The strong partial outer graphical derivative of f at $\bar{x}$ for 0 is positive-definite in the sense that*

$$v' \in \tilde{D}^2_{xx}f(\bar{x},\bar{w}|0)(x') \Rightarrow \langle v', x' \rangle > 0 \text{ unless } x' = 0.$$

Moreover, under either of these equivalent conditions and for any parameter pair $(v,w) \in V \times W$, the unique stationary point $x(v,w)$ in the set $SP_{tilt}(v,w) \cap X$ is also the unique optimal solution to

$$\min\{f(x,w) - \langle v, x \rangle\} \text{ over all } x \in X.$$

Remark. Notice that just as we have seen previously, even though Proposition 4.2.1 involves the tilted stationary point mapping SP_{tilt} (2.7), its condition (ii) is automatically a sufficient condition for existence and uniqueness of the stationary points associated with the untilted problem (1.2), since these points are represented by the mapping SP (2.4) which satisfies $SP(w) = SP_{tilt}(0,w)$. Moreover, the gap between the sufficient condition (ii) and the existence and uniqueness of stationary points is identified by Proposition 4.2.1 via the properties given in condition (i).

Notice finally that Proposition 4.2.1 also identifies conditions which guarantee that stationary points are locally optimal solutions.

If we combine Propositions 4.2.1 and 3.1.2, we get the following stability result for optimal solutions.

COROLLARY 4.2.1 (see Theorem 1.1 of [20]). *If f is fully amenable in x at $\bar{x}$ with compatible parameterization in w at $\bar{w}$ and if $0 \in \partial_x f(\bar{x},\bar{w})$, then under the second-order conditions* (ii) *of Proposition* 4.2.1 *and* (i) *of Proposition* 3.1.2, *there exist neighborhoods $X \subseteq \mathbb{R}^n$ of $\bar{x}$ and $W \in \mathbb{R}^d$ of $\bar{w}$ such that for each parameter*

$w \in W$, there exists a unique solution $x(w)$ to

$$\min\{f(x,w)\} \text{ over all } x \in X,$$

and this solution function $x(w)$ is calm in the sense of (3.1)*. Moreover, $x(w)$ is also the unique stationary point in the set $SP(w) \cap X$.*

Remark. We have stated both Proposition 4.2.1 and Corollary 4.2.1 in terms of the fully amenable essential objective functions we studied in section 2.2; however, they were developed in [20] for an even broader class of objective functions (though one not as broad as those covered by Proposition 4.1.2). A similar result was developed in [25] where second-order coderivative conditions were used instead, and the stronger property of Lipschitz continuity was obtained.

4.3. Existence and uniqueness of solutions to nonlinear programs. In the case of the nonlinear program (1.3), the results in the preceding section can be translated into more familiar terms using the formula for the outer graphical derivative of $\partial_x f$ from Proposition 2.3.1.

PROPOSITION 4.3.1 (see Theorem 4.2 of [20]). *For the nonlinear program* (1.3) *and its associated stationary point multifunction SP* (2.18)*, suppose the Mangasarian–Fromovitz constraint qualification* (2.17) *holds at $(\bar{x}, \bar{w})$ with $\bar{x} \in SP(\bar{w})$, and the second-order condition holds that:*

(i) *For any sequences $x \to \bar{x}$, $w \to \bar{w}$, and $v \to 0$ with $x \in C(w)$, and any convergent sequence of points $x' \in \cap_{i\in[1,m]}$ with $_{g_i(x,w)=0}\nabla_x g_i(x,w)^\perp$ with nonzero limit $\bar{x}'$, together with any corresponding convergent sequence of multipliers y which each maximizes the value $\langle x'\nabla^2_{xx}L(x,y,w)\cdot x'\rangle$ over all choices of multipliers $y \in N_K\big(G(x,w)\big)$ with $\nabla_x L(x,y,w) = v$, the Hessian at the limit multiplier $\bar{y}$ satisfies $\langle \bar{x}', \nabla^2_{xx}L(\bar{x},\bar{y},\bar{w})\bar{x}'\rangle > 0$.*

Then there exist neighborhoods $X \subseteq \mathbb{R}^n$ of $\bar{x}$ and $W \subseteq \mathbb{R}^d$ of $\bar{w}$ such that for every parameter $w \in W$, there is exactly one stationary point in the set $SP(w) \cap X$, and it is, moreover, the unique solution to the problem

$$\min\{g_0(x,w)\} \text{ over all } x \in C(w) \cap X.$$

Moreover, condition (i) *above is equivalent to the pair of conditions* (i)(a) *and* (i)(b) *from Proposition* 4.2.1 *applied to the associated stationary point mapping SP_{tilt}* (2.19)*.*

If in addition the second-order condition (i) *of Proposition* 3.2.1 *holds, then the function $x(w)$ is calm in the sense of* (3.1)*.*

Remark. Together, the second-order conditions (i) in Propositions 4.3.1 and 3.2.1 actually characterize a slightly stronger stability property than calmness (see [20]), so the gap between these conditions and the calmness property is known. Moreover, it is shown in [20] that the second-order conditions in Propositions 4.3.1 and 3.2.1 are both weaker than the general strong second-order condition (from [11]):

> For every multiplier $y \in Y(\bar{x},\bar{w},0,u)$ for some $u \in \mathbb{R}^d$, the Hessian $\nabla^2_{xx}L(\bar{x},\bar{y},\bar{w})$ is positive-definite on the subspace of vectors perpendicular to every $\nabla_x g_i(\bar{x},\bar{w})$ with $i \in I_1$ (2.32).

This general strong second-order condition when paired with the Mangasarian–Fromovitz constraint qualification has been known for some time (via [48] and [16]) to imply the calmness property, and the fact that the conditions in Proposition 4.3.1 are weaker than the general strong second-order condition means that this proposition is an improvement on this previous sufficient condition for calmness.

Example program revisited. We return once more to the nonlinear program introduced in section 2.3, but this time to see how the condition (i) in Proposition 4.3.1 works out for this example. Notice that in this case, the final inner product whose positivity needs to be verified satisfies

$$\langle \bar{x}', \nabla^2_{xx} L(\bar{x}, \bar{y}, \bar{w}) \bar{x}' \rangle = 2(\bar{x}'_1)^2 + 2(1 - \bar{y}_1)(\bar{x}'_2)^2, \tag{4.2}$$

and that $\bar{y}_1 \in [0, 1]$. We consider only nonzero limit points $\bar{x}'$, and notice that in the case when $\bar{x}'_1 \neq 0$, the inner product (4.2) is positive no matter what the choice of multiplier since $\bar{y}_1 \in [0, 1]$. On the other hand, if $\bar{x}'_1 = 0$, then $\bar{x}'$ being nonzero implies that $\bar{x}'_2 \neq 0$, which means that we can assume $x'_2 \neq 0$ for the whole sequence of $x' \to \bar{x}'$. In this case, since the inner product to be maximized by the multiplier y is given by

$$\langle x', \nabla^2_{xx} L(x, y, w) x' \rangle = 2(x'_1)^2 + 2(1 - y_1)(x'_2)^2$$

we can conclude that $y_1 = 0$ for each multiplier in the sequence, so the limit multiplier $\bar{y}$ satisfies $\bar{y}_1 = 0$. It follows that in this case too, the inner product (4.2) is positive, so this example program satisfies the conditions of Proposition 4.3.1, and as a result has exactly one stationary point (and this stationary point also happens to be a locally optimal solution) for all values of w near $\bar{w} = 0$.

It is interesting to note that the general strong second-order condition does not hold for this example program (for the multiplier $\bar{y} = (1, 0)$, the Hessian $\nabla^2_{xx} L(\bar{x}, \bar{y}, \bar{w})$ is only positive semidefinite on the appropriate subspace), so this optimization problem falls in the gap between that sufficient condition and the stability properties promised by Proposition 4.3.1.

Remark. Of course it would be nice if the condition (i) from Proposition 4.3.1 could be refined in such a way as to depend only on information at the base point like the general strong second-order condition and not on the limits of nearby parameter values. In particular, such a refinement would likely make condition (i) less daunting to verify in many situations. However, such refinements are generally not possible without creating a gap and sacrificing the characterization. This fact can be seen by considering two different parameterizations of any well-behaved unperturbed problem: One parameterization is trivial, with each perturbed problem being the same as the unperturbed model, and the other parameterization is anything that creates a lack of the kind of stability covered by Proposition 4.3.1. For instance, consider the objective function $g_0(x_1, x_2) := x_1^2/2$ with parameterized constraint set

$$C(w) := \{(x_1, x_2) \in \mathbb{R}^2 : g_1(x, w) := x_2 - w^{2k} \leq 0 \text{ and } g_2(x, w) := -x_2 - w^{2k} \leq 0\}$$

for any positive integer k. For the unperturbed problem with $\bar{w} = 0$, the constraint set reduces to $C(0) = \{(x_1, x_2) \in \mathbb{R}^2 : x_2 = 0\}$, and the minimization of g_0 over $C(0)$ has a unique solution at $\bar{x} = (0, 0)$. Of course the trivially parameterized problem also has this property, so it exhibits the kind of stability covered by Proposition 4.3.1, and it even satisfies the general strong second-order condition. However, the minimization of g_0 over $C(w)$ for $w \neq 0$ has multiple solutions, so it fails to be stable at $\bar{w}$ in the sense of Proposition 4.3.1. In this case, only the derivatives of order $2k$ with respect to w of g_1 and g_2 evaluated at the base point $(\bar{x}, \bar{w})$ distinguish the two different parameterizations. Since k is any positive integer, no test using only derivatives evaluated at the base point will be able to distinguish the stability discrepancies illustrated in this example.

4.4. Existence and uniqueness of KKT pairs. We again consider the KKT pairs associated with the optimization model from section 2.4

$$\min\{g_0(x,w) + g\big(G(x,w)+v_2\big) - \langle v_1, x\rangle\} \text{ over } x \in \mathbb{R}^n.$$

Unlike the multifunctions studied in the preceding two sections, the particular structure of the KKT pair multifunction allows us to directly apply the fundamental characterizations in terms of the coderivative and strict derivative from section 4. In fact, for the KKT pair multifunction, the different characterizations from section 4 are both equivalent to each other and to the existence of a unique local selection that is Lipschitz continuous.

PROPOSITION 4.4.1. *If the mapping $\nabla_x L$ is $\mathcal{C}^1$ at $(\bar{x},\bar{y},\bar{w})$ and $(\bar{x},\bar{y}) \in KKT(0,\bar{w})$ for the KKT pair multifunction* (2.24)*, then the following are equivalent:*

(i) *The pair $(v',w') = (0,0)$ is the only solution in $\mathbb{R}^{n+m+d}$ to the system*

$$\begin{aligned}
\nabla_{xx}L(\bar{x},\bar{y},\bar{w})^* \cdot v_1' &= \nabla_x G(\bar{x},\bar{w})^* \cdot v_2',\\
\nabla_{xy}L(\bar{x},\bar{y},\bar{w})^* \cdot v_1' &\in D^*\Big(\big(\partial g\big)^{-1}\Big)\big(\bar{y}|G(\bar{x},\bar{w})\big)(-v_2'),\\
w' + \nabla_{xw}L(\bar{x},\bar{y},\bar{w})^* \cdot v_1' &= \nabla_w G(\bar{x},\bar{w})^* \cdot v_2'.
\end{aligned}$$

(ii) *The pair $(x',y') = (0,0)$ is the only solution in $\mathbb{R}^{n+m}$ to the system*

$$\begin{aligned}
0 &= \nabla^2_{xx}L(\bar{x},\bar{y},\bar{w})\cdot x' + \sum_{i=1}^m y_i' \nabla_x g_i(\bar{x},\bar{w}),\\
\nabla_x G(\bar{x},\bar{w})\cdot x' &\in D_*\Big(\big(\partial g\big)^{-1}\Big)\big(\bar{y}|G(\bar{x},\bar{w})\big)(y').
\end{aligned}$$

(iii) *There exist neighborhoods $X \times Y$ of $(\bar{x},\bar{y}) \in \mathbb{R}^{n+m}$ and $V \times W$ of $(0,\bar{w})$ in $\mathbb{R}^{n+d}$ such that for all pairs $(v,w) \in V \times W$, there is exactly one KKT pair $\big(x(v,w),y(v,w)\big)$ in the set $KKT(v,w) \cap (X \times Y)$, and, moreover, this KKT pair function is Lipschitz continuous on $V \times W$.*

Moreover, under any of these equivalent conditions, if the subgradient multifunction ∂g is protodifferentiable at $G(\bar{x},\bar{w})$ for $\bar{y} \in \partial g\big(G(\bar{x},\bar{w})\big)$, then the KKT pair $\big(x(v,w),y(v,w)\big)$ is B-differentiable at $(0,\bar{w})$.

Proof. The equivalence between (i) and (iii) follows from Theorem 4.1 and [9, Theorem 3] since the coderivative image set $D^*\big(KKT\big)(0,\bar{w}|\bar{x},\bar{y})(x',y')$ consists of all pairs $(v',w') \in \mathbb{R}^{n+d}$ that satisfy the system

$$\begin{aligned}
-x' &= -\nabla_{xx}L(\bar{x},\bar{y},\bar{w})^* \cdot v_1' + \nabla_x G(\bar{x},\bar{w})^* \cdot v_2',\\
-y' &\in -\nabla_{xy}L(\bar{x},\bar{y},\bar{w})^* \cdot v_1' + D^*\Big(\big(\partial g\big)^{-1}\Big)\big(\bar{y}|G(\bar{x},\bar{w})\big)(-v_2'),\\
w' &= -\nabla_{xw}L(\bar{x},\bar{y},\bar{w})^* \cdot v_1' + \nabla_w G(\bar{x},\bar{w})^* \cdot v_2'.
\end{aligned}$$

To prove the equivalence between (ii) and (iii), we use a certain linearization $L_{KKT} : \mathbb{R}^n \rightrightarrows \mathbb{R}^{n+m}$ of the KKT pair multifunction:

$$L_{KKT}(v) := \big\{(x,y) : v \in H(\bar{x},\bar{y},\bar{w}) + \nabla_{xy}H(\bar{x},\bar{y},\bar{w})\cdot\big((x,y)-(\bar{x},\bar{y})\big) + N(x,y)\big\},$$

where the mapping $H : \mathbb{R}^{n+m+d} \to \mathbb{R}^{n+m}$ is defined by

$$H(x,y,w) := \Big(\nabla_x L(x,y,w), -G(x,w)\Big)$$

and the multifunction $N : \mathbb{R}^{n+m} \rightrightarrows \mathbb{R}^{n+m}$ is defined by

$$N(x,y) := \Big(\{0\}^n \times \big(\partial g\big)^{-1}(y)\Big).$$

Notice that the multifunction L_{KKT} is the same as the KKT pair multifunction except that the former uses a linearization of H instead of H itself. Moreover, the graph of L_{KKT} is a kernel inverting Lipschitzian manifold (see [21]) so it is covered by our Proposition 4.0.3. In [21, Theorem 4.1] and [9, Propositions 1 and 2] it was shown that property (iii) above for the KKT pair multifunction is equivalent to the same property for the linearization L_{KKT}, and in [21, Theorem 4.1] it was shown that the nonsingularity condition (ii) is equivalent to the same condition applied to the linearization multifunction L_{KKT}. This means that the characterization recorded in Proposition 4.0.3 also holds for the KKT pair multifunction, so the result here follows from the calculus rules in Theorem 4.3, which allow us to deduce that the strict derivative image set $D_*\big(KKT\big)(0,\bar w|\bar x,\bar y)(v',w')$ consists of all pairs $(x',y') \in \mathbb{R}^{n+m}$ that satisfy the system

$$\begin{aligned} v_1' &= \nabla_{x(x,y,w)} L(\bar x,\bar y,\bar w)\cdot(x',y',w'),\\ v_2' &\in -\nabla G(\bar x,\bar w)\cdot(x',w') + D_*\Big(\big(\partial g\big)^{-1}\Big)\big(\bar y|G(\bar x,\bar w)\big)(y'). \end{aligned}$$

The B-differentiability follows immediately from Proposition 2.4.2. □

Remark. Notice the similarity between condition (ii) here and condition (i) from Proposition 3.3.1 characterizing the calmness of local KKT pair selections: The only difference is that the generalized derivative used in Proposition 3.3.1 is the outer graphical derivative. This reflects the fact that these arose from similar nonsingularity conditions which were derived in terms of the different generalized derivatives.

4.5. Existence and uniqueness of KKT pairs for nonlinear programs. For nonlinear programs with canonical perturbations (2.30), the results of the preceding section can be translated into even more specific terms by working out formulas for either the coderivative or strict derivative of the inverse of the normal cone multifunction N_K (2.13). We again use the sets of indices (2.32).

PROPOSITION 4.5.1. *For the canonically perturbed nonlinear program* (2.30) *and its associated KKT pair multifunction* (2.33)*, if the Mangasarian–Fromovitz constraint qualification* (2.17) *holds at* $(\bar x,\bar w)$ *with* $(\bar x,\bar y) \in KKT(0,\bar w)$*, then the following are equivalent:*

(i) *The pair of conditions hold that*

(a) *The vectors* $\nabla_x g_i(\bar x,\bar w)$ *for* $i \in I_1 \cup I_2$ *are linearly independent;*

(b) *For each partition of* $\{1,\dots,m\}$ *into index sets* I_1', I_2', I_3' *with* $I_1 \subseteq I_1' \subseteq I_1 \cup I_2$ *and* $I_3 \subseteq I_3' \subseteq I_3 \cup I_2$*, the cone* $K(I_1',I_2') \subseteq \mathbb{R}^n$ *consisting of all vectors* x' *satisfying*

$$\langle \nabla_x g_i(\bar x,\bar w), x'\rangle \begin{cases} = 0 & \text{for } i \in I_1',\\ \le 0 & \text{for } i \in I_2', \end{cases}$$

should be such that

$$x' \in K(I_1',I_2') \text{ and } \nabla^2_{xx} L(\bar x,\bar y,\bar w)\cdot x' \in K(I_1',I_2')^* \quad \Rightarrow \quad x' = 0.$$

(ii) *The pair* $(x',y') = (0,0)$ *is the only solution in* $\mathbb{R}^{n+m}$ *to the system*

$$0 = \nabla^2_{xx} L(\bar x,\bar y,\bar w)\cdot x' + \sum_{i=1}^m y_i' \nabla_x g_i(\bar x,\bar w),$$

$$\langle \nabla_x g_i(\bar{x}, \bar{w}), x' \rangle = 0 \text{ for } i \in I_1,$$
$$y_i' \langle \nabla_x g_i(\bar{x}, \bar{w}), x' \rangle \geq 0 \text{ for } i \in I_2,$$
$$y_i' = 0 \text{ for } i \in I_3.$$

(iii) *There exist neighborhoods $X \times Y$ of $(\bar{x}, \bar{y}) \in \mathbb{R}^{n+m}$ and $V \times W$ of $(0, \bar{w})$ in $\mathbb{R}^{n+d}$ such that for all pairs $(v, w) \in V \times W$, there is exactly one KKT pair $\big(x(v,w), y(v,w)\big)$ in the set $KKT(v,w) \cap (X \times Y)$, and, moreover, this KKT pair function is Lipschitz continuous on $V \times W$.*

Moreover, under any of these equivalent conditions, the KKT pair $\big(x(v,w), y(v,w)\big)$ is B-differentiable at $(0, \bar{w})$.

Proof. The equivalence between (i) and (iii) was established in [9, Theorem 5], where it was essentially shown that condition (i) above is equivalent to condition (i) from our Proposition 4.4.1. The equivalence between (ii) and (iii) follows directly from our Proposition 4.4.1 since the strict derivative of the inverse of N_K has $z' \in D_*((N_K)^{-1})(\bar{y}|G(\bar{x}, \bar{w}))(y')$ if and only if

$$y_i' = 0 \text{ for } i \in I_1,$$
$$y_i' \, z_i' \geq 0 \text{ for } i \in I_2,$$
$$z_i' = 0 \text{ for } i \in I_3,$$

(see [23]) and since the assumption that $(\bar{x}, \bar{y})$ is a KKT pair for $(0, \bar{w})$ implies that for $i \in [1, s]$, $g_i(\bar{x}, \bar{w}) \leq 0$ and $\bar{y}$ satisfies $\bar{y}_i \, g_i(\bar{x}, \bar{w}) = 0$ with $\bar{y}_i \geq 0$. □

Notice that the assumed satisfaction of the Mangasarian–Fromovitz constraint qualification is guaranteed by the linear independence condition (i)(a). It is an easy exercise to show that any pair satisfying the system in condition (i) of Proposition 3.4.1 also satisfies the system in condition (ii) above. This is consistent with the fact that condition (i) of Proposition 3.4.1 characterizes the calmness of local selections of KKT pairs, which is a weaker property than (iii) of Proposition 4.5.1. Notice that Proposition 4.5.1 does not address the issue of the primal component of the KKT pair being a locally optimal solution. For this, we need a positive-definiteness assumption.

PROPOSITION 4.5.2 (see Theorem 6 of [9]). *Under the assumptions of Proposition* 4.5.1*, the following are equivalent:*

(i) *There exist neighborhoods $X \times Y$ of $(\bar{x}, \bar{y}) \in \mathbb{R}^{n+m}$ and $V \times W$ of $(0, \bar{w})$ in $\mathbb{R}^{n+d}$ such that for all pairs $(v, w) \in V \times W$, there is exactly one KKT pair $\big(x(v,w), y(v,w)\big)$ in the set $KKT(v,w) \cap (X \times Y)$, this KKT pair function is Lipschitz continuous on $V \times W$, and, moreover, $x(v,w)$ is a locally optimal solution to the canonically perturbed nonlinear program* (2.30).

(ii) *The constraint gradients $\nabla_x g_i(\bar{x}, \bar{w})$ for $i \in I_1 \cup I_2$ are linearly independent and the strong second-order sufficient condition for local optimality holds that $\nabla^2_{xx} L(\bar{x}, \bar{y}, \bar{w})$ is positive-definite on the subspace of vectors perpendicular to all constraint gradients $\nabla_x g_i(\bar{x}, \bar{w})$ for $i \in I_1$.*

Moreover, under either of these equivalent conditions, the KKT pair $(x(v,w), y(v,w))$ is B-differentiable at $(0, \bar{w})$.

Remark. Notice that under the linear independence assumption in (ii), there is only one multiplier $\bar{y}$ which can be paired with $\bar{x}$ as a KKT pair for the parameters $(0, \bar{w})$ (which condition is equivalent to the strict Mangasarian–Fromovitz constraint qualification). Under these circumstances, the strong second-order sufficient condition here is equivalent to Robinson's general strong second-order condition (see section 4.3), so condition (ii) of Proposition 4.5.2 is stronger than the assumptions in Proposition 4.3.1. This is consistent with the fact that the properties identified in condition

(i) of Proposition 4.5.2 are stronger than the calmness of the unique local stationary point selection guaranteed by Proposition 4.3.1 (recall that under the Mangasarian–Fromovitz constraint qualification, each of the stationary points associated with the nonlinear program (1.4) can be paired with a multiplier to form a KKT pair).

5. Conclusions and related work. Each of the main three sections in the body of this paper follow the same basic pattern: The most general principles are established first (e.g., Theorems 2.1, 3.1, 4.1, and 4.2), and the rest of the work is computing the appropriate generalized derivatives in order to translate the general principles into verifiable conditions in the language of a particular optimization model (e.g., nonlinear programming). The results for stationary points and stationary point-multiplier pairs mirror each other because they are based on the same general principles; however, they are different too because the generalized derivatives work out differently in each of these situations. One common trick used in the sections involving stationary points is to characterize the sensitivity properties of the tilted version of the optimization problem (2.6) first and then to translate these characterizations into sufficient conditions for the untilted case. A similar approach is used in sections involving KKT stationary point-multiplier pairs, where the presence of canonical perturbations is exploited. We can always recover results about the untilted (or uncanonically perturbed) model by setting the new parameters equal to zero, and, moreover, the gap created by the resulting sufficient condition is precisely identified in terms of the augmented model.

Others who have used outer graphical derivatives to obtain results about the calmness of local selections include Qiu and Magnanti [44], [18], and Klatte and Kummer [15]. A good survey of sensitivity results in nonlinear programming obtained via this and similar methods can be found in [10], and a survey of other results in nonlinear programming obtained by these and more traditional approaches can be found in [11]. There has also been some interest recently in sensitivity results for nonlinear programs under relaxed differentiability assumptions (see [14], for example). Of course the generality of our approach means that our results could be applied in such cases by computing the appropriate generalized derivatives under the new assumptions.

A slightly different but complementary approach to studying the sensitivity properties of inverse multifunctions is based on the work of Robinson [46] and [47]. Instead of using the outer graphical derivative, this work relies on a different kind of linearization of the inverse multifunction (we actually use this linearization in our proof of Proposition 4.4.1). For many important cases of interest (including the KKT pair multifunction), the local single-valuedness and Lipschitz continuity of the linearization implies the same properties for the original inverse multifunction. The focus of this approach then is on establishing conditions under which the linearization is locally single-valued and Lipschitz continuous, a property which Robinson called "strong regularity." This approach has also yielded B-differentiability results for certain inverse multifunctions when they are single-valued [50], as well as results about when stationary points for optimization problems are optimal solutions.

In Theorem 4.1, we recorded Mordukhovich's characterization of "Aubin continuity" in terms of coderivatives [32, Theorem 5.4], but since our focus in the present paper is on the continuity property of selection calmness, we did not exploit this fundamental characterization more. Coderivatives and the characterization [32, Theorem 5.4] have been used effectively to develop conditions guaranteeing Lipschitzian properties of optimal solutions, stationary points, and related objects (see, for exam-

ple, [32], [34], [35], [36], [38], [52], and [25]). Many of these results depend on the coderivative calculus developed in [37].

A good example of a fundamentally different approach to sensitivity analysis is surveyed in [5], where quite different sensitivity properties are studied. These authors start from a generalized nonlinear program, and give sufficient conditions for directional differentiability and calmness of nearly optimal solutions along fixed parameter directions.

One very satisfying aspect of our approach is that even when applied to the well-studied case of nonlinear programming, our fundamental rules have generated new conditions for previously sought conditions. For example, Propositions 3.2.1 and 3.2.2 provide unprecedented conditions guaranteeing the calmness of local stationary point selections. Proposition 4.3.1 is also unprecedented, and improves the combined results of [48] and [16] giving the existence, uniqueness, and calmness of local solutions. Related results giving true Lipschitz continuity include a theorem of Robinson's [47] which stated that the linear independence of the gradients of all the binding constraints, together with the general strong second-order condition, were enough to ensure the existence, uniqueness, and Lipschitz continuity of locally optimal solutions. A long-standing question in this area (posed formally by Robinson in [48]) is whether a weaker constraint qualification can replace linear independence and still ensure this kind of solution stability. One answer to this question was provided in Liu [30] and Ralph and Dempe [45] where a "constant rank condition" combined with the Mangasarian–Fromovitz constraint qualification and the general strong second-order condition were shown to ensure the existence, uniqueness, and Lipschitz continuity of locally optimal solutions to the nonlinear program (1.3). Proposition 4.3.1 shows that the Mangasarian–Fromovitz constraint qualification together with a pair of weaker second-order conditions suffice to ensure the slightly weaker stability property of calmness. However, Robinson [48] provided an example to show that the assumptions in Proposition 4.3.1 are not enough to ensure the local Lipschitz continuity of locally optimal solutions to (1.3).

Finally, the success of our approach is not limited to the nonlinear programs on which we have focused our attention here. As another example, the same basic approach has been applied successfully to carry out a reasonably complete sensitivity analysis of solutions to parameterized nonlinear complementarity problems [22].

Acknowledgments. The author is grateful to Asen Dontchev and Boris Mordukhovich for their comments during the preparation of this paper.

REFERENCES

[1] J. P. Aubin, *Lipschitz behavior of solutions to convex minimization problems*, Math. Oper. Res., 9 (1984), pp. 87–111.
[2] J. P. Aubin and I. Ekeland, *Applied Nonlinear Analysis*, Wiley-Interscience, New York, 1984.
[3] J. P. Aubin and H. Frankowska, *Set-valued Analysis*, Birkhäuser, Boston, 1990.
[4] J. F. Bonnans, *Local analysis of Newton-type methods for variational inequalities and nonlinear programming*, Appl. Math. Optim., 29 (1994), pp. 161–186.
[5] J. F. Bonnans and A. Shapiro, *Optimization problems with perturbations: A guided tour*, SIAM Rev., 40 (1998), pp. 228–264.
[6] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.
[7] A. L. Dontchev, *Characterizations of Lipschitz stability in optimization*, in Recent Developments in Well-Posed Problems, R. Lucchetti and J. Revalski, eds., Kluwer, Dordrecht, The Netherlands, 1995, pp. 95–115.
[8] A. L. Dontchev and W. W. Hager, *Implicit functions, Lipschitz maps, and stability in optimization*, Math. Oper. Res., 19 (1994), pp. 297–326.

[9] A. L. Dontchev and R. T. Rockafellar, *Characterizations of strong regularity for variational inequalities over polyhedral convex sets*, SIAM J. Optim., 6 (1996), pp. 1087–1105.
[10] A. L. Dontchev and R. T. Rockafellar, *Characterizations of Lipschitzian stability in nonlinear programming*, in Mathematical Programming with Data Perturbations, Lecture Notes in Pure and Appl. Math. 195, Dekker, New York, 1998, pp. 65–82.
[11] A. V. Fiacco and J. Kyparisis, *Sensitivity analysis in nonlinear programming under second order assumptions*, in Systems and Optimization (Enschede, 1984), Lecture Notes in Control and Inform. Sci. 66, Springer, New York, 1985, pp. 74–97.
[12] M. S. Gowda and J.-S. Pang, *Stability analysis of variational inequalities and nonlinear complementarity problems, via the mixed linear complementarity problem and degree theory*, Math. Oper. Res., 19 (1994), pp. 831–879.
[13] A. J. King and R. T. Rockafellar, *Sensitivity analysis for nonsmooth generalized equations*, Math. Programming, 55 (1992), pp. 193–212.
[14] D. Klatte, *Nonlinear optimization under data perturbations*, in Modern Methods of Optimization, W. Krabs and J. Zowe, eds., Springer-Verlag, New York, 1992, pp. 204–235.
[15] D. Klatte and B. Kummer, *Generalized Kojima-functions and Lipschitz stability of critical points*, Comput. Optim. Appl., 13 (1999), pp. 61–85.
[16] M. Kojima, *Strongly stable stationary solutions in nonlinear programs*, in Analysis and Computation of Fixed Points, S. M. Robinson, ed., Academic Press, New York, 1980, pp. 93–138.
[17] B. Kummer, *Lipschitzian inverse functions, directional derivatives, and applications in $\mathcal{C}^{1,1}$ optimization*, J. Optim. Theory Appl., 70 (1991), pp. 561–582.
[18] J. Kyparisis, *Parametric variational inequalities with multivalued solution sets*, Math. Oper. Res., 17 (1992), pp. 341–364.
[19] A. B. Levy, *Nonsingularity conditions for multifunctions*, Set-Valued Anal., 7 (1999), pp. 89–99.
[20] A. B. Levy, *Calm minima in parameterized finite-dimensional optimization*, SIAM J. Optim., 11 (2000), pp. 160–178.
[21] A. B. Levy, *Lipschitzian multifunctions and a Lipschitzian inverse mapping theorem*, Math. Oper. Res., to appear.
[22] A. B. Levy, *Stability of solutions to parameterized nonlinear complementarity problems*, Math. Program., 85 (1999), pp. 397–406.
[23] A. B. Levy, *Implicit multifunction theorems for the sensitivity analysis of variational conditions*, Math. Programming, 74 (1996), pp. 333–350.
[24] A. B. Levy, *Errata in "Implicit multifunction theorems for the sensitivity analysis of variational conditions,"* Math. Programming, 86 (1999), pp. 439–441.
[25] A. B. Levy, R. A. Poliquin, and R. T. Rockafellar, *Stability of locally optimal solutions*, SIAM J. Optim., 10 (2000), pp. 580–604.
[26] A. B. Levy and R. T. Rockafellar, *Sensitivity analysis of solutions to generalized equations*, Trans. Amer. Math. Soc., 345 (1994), pp. 661–671.
[27] A. B. Levy and R. T. Rockafellar, *Sensitivity of solutions in nonlinear programming problems with nonunique multipliers*, in Recent Advances in Nonsmooth Optimization, D.-Z. Du, L. Qi, and R. S. Womersley, eds., World Scientific, River Edge, NJ, 1995, pp. 215–223.
[28] A. B. Levy and R. T. Rockafellar, *Variational conditions and the proto-differentiation of partial subgradient mappings*, Nonlinear Anal., 26 (1995), pp. 1951–1964.
[29] A. B. Levy and R. T. Rockafellar, *Proto-derivatives and the geometry of solution mappings in nonlinear programming*, in Nonlinear Optimization and Applications, Plenum, New York, 1996, pp. 343–365.
[30] J. Liu, *Sensitivity analysis in nonlinear programs and variational inequalities via continuous selections*, SIAM J. Control Optim., 33 (1995), pp. 1040–1060.
[31] B. Mordukhovich, *Approximation Methods in Problems of Optimization and Control*, Nauka, Moscow, Russia, 1988.
[32] B. Mordukhovich, *Sensitivity analysis in nonsmooth optimization*, in Theoretical Aspects of Industrial Design, SIAM Proceedings in Applied Mathematics 58, D. A. Field and V. Komkov, eds., SIAM, Philadelphia, 1992, pp. 32–46.
[33] B. Mordukhovich, *Complete characterization of openness, metric regularity, and Lipschitzian properties of multifunctions*, Trans. Amer. Math. Soc., 340 (1993), pp. 1–35.
[34] B. Mordukhovich, *Lipschitzian stability of constraint systems and generalized equations*, Nonlinear Anal., 22 (1994), pp. 173–206.
[35] B. Mordukhovich, *Stability theory for parametric generalized equations and variational inequalities via nonsmooth analysis*, Trans. Amer. Math. Soc., 343 (1994), pp. 609–657.

[36] B. Mordukhovich, *Sensitivity analysis for constraint and variational systems by means of set-valued differentiation*, Optimization, 31 (1994), pp. 13–46.
[37] B. Mordukhovich, *Generalized differential calculus for nonsmooth and set-valued mappings*, J. Math. Anal. Appl., 183 (1994), pp. 250–288.
[38] B. Mordukhovich, *Coderivatives of set-valued mappings: Calculus and applications*, Nonlinear Anal., 30 (1997), pp. 3059–3070.
[39] J.-S. Pang, *A degree-theoretic approach to parametric nonsmooth equations with multivalued perturbed solution sets*, Math. Programming, 62 (1993), pp. 359–383.
[40] J.-S. Pang, *Necessary and sufficient conditions for solution stability of parametric nonsmooth equations*, in Recent Advances in Nonsmooth Optimization, D.-Z. Du, L. Qi, and R. S. Womersley, eds., World Scientific, River Edge, NJ, 1995, pp. 261–288.
[41] R. A. Poliquin and R. T. Rockafellar, *A calculus of epi-derivatives applicable to optimization*, Canad. J. Math., 45 (1993), pp. 879–896.
[42] R. A. Poliquin and R. T. Rockafellar, *Proto-derivative formulas for basic subgradient mappings in mathematical programming*, Set-Valued Anal., 2 (1994), pp. 275–290.
[43] R. A. Poliquin and R. T. Rockafellar, *Proto-derivatives of partial subgradient mappings*, J. Convex Anal., 4 (1997), pp. 221–234.
[44] Y. Qiu and T. L. Magnanti, *Sensitivity analysis for variational inequalities*, Math. Oper. Res., 17 (1992), pp. 61–76.
[45] D. Ralph and S. Dempe, *Directional derivatives of the solutions of a parametric nonlinear program*, Math. Programming, 70 (1995), pp. 159–172.
[46] S. M. Robinson, *Generalized equations and their solutions, part* I: *Basic theory*, Math. Programming Stud., 10 (1979), pp. 128–141.
[47] S. M. Robinson, *Strongly regular generalized equations*, Math. Oper. Res., 5 (1980), pp. 43–62.
[48] S. M. Robinson, *Generalized equations and their solutions, part* II: *Applications to nonlinear programming*, Math. Programming Stud., 19 (1982), pp. 200–221.
[49] S. M. Robinson, *Local structure of feasible sets in nonlinear programming, part* III: *Stability and sensitivity*, Math. Programming Stud., 30 (1987), pp. 45–66.
[50] S. M. Robinson, *An implicit-function theorem for a class of nonsmooth functions*, Math. Oper. Res., 16 (1991), pp. 292–309.
[51] R. T. Rockafellar, *Perturbation of generalized Kuhn-Tucker points in finite-dimensional optimization*, in Nonsmooth Analysis and Related Topics, F. H. Clarke et al., eds., Plenum Press, New York, 1989, pp. 393–402.
[52] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, New York, 1998.
[53] L. Thibault, *Subdifferentials of compactly Lipschitzian vector-valued functions*, Ann. Mat. Pura Appl. (4), 125 (1980), pp. 157–192.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 1, pp. 39–53

A REMARK ON NULL EXACT CONTROLLABILITY OF THE HEAT EQUATION*

XU ZHANG†

Abstract. It is well known that the heat equation $u_t - \Delta u = f\chi_\omega$ in $(0,T)\times\Omega$ with homogeneous Dirichlet boundary conditions is null exactly controllable for any $T > 0$ and any open nonempty subset ω of Ω. In this note we show that this property may be obtained as a singular limit of the exact controllability properties of singularly perturbed damped wave equations with a changing controller.

Key words. controllability, singular perturbation, heat equation, wave equation

AMS subject classifications. 93B05, 93B07, 93D15

PII. S0363012900371691

1. Introduction and main result.

Let us consider the following controlled heat equation:

$$\begin{cases} u_t - \Delta u = f\chi_\omega & \text{in} \quad Q, \\ u = 0 & \text{on} \quad \Sigma, \\ u(0) = u^0 & \text{in} \quad \Omega. \end{cases} \tag{1.1}$$

In (1.1), $Q \triangleq (0,T)\times\Omega$ and $\Sigma \triangleq (0,T)\times\Gamma$, where $T > 0$ and Ω is a bounded domain of $\mathbb{R}^n$ with C^∞ boundary $\Gamma \triangleq \partial\Omega$, $u = u(t,x)$ is the *state*, $f = f(t,x)$ is the *control*, and χ_ω denotes the characteristic function of the open nonempty subset ω of Ω where the control is supported.

It is well known (see [4] and [6]) that (1.1) is null exactly controllable for any given $T > 0$ and any nonempty open subset ω of Ω; i.e., for any given $u^0 \in L^2(\Omega)$, one can find a control $f \in L^2((0,T)\times\omega)$ such that the weak solution $u(\cdot) \in C([0,T]; L^2(\Omega)) \cap C((0,T]; H_0^1(\Omega))$ of (1.1) satisfies

$$u(T) = 0. \tag{1.2}$$

The corresponding f is called a null-control for (1.1) with initial state u^0.

In this note we will show that the above property of (1.1) may be obtained as a singular limit of the exact controllability properties of the following damped, singularly perturbed wave equation with a *changing* controller ω_ε:

$$\begin{cases} \varepsilon u_{\varepsilon,tt} - \Delta u_\varepsilon + u_{\varepsilon,t} = f_\varepsilon \chi_{\omega_\varepsilon} & \text{in} \quad Q, \\ u_\varepsilon = 0 & \text{on} \quad \Sigma, \\ u_\varepsilon(0) = u^0,\ u_{\varepsilon,t}(0) = u^1 & \text{in} \quad \Omega. \end{cases} \tag{1.3}$$

In (1.3), $\varepsilon > 0$ is a *small* parameter (which is devoted to tend to zero), $(u_\varepsilon, u_{\varepsilon,t}) = \big(u_\varepsilon(t,x), u_{\varepsilon,t}(t,x)\big)$ is the *state*, $f_\varepsilon = f_\varepsilon(t,x)$ is the *control*, and ω_ε is an open nonempty

*Received by the editors May 1, 2000; accepted for publication (in revised form) October 24, 2000; published electronically May 3, 2001. This research was partially supported by the NSF of China under grant 19901024.

http://www.siam.org/journals/sicon/40-1/37169.html

†School of Mathematics, Sichuan University, Chengdu 610064, Sichuan Province, China (xuzhang@fudan.edu).

subset of Ω where the control is supported. The exact controllability of (1.3) is formulated as follows: For any given (u^0, u^1), $(v^0, v^1) \in H_0^1(\Omega) \times L^2(\Omega)$, one can find a control $f_\varepsilon \in L^2((0,T) \times \omega_\varepsilon)$ such that the weak solution $u_\varepsilon(\cdot) \in C([0,T]; H_0^1(\Omega)) \cap C^1([0,T]; L^2(\Omega))$ of (1.3) satisfies

$$u_\varepsilon(T) = v^0, \qquad u_{\varepsilon,t}(T) = v^1. \tag{1.4}$$

If the final state (v^0, v^1) is $(0,0)$, the corresponding f_ε is called a null-control of (1.3) with initial state (u^0, u^1). We refer the reader to [7], [3], [11], and the references therein for an extensive study of the theory of exact controllability.

In the case $\omega_\varepsilon \equiv \omega$, i.e., the controller of system (1.3) is fixed, [9] considered the above singular pertubation problem under the following geometric control condition: [1]

$$\omega = \Omega \cap \mathcal{O}_\delta(\Gamma(x_0)), \tag{1.5}$$

where $\delta > 0$ is a constant, $\Gamma(x_0) = \{x \in \Gamma \,|\, (x - x_0) \cdot \nu(x) > 0\}$, $x_0 \in \mathbb{R}^n$, $\nu(x)$ denotes the outward unit normal to Ω at $x \in \Gamma$, and $\cdot$ denotes the scalar product in $\mathbb{R}^n$. More precisely, [9] proved the following result.

THEOREM A. *Let $T > 0$ and Ω be a bounded domain of $\mathbb{R}^n$ of class C^∞. Let $\omega_\varepsilon \equiv \omega$ with ω being given by* (1.5)*. Then, there exists an $\varepsilon_0 = \varepsilon_0(T, \Omega, \omega) > 0$ such that for any $0 < \varepsilon < \varepsilon_0$, system* (1.3) *is uniformly exactly controllable in time T. Furthermore, for any fixed $(u^0, u^1) \in H_0^1(\Omega) \times L^2(\Omega)$, the null-controls f_ε of* (1.3) *may be chosen such that*

$$f_\varepsilon \to f \text{ in } L^2((0,T) \times \omega) \text{ as } \varepsilon \to 0, \tag{1.6}$$

f being a null-control for the limit heat equation (1.1) *with initial datum u^0.*

Very recently, [10] generalized Theorem A by considering a more general geometric control condition introduced in [8]. It is also expected that Theorem A still holds under the sharp geometric control condition of [1] on ω although this is as of now an open problem.

However, we note that for the fixed controller case, according to the result in [1], some geometric conditions are needed on ω, the so-called *geometric optics condition*, to guarantee the exact controllability of (1.3). Therefore one may not expect the result of Theorem A to hold for any open nonempty subset ω of Ω.

In this note, we will allow the controller ω_ε of (1.3) to change as ε tends to zero. More precisely, for any given open nonempty subset ω of Ω, we suitably select a family of controllers ω_ε such that (1.3) with this controller is exactly controllable and

$$\lim_{\varepsilon \to 0} \omega_\varepsilon \setminus \omega = \emptyset.$$

Then, we can obtain the null exact controllability of (1.1) by considering the singular limit of the exact controllability properties of (1.3) with these controllers ω_ε.

This is precisely the main result of this note.

THEOREM B. *Let $T > 0$ and Ω be a bounded domain of $\mathbb{R}^n$ with C^∞ boundary Γ. Let ω be any given open nonempty subset of Ω and $\omega_\varepsilon = \omega \cup \big(\Omega \cap \mathcal{O}_\varepsilon(\Gamma)\big)$. Then, for any $(u^0, u^1) \in H_0^1(\Omega) \times L^2(\Omega)$ fixed, one can choose the null-controls f_ε of* (1.3) *such that*

$$f_\varepsilon \chi_{\omega_\varepsilon} \to f \chi_\omega \text{ in } L^2(Q) \text{ as } \varepsilon \to 0, \tag{1.7}$$

[1] For any $M \subseteq \mathbb{R}^n$ and $\eta > 0$, we put $\mathcal{O}_\eta(M) = \{y \in \mathbb{R}^n \,|\, |y - x| < \eta \text{ for some } x \in M\}$.

where f is a null-control for (1.1) *with initial datum u^0. Furthermore, it holds that*

$$\begin{array}{l} u_\varepsilon \to u \text{ strongly in } L^2(0,T;H_0^1(\Omega)) \cap C([0,T];L^2(\Omega)), \\ u_{\varepsilon,t} \to u_t \text{ strongly in } L^2(Q), \end{array} \tag{1.8}$$

where u_ε and u are the corresponding solutions of (1.3) *and* (1.1)*, respectively.*

Remark 1.1. Of course, one can select the controllers ω_ε of (1.3) in a different way. For example, we may take $\omega_\varepsilon = \omega \cup \big(\Omega \cap \mathcal{O}_\varepsilon(\Gamma(x_0))\big)$, where x_0 is a given point in $\mathbb{R}^n$.

Remark 1.2. As noted in [9], Theorem B provides the null-control of the limit heat equation only when $u^0 \in H_0^1(\Omega)$. However, due to the regularizing effect of the heat equation, as soon as we let the heat equation evolve freely (without control) during an arbitrarily short time interval, even if the initial datum lies in $L^2(\Omega)$, the solution enters $H_0^1(\Omega)$, and then the result above applies.

The proof of Theorem B follows the main steps developed in [9]. When doing this we need the following observability estimate for the solutions of the adjoint system:

$$\left\{ \begin{array}{lll} \varepsilon\varphi_{tt} - \Delta\varphi - \varphi_t = 0 & \text{in} & Q, \\ \varphi = 0 & \text{on} & \Sigma, \\ \varphi(T) = \varphi_0,\ \varphi_t(T) = \varphi_1 & \text{in} & \Omega. \end{array} \right. \tag{1.9}$$

THEOREM C. *Let Ω be a bounded domain of $\mathbb{R}^n$ with C^2 boundary Γ. Then for any $T > 0$ there exist two positive constants $\varepsilon_2 = \varepsilon_2(T,\Omega)$ and $C = C(T,\Omega)$ such that*

$$|\varphi_0|^2_{L^2(\Omega)} + \varepsilon|\varphi_1|^2_{H^{-1}(\Omega)} \le Ce^{C/\sqrt{\varepsilon}} \int_0^T \int_{\Omega\cap\mathcal{O}_\varepsilon(\Gamma)} \varphi^2 dxdt \tag{1.10}$$

for all $0 < \varepsilon < \varepsilon_2$ and every solution of (1.9).

The rest of this paper is organized as follows. In section 2, we reduce the proof of Theorem B to the explicit observability estimate (1.10) (in Theorem C). In section 3, we give the proof of Theorem C, which is performed using global Carleman inequalities.

2. Proof of Theorem B. This section is devoted to proving Theorem B. Similar to [9], we will reduce our problem to the sharp observability estimate on the eigenfunctions of the Laplacian in [6] and the explicit observability estimate in our Theorem C. For the reader's convenience, we will give the outline of the proof.

Step 1. *Some notations.*

First of all, we introduce the spectrum of the Laplacian

$$\left\{ \begin{array}{ll} -\Delta e_k = \mu_k e_k & \text{in} \quad \Omega, \\ e_k = 0 & \text{on} \quad \Gamma. \end{array} \right. \tag{2.1}$$

We know that (2.1) admits an increasing sequence of positive eigenvalues of finite multiplicity

$$0 < \mu_1 < \mu_2 \le \cdots \le \mu_k \le \cdots$$

tending to infinity. The eigenfunctions $\{e_k\}_{k\ge1}$ may be chosen to constitute an orthonormal basis of $L^2(\Omega)$.

Next, let us re-endow the Hilbert space $H = H_0^1(\Omega) \times L^2(\Omega)$ with the norm

$$|(u,v)|^2_{H_0^1(\Omega)\times L^2(\Omega)} = |u|^2_{H_0^1(\Omega)} + \varepsilon|v|^2_{L^2(\Omega)}. \tag{2.2}$$

We shall denote the norm (2.2) as $|\cdot|_{H^\varepsilon}$ to make explicit its dependence on ε. The space H endowed with this norm will be denoted by H^ε.

Finally, let us introduce the following subspaces of the space H^ε:

$$H_p^\varepsilon = \{U = (u, v) \in H^\varepsilon : u, v \in \text{span}_{1\le k\le k(\varepsilon)}(e_k)\}, \text{ where } k(\varepsilon) \text{ is such that}$$

$$\frac{1}{4\mu_{k(\varepsilon)+1}} < \varepsilon \le \frac{1}{4\mu_{k(\varepsilon)}};$$

$$H_h^\varepsilon = \{U = (u, v) \in H^\varepsilon : u, v \in \text{span}_{k\ge k(\varepsilon)+1}(e_k)\}, \text{ where } k(\varepsilon) \text{ is as above.}$$

Given $(u, v) \in H^\varepsilon$ we denote its orthogonal projections over H_p^ε by $\pi_p^\varepsilon(u, v)$.

Step 2. *Some preliminaries.*

First of all, we need the following theorem, which concerns the uniform boundedness of the controls of the parabolic component of solutions of (1.3), which is proved in [9] by means of the estimate of [6].

THEOREM 2.1. *Let Ω be a bounded domain of $\mathbb{R}^n$ of class C^∞. Let ω be any open nonempty subset of Ω. Then, there exist two positive constants $\varepsilon_3 = \varepsilon_3(T, \Omega, \omega)$ and $C = C(T, \Omega, \omega)$ such that for all $0 < \varepsilon < \varepsilon_3$ and $(u^0, u^1) \in H^\varepsilon$, there exists a control $f_\varepsilon \in L^2((0, T) \times \omega)$ such that the solution u_ε of*

$$\left\{\begin{array}{lll} \varepsilon u_{\varepsilon,tt} - \Delta u_\varepsilon + u_{\varepsilon,t} = f_\varepsilon \chi_\omega & in & Q, \\ u_\varepsilon = 0 & on & \Sigma, \\ u_\varepsilon(0) = u^0,\ u_{\varepsilon,t}(0) = u^1 & in & \Omega \end{array}\right. \tag{2.3}$$

satisfies

$$\pi_p^\varepsilon(u_\varepsilon(T), u_{\varepsilon,t}(T)) = 0 \tag{2.4}$$

and

$$|f_\varepsilon|_{L^2((0,T)\times\omega)} \le C|(u^0, u^1)|_{H^\varepsilon} \quad \forall\ (u^0, u^1) \in H^\varepsilon. \tag{2.5}$$

Next, let us recall the following known result (see [9]), which is devoted to analyzing the dissipativity of (1.3) over the solutions whose parabolic component vanishes.

THEOREM 2.2. *Let u_ε be a solution of* (1.3) *with $f_\varepsilon \equiv 0$ with initial data $(u^0, u^1) \in H_h^\varepsilon$. Then $(u_\varepsilon(t), u_{\varepsilon,t}(t)) \in H_h^\varepsilon$ for all $t \ge 0$ and*

$$|(u_\varepsilon(t), u_{\varepsilon,t}(t))|_{H^\varepsilon} \le 2^{3/2} e^{-\frac{t}{4\varepsilon}} |(u^0, u^1)|_{H^\varepsilon}. \tag{2.6}$$

Finally, by means of Theorem C (see section 3 for its proof), similar to the proof of Proposition 5.1 in [9], one gets the following result.

THEOREM 2.3. *Let Ω be a bounded domain of $\mathbb{R}^n$ of class C^2. Then, there exists a positive constant $C = C(T, \Omega)$ such that for all $0 < \varepsilon < \varepsilon_2$ (recall Theorem C for ε_2) and $(u^0, u^1) \in H^\varepsilon$, there exists a control $f_\varepsilon \in L^2((0, T) \times (\Omega \cap \mathcal{O}_\varepsilon(\Gamma)))$ such that the solution u_ε of*

$$\left\{\begin{array}{lll} \varepsilon u_{\varepsilon,tt} - \Delta u_\varepsilon + u_{\varepsilon,t} = f_\varepsilon \chi_{\Omega\cap\mathcal{O}_\varepsilon(\Gamma)} & in & Q, \\ u_\varepsilon = 0 & on & \Sigma, \\ u_\varepsilon(0) = u^0,\ u_{\varepsilon,t}(0) = u^1 & in & \Omega \end{array}\right. \tag{2.7}$$

satisfies

$$u_\varepsilon(T) = u_{\varepsilon,t}(T) = 0 \tag{2.8}$$

and

$$|f_\varepsilon|_{L^2((0,T)\times(\Omega\cap\mathcal{O}_\varepsilon(\Gamma)))} \le Ce^{C/\sqrt{\varepsilon}}|(u^0,u^1)|_{H^\varepsilon} \quad \forall\, (u^0,u^1)\in H^\varepsilon. \tag{2.9}$$

Step 3. *Description of the control strategy.*

Let us describe our control strategy. Given $T>0$, denote (recall Theorem 2.1 and Theorem 2.3 for ε_3 and ε_2, respectively)

$$\varepsilon_1 = \varepsilon_1(T,\Omega,\omega) \stackrel{\triangle}{=} \min(\varepsilon_2(T/3,\Omega), \varepsilon_3(T/3,\Omega,\omega)). \tag{2.10}$$

We divide the time interval $[0,T]$ in three subintervals $[0,T] = I_1\cup I_2\cup I_3$ with $I_1=[0,T/3]$, $I_2=[T/3,2T/3]$, and $I_3=[2T/3,T]$. Given an initial datum $(u^0,u^1)\in H_0^1(\Omega)\times L^2(\Omega)$ to be controlled, we proceed as follows.

• **First step.** In the first time interval I_1 we control the parabolic component of the solutions. By Theorem 2.1, for any $\varepsilon\in(0,\varepsilon_1)$ we can build a control $f_{1,\varepsilon}\in L^2((0,T/3)\times\omega)$ such that the solution of

$$\begin{cases} \varepsilon u_{\varepsilon,tt} - \Delta u_\varepsilon + u_{\varepsilon,t} = f_{1,\varepsilon}\chi_\omega & \text{in} \quad (0,T/3)\times\Omega,\\ u_\varepsilon = 0 & \text{on} \quad (0,T/3)\times\Gamma,\\ u_\varepsilon(0)=u^0,\ u_{\varepsilon,t}(0)=u^1 & \text{in} \quad \Omega \end{cases} \tag{2.11}$$

satisfies

$$\pi_p^\varepsilon(u_\varepsilon(T/3),\, u_{\varepsilon,t}(T/3)) = 0 \tag{2.12}$$

and

$$|f_{1,\varepsilon}|_{L^2(\omega\times(0,T/3))} \le C|(u^0,u^1)|_{H^\varepsilon} \quad \forall\,(u^0,u^1)\in H^\varepsilon, \quad \forall\, 0<\varepsilon<\varepsilon_1 \tag{2.13}$$

for some constant $C=C(T,\Omega,\omega)>0$.

In view of the uniform bound (2.13) of the control, by classical energy estimates, we have

$$|(u_\varepsilon(T/3),u_{\varepsilon,t}(T/3))|_{H^\varepsilon} \le C|(u^0,u^1)|_{H^\varepsilon} \quad \forall\,(u^0,u^1)\in H^\varepsilon, \quad \forall\, 0<\varepsilon<\varepsilon_1. \tag{2.14}$$

• **Second step.** In the time interval I_2 we let the equation evolve freely; i.e., we solve

$$\begin{cases} \varepsilon u_{\varepsilon,tt} - \Delta u_\varepsilon + u_{\varepsilon,t} = 0 & \text{in} \quad (T/3,2T/3)\times\Omega,\\ u_\varepsilon = 0 & \text{on} \quad (T/3,2T/3)\times\Gamma,\\ u_\varepsilon(T/3) = v_\varepsilon^0 \stackrel{\triangle}{=} u_\varepsilon(T/3), \quad u_{\varepsilon,t}(T/3) = v_\varepsilon^1 \stackrel{\triangle}{=} u_{\varepsilon,t}(T/3) & \text{in} \quad \Omega. \end{cases} \tag{2.15}$$

By Theorem 2.2, we see that

$$\pi_p^\varepsilon(u_\varepsilon(t),\, u_{\varepsilon,t}(t)) = 0 \quad \forall\, T/3\le t\le 2T/3 \tag{2.16}$$

and

$$|(u_\varepsilon(t),\, u_{\varepsilon,t}(t))|_{H^\varepsilon} \le Ce^{-\frac{(t-T/3)}{4\varepsilon}}|(v_\varepsilon^0,v_\varepsilon^1)|_{H^\varepsilon} \quad \forall\, T/3\le t\le 2T/3, \quad \forall\, 0<\varepsilon<\varepsilon_1 \tag{2.17}$$

for some generic constant $C > 0$. Thus, according to (2.14), we get

$$|(u_\varepsilon(2T/3),\, u_{\varepsilon,t}(2T/3))|_{H^\varepsilon} \le Ce^{\frac{-T}{12\varepsilon}}|(u^0,\, u^1)|_{H^\varepsilon} \quad \forall\, 0 < \varepsilon < \varepsilon_1. \tag{2.18}$$

• **Third step.** In this last step we control the whole solution to zero. By Theorem 2.3, for any $\varepsilon \in (0, \varepsilon_1)$, one can find a control $f_{2,\varepsilon} \in L^2((2T/3, T) \times (\Omega \cap \mathcal{O}_\varepsilon(\Gamma)))$ such that the solution of

$$\begin{cases} \varepsilon u_{\varepsilon,tt} - \Delta u_\varepsilon + u_{\varepsilon,t} = f_{2,\varepsilon}\chi_{\Omega\cap\mathcal{O}_\varepsilon(\Gamma)} & \text{in} \quad (2T/3, T) \times \Omega, \\ u_\varepsilon = 0 & \text{on} \quad (2T/3, T) \times \Gamma, \\ u_\varepsilon(2T/3) = w_\varepsilon^0 \triangleq u_\varepsilon(2T/3), \quad u_{\varepsilon,t}(2T/3) = w_\varepsilon^1 \triangleq u_{\varepsilon,t}(2T/3) & \text{in} \quad \Omega \end{cases} \tag{2.19}$$

satisfies

$$u_\varepsilon(T) \equiv u_{\varepsilon,t}(T) \equiv 0 \tag{2.20}$$

and

$$|f_{2,\varepsilon}|_{L^2(\omega\times(2T/3,T))} \le Ce^{C/\sqrt{\varepsilon}}|(w_\varepsilon^0, w_\varepsilon^1)|_{H^\varepsilon} \quad \forall\, 0 < \varepsilon < \varepsilon_1 \tag{2.21}$$

for some constant $C = C(T, \Omega) > 0$.

Combining (2.18) and (2.21), we get

$$|f_{2,\varepsilon}|_{L^2((2T/3,T)\times(\Omega\cap\mathcal{O}_\varepsilon(\Gamma)))} \le Ce^{C/\sqrt{\varepsilon}}e^{\frac{-T}{12\varepsilon}}|(u^0, u^1)|_{H^\varepsilon} \quad \forall\, 0 < \varepsilon < \varepsilon_1 \tag{2.22}$$

for all initial data $(u^0, u^1) \in H^\varepsilon$.

This shows that the control

$$f_\varepsilon = \begin{cases} f_{1,\varepsilon}, & 0 \le t \le T/3, \\ 0, & T/3 \le t \le 2T/3, \\ f_{2,\varepsilon}, & 2T/3 \le t \le T, \end{cases} \tag{2.23}$$

is such that the null controllability condition (2.20) holds and, according to (2.13) and (2.22),

$$|f_\varepsilon\chi_{\omega_\varepsilon}|_{L^2(Q)} \le C|(u^0, u^1)|_{H^\varepsilon}. \tag{2.24}$$

Thus we get the uniformly exact controllability of system (1.3).

Step 4. *Completion of the proof.*

Now, by means of the uniform estimate (2.24), proceeding exactly as section 7 in [9], one gets (1.7). Also, similar to Appendix D in [9] (with numerous but small changes), it is easy to obtain (1.8). Thus, the proof of Theorem B is completed. □

3. Proof of Theorem C. We now prove Theorem C.

We first observe that by the change of the time variable $t \to \tau = (T - t)/\sqrt{\varepsilon}$, system (1.9) becomes

$$\begin{cases} \varphi_{\tau\tau} - \Delta\varphi + \frac{1}{\sqrt{\varepsilon}}\varphi_\tau = 0 & \text{in} \quad (0, \widehat{T}) \times \Omega, \\ \varphi = 0 & \text{on} \quad (0, \widehat{T}) \times \Gamma, \\ \varphi(0) = \varphi_0, \quad \varphi_\tau(0) = -\sqrt{\varepsilon}\varphi_1 & \text{in} \quad \Omega, \end{cases} \tag{3.1}$$

where $\widehat{T} = \frac{T}{\sqrt{\varepsilon}}$. To simplify the notation we rewrite it as

$$\begin{cases} w_{tt} - \Delta w + kw_t = 0 & \text{in} \quad Q, \\ w = 0 & \text{on} \quad \Sigma, \\ w(0) = w_0, \quad w_t(0) = w_1 & \text{in} \quad \Omega. \end{cases} \tag{3.2}$$

It is sufficient to derive the following estimate.

THEOREM 3.1. *Let $t_0 \in (2\mathrm{diam}\,\Omega, \infty)$ be given. Then there is a positive constant $C = C(t_0, \Omega)$ such that*

$$|w_0|^2_{L^2(\Omega)} + |w_1|^2_{H^{-1}(\Omega)} \le C\delta^{-6}e^{Ck}\int_0^T\int_{\Omega\cap\mathcal{O}_\delta(\Gamma)} w^2 dxdt \tag{3.3}$$
$$\forall\, k > 0, T \ge t_0, \delta > 0, \text{ and } (w_0, w_1) \in L^2(\Omega)\times H^{-1}(\Omega),$$

where $w(\cdot) \in C([0,T]; L^2(\Omega)) \cap C^1([0,T]; H^{-1}(\Omega))$ is the weak solution of (3.2).

In order to prove Theorem 3.1, we need some preliminaries.

In what follows, we use the notations

$$f_i = f_i(x) \triangleq \frac{\partial f(x)}{\partial x_i},\ i = 1, 2, \dots, n;\ \sum_i \triangleq \sum_{i=1}^n;\ \sum_{i,j} \triangleq \sum_{i,j=1}^n.$$

(On the other hand, x_i is always the ith coordinate of the point x.)

The following three lemmas can be found, for example, in [9] and [10].

LEMMA 3.2. *Let $\lambda > 0$, $\alpha \in (0,1)$, and $k \in \mathbb{R}$ be constant. Let $x_0 \in \mathbb{R}^n$, $T > 0$, and*

$$\begin{cases} \phi(t,s,x) = \frac{1}{2}\left[|x-x_0|^2 - \alpha(t-T/2)^2 - \alpha(s-T/2)^2\right], \\ \ell = \lambda\phi, \\ \Psi = (n-1+\alpha)\lambda. \end{cases} \tag{3.4}$$

Let $z = z(t,s,x) \in C^2(\mathbb{R}^{2+n})$. Denote

$$v \triangleq \theta z \quad \text{with} \quad \theta = e^{\ell}. \tag{3.5}$$

Then

$$\begin{aligned} &\theta^2|z_{tt} + z_{ss} - \Delta z + k(z_t + z_s)|^2 \\ &\ge \Big[(k-2\ell_t)(v_t^2 - v_s^2 + \textstyle\sum_j v_j^2) + 2(k-2\ell_s)v_tv_s \\ &\qquad +4\textstyle\sum_j(\ell_j v_t v_j) + 2\Psi v_t v + (k-2\ell_t)(A+\Psi)v^2\Big]_t \\ &\quad +\Big[(k-2\ell_s)(v_s^2 - v_t^2 + \textstyle\sum_j v_j^2) + 2(k-2\ell_t)v_tv_s \\ &\qquad +4\textstyle\sum_j(\ell_j v_s v_j) + 2\Psi v_s v + (k-2\ell_s)(A+\Psi)v^2\Big]_s \\ &\quad -2\textstyle\sum_j\Big[2\sum_i(\ell_i v_i v_j) - \ell_j\sum_i v_i^2 + (k-2\ell_t)v_tv_j + (k-2\ell_s)v_sv_j \\ &\qquad +\Psi v_j v + \ell_j(v_t^2 + v_s^2) - (A+\Psi)\ell_j v^2\Big]_j \\ &\quad +2(1-\alpha)\lambda\Big(v_t^2 + v_s^2 + \textstyle\sum_j v_j^2\Big) + Bv^2, \end{aligned} \tag{3.6}$$

where

$$\begin{aligned} A = {}&\lambda^2\left[\alpha^2(t-T/2)^2 + \alpha^2(s-T/2)^2 - |x-x_0|^2\right] \\ &+\alpha\lambda k(t+s-T) + (1+\alpha)\lambda \end{aligned} \tag{3.7}$$

and

$$\begin{aligned} B = {}&2(3+\alpha)\lambda^3|x-x_0|^2 - 2\alpha^2\lambda^3(1+3\alpha)\left[(t-T/2)^2 + (s-T/2)^2\right] \\ &-2\alpha\lambda k^2 - 2\alpha\lambda^2 k(1+3\alpha)(t+s-T) - \left[n^2 + 4\alpha n + 1 + 2\alpha + 5\alpha^2\right]\lambda^2. \end{aligned} \tag{3.8}$$

LEMMA 3.3. *Let $0 \le S_1 < S_2 < T_2 < T_1 \le T$ and $k \in \mathbb{R}$ be given. Then there is a constant $C > 0$, which is independent of k, such that*

$$\int_{S_2}^{T_2} |w_t(t,\cdot)|^2_{H^{-1}(\Omega)} dt \le C(1+k^2) \int_{S_1}^{T_1} |w(t,\cdot)|^2_{L^2(\Omega)} dt, \tag{3.9}$$

where $w(\cdot)$ is the weak solution of system (3.2).

LEMMA 3.4. *For any $k \in \mathbb{R}$, it holds that*

$$E(t) \le E(s) e^{2|k|T} \qquad \forall\, t, s \in [0,T], \tag{3.10}$$

where

$$E(t) \triangleq \frac{1}{2} \left(|w_t(t,\cdot)|^2_{H^{-1}(\Omega)} + |w(t,\cdot)|^2_{L^2(\Omega)} \right), \tag{3.11}$$

with $w(\cdot)$ being the weak solution of system (3.2).

The following lemma can be found, for example, in [2].

LEMMA 3.5. *If X is an open nonempty subset in $\mathbb{R}^n$ and K is a compact subset of X, then one can find $\xi \in C_0^\infty(X)$ with $0 \le \xi \le 1$ so that $\xi = 1$ on a neighborhood of K. And it holds that*

$$\sup_{x \in X} |\partial^\beta \xi(x)| \le C_\beta \delta^{-|\beta|}, \tag{3.12}$$

where $\beta = (\beta_1, \dots, \beta_n)$ is a multi-index with nonnegative integer components and $|\beta| = \beta_1 + \cdots + \beta_n$, $\delta = \operatorname{dist}(K, \partial X)/4$ and C_β depends only on β and n.

Now, let us prove Theorem 3.1.

Proof of Theorem 3.1. We divide the proof into several steps.

Step 1. *Selection of pseudoconvex function.*

First of all, we see easily that it suffices to prove Theorem 3.1 under the condition

$$T = t_0.$$

The main idea of our proof is to use the pointwise estimate (3.6) in Lemma 3.2. For this purpose, we need to choose some suitable pseudoconvex function ϕ, that is, to choose x_0 and α.

By $T > 2\operatorname{diam}\Omega$, one can find a point $x_0 \in \mathbb{R}^n \backslash \overline{\Omega}$ such that $T > 2\max_{x\in\Omega} |x - x_0|$. Put

$$R_0 \triangleq \min_{x\in\Omega} |x - x_0|, \qquad R_1 \triangleq \max_{x\in\Omega} |x - x_0|. \tag{3.13}$$

Then $R_0 > 0$ and $T > 2R_1$. Thus we can choose an $\alpha \in (3/4, 1)$ such that

$$R_1^2 < \alpha (T/2)^2. \tag{3.14}$$

Having chosen x_0 and α as above, we next introduce the desired pseudoconvex function ϕ by setting

$$\phi = \phi(t,s,x) \triangleq \frac{1}{2} \left[|x - x_0|^2 - \alpha (t - T/2)^2 - \alpha (s - T/2)^2 \right]. \tag{3.15}$$

Step 2. *Some notations and transformations.*
We need the following notations. Denote

$$\begin{cases} \mathcal{Q} \triangleq (0,T)\times Q, \quad \mathcal{S} \triangleq (0,T)\times \Sigma, \\ T_i \triangleq T/2 - b_i T, \quad T_i' \triangleq T/2 + b_i T, \\ \mathcal{Q}_i \triangleq (T_i,T_i')\times(T_i,T_i')\times\Omega, \quad \mathcal{S}_i \triangleq (T_i,T_i')\times(T_i,T_i')\times\Gamma \end{cases} \tag{3.16}$$

and

$$\Lambda_j \triangleq \left\{ (t,s,x)\in \mathcal{Q} \;\middle|\; 2\phi(t,s,x) \geq \frac{R_0^2}{j+2} \right\}, \tag{3.17}$$

where $i = 0,1,2,3$; $j = 0,1,2$ and $0 < b_0 < b_1 < b_2 < b_3 < 1/2$ will be given below.

In order to determine b_i $(i = 0,1,2,3)$, we need an idea in [5]. First of all, by (3.13)–(3.15), one sees that

$$\phi(0,s,x) = \phi(T,s,x) \leq \frac{1}{2}\left(R_1^2 - \frac{\alpha T^2}{4}\right) < 0 \quad \forall\, (s,x)\in Q. \tag{3.18}$$

Thus, one can find a $b_1 \in (0,1/2)$ (close to $1/2$) such that (recall (3.16)–(3.17) for $\mathcal{Q}_1$, T_1, T_1', and Λ_2)

$$\Lambda_2 \subset \mathcal{Q}_1 \tag{3.19}$$

and for any $(t,s,x) \in (((0,T_1)\cup(T_1',T))\times Q)\cup((0,T)\times((0,T_1)\cup(T_1',T))\times\Omega)$ it holds that

$$\phi(t,s,x) < 0. \tag{3.20}$$

Next, noting that $R_0 > 0$ and $\{T/2\}\times\{T/2\}\times\Omega \subset \Lambda_0$, thus one finds a small $b_0 \in (0,b_1)$ such that (recall (3.17) and (3.16) for Λ_0 and $\mathcal{Q}_0$, respectively)

$$\mathcal{Q}_0 \subset \Lambda_0. \tag{3.21}$$

Finally, we fix any two numbers b_2 and b_3 satisfying $b_1 < b_2 < b_3 < 1/2$.

Now, we note that (recall (3.8) for B)

$$B = B\chi_{\Lambda_2}(t,s,x) + B\chi_{\mathcal{Q}\setminus\Lambda_2}(t,s,x). \tag{3.22}$$

First, we will prove Theorem 3.1 under the condition $k > 0$ being big enough. For this purpose, let us take

$$k = \beta\lambda, \tag{3.23}$$

where $\beta \in (0,1)$ is a constant to be determined later. Note that by (3.8) and (3.23), we have

$$\begin{aligned} B\chi_{\Lambda_2}(t,s,x) &\geq \Big\{ 2\alpha(1+3\alpha)\Big[|x-x_0|^2 - \alpha(t-T/2)^2 - \alpha(s-T/2)^2\Big]\lambda^3 \\ &\quad -2\alpha[\beta + (1+3\alpha)(t+s-T)]\beta\lambda^3 \\ &\quad -[n^2 + 4\alpha n + 1 + 2\alpha + 5\alpha^2]\lambda^2\Big\}\chi_{\Lambda_2}(t,s,x) \\ &\geq \Big\{ \tfrac{1}{2}\alpha(1+3\alpha)R_0^2\lambda^3 - C\beta\lambda^3 - \Big[n^2 + 4\alpha n + 1 + 2\alpha + 5\alpha^2\Big]\lambda^2\Big\}\chi_{\Lambda_2}(t,s,x), \end{aligned} \tag{3.24}$$

where $C = C(\alpha, T) > 0$ is a constant, independent of β. Thus, by (3.23)–(3.24) and (3.8), one sees easily that there exist a sufficiently small constant $\beta \in (0,1)$ and a sufficiently large constant $\lambda_1 > 1$ such that for any $\lambda > \lambda_1$, it holds that

$$B\chi_{\Lambda_2}(t,s,x) \geq c_0\lambda^3\chi_{\Lambda_2}(t,s,x) \tag{3.25}$$

and

$$\left|B\chi_{\mathcal{Q}\setminus\Lambda_2}(t,s,x)\right| \leq C\lambda^3 \tag{3.26}$$

for some constants $c_0 > 0$ and $C > 0$.

Finally, we need some transformations. For any fixed $\delta > 0$, by Lemma 3.5, we can choose a function $\xi \in C_0^\infty(\mathbb{R}^n; [0,1])$ such that

$$\begin{cases} \xi \equiv 1 & \text{on } \Omega \setminus \mathcal{O}_{\delta/2}(\Gamma); \\ \xi \equiv 0 & \text{on } \Omega \cap \mathcal{O}_{\delta/3}(\Gamma). \end{cases} \tag{3.27}$$

Set

$$p = p(t,x) \triangleq \xi(x)w(t,x), \quad (t,x) \in Q, \tag{3.28}$$

where w is the weak solution of (3.2). Then by (3.2), it is easy to see that

$$\begin{cases} p_{tt} - \Delta p + kp_t = -w\Delta\xi - 2(\nabla w)\cdot(\nabla\xi) & \text{in } Q, \\ p = 0 & \text{on } \Sigma, \\ p \equiv 0 & \text{in } (0,T)\times(\Omega\cap\mathcal{O}_{\delta/3}(\Gamma)). \end{cases} \tag{3.29}$$

The following simple transformations will play an important role in what follows. Put

$$z(t,s,x) \triangleq \int_s^t p(\tau,x)d\tau, \qquad Z(t,s,x) \triangleq \int_s^t w(\tau,x)d\tau. \tag{3.30}$$

Then $z(\cdot)$ satisfies (recall (3.16) for $\mathcal{Q}$ and $\mathcal{S}$)

$$\begin{cases} z_{tt} + z_{ss} - \Delta z + k(z_t + z_s) = -Z\Delta\xi - 2(\nabla Z)\cdot(\nabla\xi) & \text{in } \mathcal{Q}, \\ z = 0 & \text{on } \mathcal{S}, \\ z \equiv 0 & \text{in } (0,T)^2\times(\Omega\cap\mathcal{O}_{\delta/3}(\Gamma)). \end{cases} \tag{3.31}$$

Step 3. *Carleman-type estimate.*

For any given $\tau \in (T_2, T_1)$ and $\tau' \in (T_1', T_2')$ (recall (3.16) for T_i and T_i'), denote

$$\mathcal{Q}_\tau^{\tau'} \triangleq (\tau,\tau')\times(\tau,\tau')\times\Omega. \tag{3.32}$$

Let us observe (3.6), where z is given by (3.30) and ϕ is given by (3.15). Integrating (3.6) on $\mathcal{Q}_\tau^{\tau'}$, using integration by parts, and taking (3.31) into account, we arrive at the following (recall (3.5) for v and θ, and recall (3.16) for $\mathcal{Q}_2$, T_2, and T_2'):

$$\begin{aligned}
&2(1-\alpha)\lambda\textstyle\int_{\mathcal{Q}_\tau^{\tau'}}(v_t^2 + v_s^2 + \sum_i v_i^2)dxdtds + \int_{\mathcal{Q}_\tau^{\tau'}} Bv^2dxdtds \\
&\leq \textstyle\int_{\mathcal{Q}_2}\theta^2\left|Z\Delta\xi + 2(\nabla Z)\cdot(\nabla\xi)\right|^2dxdtds \\
&\quad + C\lambda^3\Big[\textstyle\int_{T_2}^{T_2'}\int_\Omega\Big(|v(\tau,s,x)|^2 + |v_t(\tau,s,x)|^2 + |v_s(\tau,s,x)|^2 + \sum_i|v_i(\tau,s,x)|^2 \\
&\quad + |v(\tau',s,x)|^2 + |v_t(\tau',s,x)|^2 + |v_s(\tau',s,x)|^2 + \textstyle\sum_i|v_i(\tau',s,x)|^2\Big)dxds \\
&\quad + \textstyle\int_{T_2}^{T_2'}\int_\Omega\Big(|v(t,\tau,x)|^2 + |v_t(t,\tau,x)|^2 + |v_s(t,\tau,x)|^2 + \sum_i|v_i(t,\tau,x)|^2 \\
&\quad + |v(t,\tau',x)|^2 + |v_t(t,\tau',x)|^2 + |v_s(t,\tau',x)|^2 + \textstyle\sum_i|v_i(t,\tau',x)|^2\Big)dxdt\Big] \quad \forall\,\lambda > \lambda_1.
\end{aligned} \tag{3.33}$$

However, recalling $v=\theta z$ with $\theta=e^{\ell}$, by (3.15) and (3.20), we get

$$
\begin{aligned}
&\int_{T_2}^{T_2'}\int_\Omega \Big(|v(\tau,s,x)|^2+|v_t(\tau,s,x)|^2+|v_s(\tau,s,x)|^2+\textstyle\sum_i|v_i(\tau,s,x)|^2\\
&\quad+|v(\tau',s,x)|^2+|v_t(\tau',s,x)|^2+|v_s(\tau',s,x)|^2+\textstyle\sum_i|v_i(\tau',s,x)|^2\Big)dxds\\
&\quad+\int_{T_2}^{T_2'}\int_\Omega \Big(|v(t,\tau,x)|^2+|v_t(t,\tau,x)|^2+|v_s(t,\tau,x)|^2+\textstyle\sum_i|v_i(t,\tau,x)|^2\\
&\quad+|v(t,\tau',x)|^2+|v_t(t,\tau',x)|^2+|v_s(t,\tau',x)|^2+\textstyle\sum_i|v_i(t,\tau',x)|^2\Big)dxdt\\
&\le C\lambda^2\Big[\int_{T_2}^{T_2'}\int_\Omega \Big(|z(\tau,s,x)|^2+|z_t(\tau,s,x)|^2+|z_s(\tau,s,x)|^2+\textstyle\sum_i|z_i(\tau,s,x)|^2\\
&\quad+|z(\tau',s,x)|^2+|z_t(\tau',s,x)|^2+|z_s(\tau',s,x)|^2+\textstyle\sum_i|z_i(\tau',s,x)|^2\Big)dxds\\
&\quad+\int_{T_2}^{T_2'}\int_\Omega \Big(|z(t,\tau,x)|^2+|z_t(t,\tau,x)|^2+|z_s(t,\tau,x)|^2+\textstyle\sum_i|z_i(t,\tau,x)|^2\\
&\quad+|z(t,\tau',x)|^2+|z_t(t,\tau',x)|^2+|z_s(t,\tau',x)|^2+\textstyle\sum_i|z_i(t,\tau',x)|^2\Big)dxdt\Big].
\end{aligned}
\tag{3.34}
$$

Further, by (3.17), (3.4)–(3.5), (3.15), and (3.25)–(3.26), we get

$$
\begin{aligned}
\int_{\mathcal{Q}_\tau^{\tau'}}Bv^2dxdtds &= \int_{\mathcal{Q}_\tau^{\tau'}\cap\Lambda_2}Bv^2dxdtds+\int_{\mathcal{Q}_\tau^{\tau'}\setminus\Lambda_2}Bv^2dxdtds\\
&\ge c_0\lambda^3\int_{\mathcal{Q}_\tau^{\tau'}\cap\Lambda_2}v^2dxdtds-C\lambda^3e^{R_0^2\lambda/4}\int_{\mathcal{Q}}z^2dxdtds\quad\forall\,\lambda>\lambda_1.
\end{aligned}
\tag{3.35}
$$

Note that by (3.17), (3.19), and (3.32), we have $\mathcal{Q}_\tau^{\tau'}\supset\Lambda_1$. Thus, by (3.35), for any $\lambda>\lambda_1$, we have

$$
\begin{aligned}
&2(1-\alpha)\lambda\int_{\mathcal{Q}_\tau^{\tau'}}(v_t^2+v_s^2+\textstyle\sum_i v_i^2)dxdtds+\int_{\mathcal{Q}_\tau^{\tau'}}Bv^2dxdtds\\
&\ge c_1\Big[\lambda\int_{\Lambda_1}(v_t^2+v_s^2+\textstyle\sum_i v_i^2)dxdtds+\lambda^3\int_{\Lambda_1}v^2dxdtds\Big]-C\lambda^3e^{R_0^2\lambda/4}\int_{\mathcal{Q}}z^2dxdtds,
\end{aligned}
\tag{3.36}
$$

where $c_1>0$ and $C>0$ are two constants, independent of λ.

Now, combining (3.33)–(3.34) and (3.36), we conclude that for any $\lambda>\lambda_1$, it holds that

$$
\begin{aligned}
&\int_{\Lambda_1}(v_t^2+v_s^2+\textstyle\sum_i v_i^2)dxdtds+\lambda^2\int_{\Lambda_1}\theta^2v^2dxdtds\\
&\le C\lambda^{-1}\Big\{e^{C\lambda}\int_{\mathcal{Q}_2}|Z\Delta\xi+2(\nabla Z)\cdot(\nabla\xi)|^2dxdtds\\
&\quad+\lambda^5\Big[\int_{T_2}^{T_2'}\int_\Omega \Big(|z(\tau,s,x)|^2+|z_t(\tau,s,x)|^2+|z_s(\tau,s,x)|^2+\textstyle\sum_i|z_i(\tau,s,x)|^2\\
&\quad+|z(\tau',s,x)|^2+|z_t(\tau',s,x)|^2+|z_s(\tau',s,x)|^2+\textstyle\sum_i|z_i(\tau',s,x)|^2\Big)dxds\\
&\quad+\int_{T_2}^{T_2'}\int_\Omega \Big(|z(t,\tau,x)|^2+|z_t(t,\tau,x)|^2+|z_s(t,\tau,x)|^2+\textstyle\sum_i|z_i(t,\tau,x)|^2\\
&\quad+|z(t,\tau',x)|^2+|z_t(t,\tau',x)|^2+|z_s(t,\tau',x)|^2+\textstyle\sum_i|z_i(t,\tau',x)|^2\Big)dxdt\Big]\\
&\quad+\lambda^3e^{R_0^2\lambda/4}\int_{\mathcal{Q}}z^2dxdtds\Big\}.
\end{aligned}
\tag{3.37}
$$

Integrating (3.37) with respect to τ and τ' from T_2 to T_1 and from T_1' to T_2', respectively, we get

$$
\begin{aligned}
&\int_{\Lambda_1}(v_t^2+v_s^2+\textstyle\sum_i v_i^2)dxdtds+\lambda^2\int_{\Lambda_1}v^2dxdtds\\
&\le C\lambda^{-1}\Big\{e^{C\lambda}\int_{\mathcal{Q}_2}|Z\Delta\xi+2(\nabla Z)\cdot(\nabla\xi)|^2dxdtds\\
&\qquad+\lambda^5\int_{\mathcal{Q}_2}\Big(z^2+z_t^2+z_s^2+\textstyle\sum_i z_i^2\Big)dxdtds+\lambda^3e^{R_0^2\lambda/4}\int_{\mathcal{Q}}z^2dxdtds\Big\}.
\end{aligned}
\tag{3.38}
$$

Consequently, by (3.4)–(3.5) and (3.15), and using (3.38), we see that for any $\lambda>\lambda_1$,

it holds that

$$\begin{aligned}&\int_{\Lambda_1}\theta^2(z_t^2+z_s^2+\textstyle\sum_i z_i^2)dxdtds\\ &\le C\lambda^{-1}\Big\{e^{C\lambda}\int_{\mathcal{Q}_2}|Z\Delta\xi+2(\nabla Z)\cdot(\nabla\xi)|^2dxdtds\\ &\qquad+\lambda^5\int_{\mathcal{Q}_2}\Big(z^2+z_t^2+z_s^2+\textstyle\sum_i z_i^2\Big)dxdtds+\lambda^3e^{R_0^2\lambda/4}\int_{\mathcal{Q}}z^2dxdtds\Big\}.\end{aligned}\tag{3.39}$$

Note that by (3.17) and (3.21), we have

$$\int_{\Lambda_1}\theta^2(z_t^2+z_s^2)dxdtds\ge\int_{\Lambda_0}\theta^2(z_t^2+z_s^2)dxdtds\ge e^{R_0^2\lambda/2}\int_{\mathcal{Q}_0}(z_t^2+z_s^2)dxdsdt.\tag{3.40}$$

Thus, by (3.39)–(3.40), we see that for any $\lambda>\lambda_1$ it holds that

$$\begin{aligned}&\int_{\mathcal{Q}_0}(z_t^2+z_s^2)dxdtds\\ &\le C_1\lambda^{-1}\Big\{e^{C_1\lambda}\int_{\mathcal{Q}_2}|Z\Delta\xi+2(\nabla Z)\cdot(\nabla\xi)|^2dxdtds\\ &\qquad+\lambda^5e^{-R_0^2\lambda/2}\int_{\mathcal{Q}_2}\Big(z^2+z_t^2+z_s^2+\textstyle\sum_i z_i^2\Big)dxdtds+\lambda^3e^{-R_0^2\lambda/4}\int_{\mathcal{Q}}z^2dxdtds\Big\},\end{aligned}\tag{3.41}$$

where $C_1>0$ is a generic constant.

Step 4. *Estimate on* $\int_{\mathcal{Q}_2}\sum_i z_i^2dxdsdt$.

Denote

$$\eta\triangleq(t-T_3)(T_3'-t)(s-T_3)(T_3'-s).\tag{3.42}$$

Multiplying the first equation of (3.31) by ηz, integrating it on $\mathcal{Q}_3$ (recall (3.16) for $\mathcal{Q}_3$), using integration by parts, by (3.23), and noting that

$$\eta(t,s)\ge(T_2-T_3)^2(T_3'-T_2)^2\quad\forall\,t\,,s\in(T_2,T_2'),$$

we get

$$\int_{\mathcal{Q}_2}\sum_i z_i^2dxdsdt\le C\left[\lambda\int_{\mathcal{Q}_3}(z_t^2+z_s^2+z^2)dxdsdt+\int_{\mathcal{Q}_3}|Z\Delta\xi+2(\nabla Z)\cdot(\nabla\xi)|^2dxdsdt\right].\tag{3.43}$$

Thus, by (3.41) and (3.43), we conclude that there is a constant $\lambda_2>\lambda_1$, which depends only on C_1 (in (3.41)) and R_0, such that

$$\begin{aligned}\int_{\mathcal{Q}_0}(z_t^2+z_s^2)dxdtds\le C\Big\{&\lambda^{-3}e^{C\lambda}\int_{\mathcal{Q}_3}|Z\Delta\xi+2(\nabla Z)\cdot(\nabla\xi)|^2dxdtds\\ &+\lambda^{-2}e^{-R_0^2\lambda/8}\int_{\mathcal{Q}}(z^2+z_t^2+z_s^2)dxdsdt\Big\}\quad\forall\,\lambda>\lambda_2.\end{aligned}\tag{3.44}$$

Step 5. *Estimate on* $\int_{\mathcal{Q}_3}|Z\Delta\xi+2(\nabla Z)\cdot(\nabla\xi)|^2dxdtds$.

Note that by (3.16), (3.27), and Lemma 3.5, using Poincáre's inequality, we have

$$\begin{aligned}\int_{\mathcal{Q}_3}|Z\Delta\xi+2(\nabla Z)\cdot(\nabla\xi)|^2dxdtds&=\int_{T_3}^{T_3'}\int_{T_3}^{T_3'}\int_{\Omega\cap\mathcal{O}_{\delta/2}(\Gamma)}|Z\Delta\xi+2(\nabla Z)\cdot(\nabla\xi)|^2dxdtds\\ &\le C\delta^{-4}\int_{T_3}^{T_3'}\int_{T_3}^{T_3'}\int_{\Omega\cap\mathcal{O}_{\delta/2}(\Gamma)}|\nabla Z|^2dxdtds.\end{aligned}\tag{3.45}$$

We need to estimate $\int_{T_3}^{T_3'}\int_{T_3}^{T_3'}\int_{\Omega\cap\mathcal{O}_{\delta/2}(\Gamma)}|\nabla Z|^2dxdsdt$. By (3.30) and (3.2), we see that $Z(\cdot)$ satisfies

$$\begin{cases}Z_{tt}+Z_{ss}-\Delta Z+k(Z_t+Z_s)=0 & \text{in }\mathcal{Q},\\ Z=0 & \text{on }\mathcal{S}.\end{cases}\tag{3.46}$$

By Lemma 3.5, we can choose a function $\zeta \in C_0^\infty(\mathbb{R}^n;[0,1])$ such that

$$\begin{cases} \zeta \equiv 1 & \text{on } \Omega \setminus \mathcal{O}_{2\delta/3}(\Gamma); \\ \zeta \equiv 0 & \text{on } \Omega \cap \mathcal{O}_{\delta/2}(\Gamma). \end{cases} \tag{3.47}$$

Denote

$$\psi \triangleq t(T-t)s(T-s)(1-\zeta). \tag{3.48}$$

Multiplying the first equation of (3.46) by $\psi^2 Z$, integrating it on $\mathcal{Q}$, by (3.46)–(3.47) and using integration by parts, and noting (3.23), we get for any $\lambda > 1$

$$\begin{aligned} &\int_{\mathcal{Q}} \psi^2 |\nabla Z|^2 dxdsdt \\ &= -\int_{\mathcal{Q}} \psi^2 \Big[Z_{tt} + Z_{ss} + k(Z_t + Z_s) \Big] Z dxdsdt - 2\int_{\mathcal{Q}} (\nabla Z)\cdot(\nabla\psi)\psi Z dxdsdt \\ &= 2\int_{\mathcal{Q}} \psi(\psi_t Z_t + \psi_s Z_s) Z dxdsdt + \int_{\mathcal{Q}} \psi^2 (Z_t^2 + Z_s^2) dxdsdt \\ &\quad -k\int_{\mathcal{Q}} \psi^2 (Z_t + Z_s) Z dxdsdt - 2\int_{\mathcal{Q}} \psi (\nabla Z)\cdot(\nabla\psi) Z dxdsdt \\ &\le C\lambda \int_{\mathcal{Q}} (1-\zeta)(Z^2 + Z_t^2 + Z_s^2) dxdsdt + C\int_{\mathcal{Q}} |\nabla\psi|^2 Z^2 dxdsdt \\ &\quad + \tfrac{1}{2}\int_{\mathcal{Q}} \psi^2 |\nabla Z|^2 dxdsdt. \end{aligned} \tag{3.49}$$

Note that by (3.47) and Lemma 3.5, we have

$$\begin{aligned} &\int_{\mathcal{Q}} |\nabla\psi|^2 Z^2 dxdsdt \le C\int_{\mathcal{Q}} |\nabla(1-\zeta)|^2 Z^2 dxdsdt \\ &= C\int_0^T \int_0^T \int_{\Omega\cap\mathcal{O}_\delta(\Gamma)} |\nabla\zeta|^2 Z^2 dxdsdt \le C\delta^{-2}\int_0^T \int_0^T \int_{\Omega\cap\mathcal{O}_\delta(\Gamma)} Z^2 dxdsdt. \end{aligned} \tag{3.50}$$

Thus, combining (3.49)–(3.50), noting (3.48), we get

$$\begin{aligned} &\int_{T_3}^{T_3'} \int_{T_3}^{T_3'} \int_{\Omega\cap\mathcal{O}_{\delta/2}(\Gamma)} |\nabla Z|^2 dxdsdt \\ &\le C(\lambda + \delta^{-2}) \int_0^T \int_0^T \int_{\Omega\cap\mathcal{O}_\delta(\Gamma)} (Z^2 + Z_t^2 + Z_s^2) dxdsdt. \end{aligned} \tag{3.51}$$

Combining (3.45) and (3.51), we get

$$\begin{aligned} &\int_{\mathcal{Q}_3} |Z\Delta\xi + 2(\nabla Z)\cdot(\nabla\xi)|^2 dxdtds \\ &\le C\delta^{-4}(\lambda + \delta^{-2}) \int_0^T \int_0^T \int_{\Omega\cap\mathcal{O}_\delta(\Gamma)} (Z^2 + Z_t^2 + Z_s^2) dxdsdt. \end{aligned} \tag{3.52}$$

Now, by (3.44) and (3.52), and noting that $\delta \in (0,1)$ and $\lambda_2 > 1$, we get

$$\begin{aligned} \int_{\mathcal{Q}_0} (z_t^2 + z_s^2) dxdtds \le C\lambda^{-2} \Big\{ &\delta^{-6} e^{C\lambda} \int_0^T \int_0^T \int_{\Omega\cap\mathcal{O}_\delta(\Gamma)} (Z^2 + Z_t^2 + Z_s^2) dxdsdt \\ &+ e^{-R_0^2\lambda/8} \int_{\mathcal{Q}} (z^2 + z_t^2 + z_s^2) dxdsdt \Big\} \quad \forall\, \lambda > \lambda_2. \end{aligned} \tag{3.53}$$

Step 6. *Completion of the proof when k large enough.*

Let us return to "w." By (3.53), (3.16), (3.28), and (3.30), one arrives at

$$\begin{aligned} \int_{T_0}^{T_0'} \int_{\Omega\setminus\mathcal{O}_{\delta/2}(\Gamma)} w^2 dxdt \le C\lambda^{-2} \Big[&\delta^{-6} e^{C\lambda} \int_0^T \int_{\Omega\cap\mathcal{O}_\delta(\Gamma)} w^2 dxdt \\ &+ e^{-R_0^2\lambda/8} \int_0^T \int_\Omega w^2 dxdt \Big] \quad \forall\, \lambda > \lambda_2. \end{aligned} \tag{3.54}$$

Now, adding both sides of (3.54) by $\int_{T_0}^{T_0'} \int_{\Omega\cap\mathcal{O}_{\delta/2}(\Gamma)} w^2 dxdt$, we conclude that

$$\begin{aligned} \int_{T_0}^{T_0'} \int_\Omega w^2 dxdt \le C\lambda^{-2} \Big[&\delta^{-6} e^{C\lambda} \int_0^T \int_{\Omega\cap\mathcal{O}_\delta(\Gamma)} w^2 dxdt \\ &+ e^{-R_0^2\lambda/8} \int_0^T E(t) dxdt \Big] \quad \forall\, \lambda > \lambda_2, \end{aligned} \tag{3.55}$$

where $E(t)$ is defined by (3.11).

However, taking two numbers S_0 and S_0' such that $T_0 < S_0 < S_0' < T_0'$, using Lemma 3.3, and noting (3.23), we see that

$$\int_{S_0}^{S_0'} E(s)ds \le C(1+\lambda^2)\int_{T_0}^{T_0'}\int_\Omega w^2 dxdt. \tag{3.56}$$

Combining (3.55) and (3.56), we arrive at

$$\int_{S_0}^{S_0'} E(s)ds \le C\left[\delta^{-6}e^{C\lambda}\int_0^T\int_{\Omega\cap\mathcal{O}_\delta(\Gamma)} w^2 dxdt + e^{-R_0^2\lambda/8}\int_0^T E(t)dxdt\right] \quad \forall\,\lambda>\lambda_2. \tag{3.57}$$

On the other hand, by Lemma 3.4 and (3.57), and noting (3.23), we get

$$E(0) \le C\left[\delta^{-6}e^{C\lambda}\int_0^T\int_{\Omega\cap\mathcal{O}_\delta(\Gamma)} w^2 dxdt + e^{(4T\beta-R_0^2/8)\lambda}E(0)\right] \quad \forall\,\lambda>\lambda_2. \tag{3.58}$$

Note that we can take $\beta>0$ small enough such that $4T\beta < R_0^2/12$. Thus by (3.58), we see that

$$E(0) \le C_2\left[\delta^{-6}e^{C_2\lambda}\int_0^T\int_{\Omega\cap\mathcal{O}_\delta(\Gamma)} w^2 dxdt + e^{-R_0^2\lambda/24}E(0)\right] \quad \forall\,\lambda>\lambda_2, \tag{3.59}$$

where $C_2>0$ is a generic constant.

Now, let us take a constant $\lambda_3>\lambda_2$ such that

$$C_2 e^{-R_0^2\lambda_3/24} < 1/2. \tag{3.60}$$

Then, by (3.59)–(3.60), we conclude that

$$E(0) \le C\delta^{-6}e^{C\lambda}\int_0^T\int_{\Omega\cap\mathcal{O}_\delta(\Gamma)} w^2 dxdt \quad \forall\,\lambda>\lambda_3. \tag{3.61}$$

Now, combining (3.61) and (3.23), we obtain the desired estimate under the condition $\lambda>\lambda_3$, or $k>k_0\triangleq\beta\lambda_3$ (recall (3.23)).

Step 7. *Completion of the proof when* $0<k\le k_0$.

Finally, let us consider the case $0<k\le k_0$. It is easy to see that (3.25)–(3.26) remains true in this case. Thus, proceeding as in (3.25)–(3.61) of the previous case the proof may be completed easily. □

REFERENCES

[1] C. Bardos, G. Lebeau, and J. Rauch, *Un exemple d'utilisation des notions de propagation pour le contrôle et la stabilisation des problèmes hyperboliques*, Rend. Sem. Mat. Univ. Politec. Torino, Fasc. Spec. (1988), pp. 11–31.

[2] L. Hörmander, *The Analysis of Linear Partial Differential Operators.* I, Distribution Theory and Fourier Analysis, Springer-Verlag, Berlin, 1983.

[3] V. Komornik, *Exact Controllability and Stabilization. The Multiplier Method*, John Wiley, Chichester, UK, Masson, Paris, 1995.

[4] A. Fursikov and O. Yu Imanuvilov, *Controllability of evolution equations*, Lecture Notes Series 34, Research Institute of Mathematics, Global Analysis Research Center, Seoul National University, Seoul, 1996.

[5] I. Lasiecka and R. Triggiani, *Carleman estimates and exact boundary controllability for a system of coupled, non-conservative second order hyperbolic equations*, in Partial Differential Equations Methods in Control and Shape Analysis, Lecture Notes in Pure and Appl. Math. 188, Marcel Dekker, New York, 1994, pp. 215–243.

[6] G. Lebeau and L. Robbiano, *Contrôle exact de l'équation de la chaleur*, Comm. Partial Differential Equations, 20 (1995), pp. 335–356.

[7] J.-L. Lions, *Contrôlabilité exacte, perturbations et systémes distribués, tome* 1, RMA Res. Notes Appl. Math. 8, Masson, Paris, 1988.

[8] K. Liu, *Locally distributed control and damping for the conservative systems*, SIAM J. Control Optim., 35 (1997), pp. 1574–1590.

[9] A. López, X. Zhang, and E. Zuazua, *Null controllability of the heat equation as singular limit of the exact controllability of dissipative wave equations*, J. Math. Pures Appl., 79 (2000), pp. 741–808.

[10] Y. Tang and X. Zhang, *A note on the singular limit of the exact controllability of dissipative wave equations*, Acta Math. Sin. (Engl. Ser.), 16 (2000), pp. 601–612.

[11] E. Zuazua, *Some problems and results on the controllability of partial differential equations*, in European Congress of Mathematics, Vol. II (Budapest, 1996), Progr. Math. 169, Birkhäuser-Verlag, Basel, 1998, pp. 276–311.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 1, pp. 54–63

A CONJUGATE POINTS THEORY FOR A NONLINEAR PROGRAMMING PROBLEM*

H. KAWASAKI†

Abstract. The conjugate point is an important global concept in the calculus of variations and optimal control. In these extremal problems, the variable is not a vector in R^n but a function. So a simple and natural question arises. Is it possible to establish a conjugate points theory for a nonlinear programming problem, Min $f(x)$ on $x \in R^n$? This paper positively answers this question. We introduce the Jacobi equation and conjugate points for the nonlinear programming problem, and we describe necessary and sufficient optimality conditions in terms of conjugate points.

Key words. conjugate points, strict conjugate points, Jacobi equation, Legendre condition, positive-definite, shortest path problem, Sylvester's criterion, principal minors

AMS subject classifications. 49K10, 90C30, 26B99

PII. S0363012900368831

1. Introduction. In this paper, we establish a conjugate points theory for a nonlinear programming problem,

$$\text{(P)} \qquad \text{Min} \quad f(x) \quad \text{on} \quad x \in R^n,$$

where $f : R^n \to R$ is assumed to be twice continuously differentiable. We will define conjugate points for (P) based on the insight in Gelfand and Fomin [1] by comparing (P) with the simplest problem in the calculus of variations,

$$\text{(SP)} \quad \begin{array}{l} \text{Min} \displaystyle\int_0^T f(t, x(t), \dot{x}(t))dt \\ \text{subject to} \quad x(0) = A,\ x(T) = B, \end{array}$$

where A and B are given points in R^n, $T > 0$ fixed, and f is a smooth function. So, we first review the classical conjugate points theory for (SP) in brief. Let $\bar{x}(t)$ be a weak minimum for (SP). Then it satisfies the Euler equation $df_{\dot{x}}(t, \bar{x}(t), \dot{\bar{x}}(t))/dt = f_x(t, \bar{x}(t), \dot{\bar{x}}(t))$ (1744) and the Legendre condition $f_{\dot{x}\dot{x}}(t, \bar{x}(t), \dot{\bar{x}}(t)) \geq 0$ (1786). Legendre attempted to prove its inverse; that is, he expected that if a feasible solution $\bar{x}(t)$ satisfies the Euler equation and the strengthened Legendre condition $f_{\dot{x}\dot{x}}(t, \bar{x}(t), \dot{\bar{x}}(t)) > 0$, then $\bar{x}(t)$ would be a weak minimum. However, his conjecture was false. Jacobi solved this problem by introducing conjugate points in 1837. For a feasible solution $\bar{x}(t)$ of (SP), conjugate points are defined via the Jacobi equation

$$\frac{d}{dt}\{\bar{f}_{\dot{x}x}(t)y(t) + \bar{f}_{\dot{x}\dot{x}}(t)\dot{y}(t)\} = \bar{f}_{xx}(t)y(t) + \bar{f}_{x\dot{x}}(t)\dot{y}(t), \tag{1.1}$$

where $\bar{f}_{\dot{x}\dot{x}}(t) := f_{\dot{x}\dot{x}}(t, \bar{x}(t), \dot{\bar{x}}(t))$, etc. A point $c \in (0, T]$ is said to be *conjugate to* $t = 0$ if there exists a nontrivial solution $y(t)$ of the Jacobi equation (1.1) on $[0, c]$ and $y(0) = y(c) = 0$.

*Received by the editors March 6, 2000; accepted for publication (in revised form) November 14, 2000; published electronically May 3, 2001. This research was partially supported by the Grant-in Aid for General Scientific Research from the Japan Society for the Promotion of Science, 11440033.
http://www.siam.org/journals/sicon/40-1/36883.html

†Graduate School of Mathematics, Kyushu University, Hakozaki 6-10-1, Fukuoka 812-8581, Japan (kawasaki@math.kyushu-u.ac.jp).

THEOREM 1.1 (Jacobi). *If $\bar{x}(t)$ is a weak minimum for the simplest problem* (SP) *and it satisfies the strengthened Legendre condition, then there is no point conjugate to $t = 0$ on $[0, T)$. Conversely, if $\bar{x}(t)$ satisfies the Euler equation and the strengthened Legendre condition, and if there is no point conjugate to $t = 0$ on $[0, T]$, then $\bar{x}(t)$ is a weak minimum.*

Recently, conjugate points have been extended to complex extremal problems such as optimal control problems and variational problems with state constraints; see, for example, [3, 4, 6, 7, 8, 9, 10, 11]. The present paper is outside of this trend. We deal with the elementary extremal problem (P). It seems to the author that one reason why researchers have not paid much attention to conjugate points for (P) lies in the following elementary results.

THEOREM 1.2. *If $\bar{x}$ is a minimum for* (P)*, then it satisfies $f'(\bar{x}) = 0$ and $f''(\bar{x}) \geq 0$. Conversely, if $\bar{x}$ satisfies $f'(\bar{x}) = 0$ and $f''(\bar{x}) > 0$, then it is a minimum for* (P)*, where $\geq$ and $>$ stand for nonnegative definite and positive-definite, respectively.*

Theorem 1.2 seems to assert that there is no room for conjugate points to play a role in (P). However, an interesting connection between Jacobi's condition and the theory of quadratic forms in R^n was discussed in Gelfand and Fomin [1, p. 125]; see section 2 of the present paper. Furthermore, the author [2, p. 8] recently found a stimulating example that strongly indicates the possibility to establish a conjugate points theory for (P).

EXAMPLE 1.1. *Let us first consider the shortest path problem on the unit sphere S in R^3. It is finding a shortest path on S joining $A = (1, 0, 0)$ and $B = (\cos T, \sin T, 0)$, where $0 < T < 2\pi$ is given. When $T > \pi$, the equatorial arc, say, AB in Figure* 1.1*, is not a weak minimum. Indeed, take another great circle arc (the broken curve in Figure* 1.1*) joining A and $C = (-1, 0, 0)$, say, AC, and join AC and the equatorial arc CB. Then it has the same length as the equatorial arc AB. However, we get a shorter curve by taking a short cut around C.*

FIG. 1.1.

According to the classical conjugate point theory, C is conjugate to A. Next, let us approximate the shortest path problem by an extremal problem in a finite-dimensional space as follows.

1. *Take a finite number of equally located longitudes ℓ_0, $\ell_1, \ldots, \ell_{n+1}$, where we assume that $A \in \ell_0$ and $B \in \ell_{n+1}$; see Figure* 1.2*.*
2. *Choose one point, say, X_k, on each ℓ_k for $k = 1, 2, \ldots n$.*
3. *Minimize the length of the polygonal curve joining $A, X_1, \ldots, X_n, B$.*

FIG. 1.2.

FIG. 1.3.

Then the above observation on the classical shortest path problem can be also applied to this extremal problem in R^n. In fact, by taking a short cut around C in Figure 1.3*, we obtain a shorter polygonal curve. Hence C can be regarded as a conjugate point in the finite-dimensional analogue.*

Though Theorem 1.2 seems to negatively answer our question, Example 1.1 seems to positively answer. In this paper, we show that the latter is right. In section 2, we first clarify the trick that Theorem 1.2 seems to negatively answer. Next, we introduce the Jacobi equation and conjugate points for (P), and we give a sufficient optimality condition in terms of conjugate points. In section 3, we define strict conjugate points, and we give a necessary optimality condition in terms of strict conjugate points. In section 4, we provide three examples and compute conjugate points. Readers will see that our conjugate points theory works very well.

2. The Jacobi equation and conjugate points: Sufficiency. In this section, we define the Jacobi equation and conjugate points for (P), and we describe sufficient optimality conditions in terms of conjugate points. In order to investigate conjugate points for (P), we should start with clarifying what corresponds to the Euler equation and the (strengthened) Legendre condition in (P), respectively. The point to achieve this purpose is that both of them are local properties. Namely, they are derived by giving a minimum $\bar{x}(t)$ an increment $\Delta x(t)$ that takes nonzero value only on a neighborhood of t. Since the variable t corresponds to the index k and the index set

$\{1,\dots,n\}$ is discrete, we may take the singleton $\{k\}$ as the smallest neighborhood of k. Hence the increment becomes $\Delta x := (0,\dots,0,\Delta x_k,0,\dots,0)$. So, defining $F(\Delta x_k) := f(\bar{x}+\Delta x)$ for any fixed index k, we get $F'(0) = f_{x_k}(\bar{x})$ and $F''(0) = f_{x_k x_k}(\bar{x})$. Therefore, we conclude that

(a) the Euler equation corresponds to $f_{x_k}(\bar{x}) = 0 \quad \forall k = 1,\dots,n$,
(b) the Legendre condition corresponds to $f_{x_k x_k}(\bar{x}) \geq 0 \quad \forall k = 1,\dots,n$,
(c) the strengthened Legendre condition corresponds to $f_{x_k x_k}(\bar{x}) > 0 \ \forall k = 1,\dots,n$.

Since there is a gap between (c) and $f''(\bar{x}) > 0$, there does exist room for conjugate points to play a role in (P).

So, let us now discuss the positive-definiteness of symmetric matrices. According to Sylvester's criterion, an $n \times n$-symmetric matrix $A = (a_{ij})$ is positive-definite if and only if its descending principal minors $|A_k|$, $(k = 1,\dots,n)$ are positive, where

$$A_k := \begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{kk} \end{pmatrix}.$$

The following lemma shows that the determinant of any square matrix is expanded with respect to the descending principal minors.

LEMMA 2.1. *For any $n \times n$-matrix $A = (a_{ij})$, its determinant is expanded as*

$$|A| = \sum_{k=0}^{n-1} \sum_{\rho} \varepsilon(\rho) a_{k+1\rho(k+1)} a_{k+2\rho(k+2)} \cdots a_{n\rho(n)} |A_k|, \tag{2.1}$$

where $|A_0| := 1$, $\varepsilon(\rho)$ denotes the sign of ρ, and the summation is taken over all permutations ρ on $\{k+1,\dots,n\}$ satisfying that there is no $\ell > k$ such that ρ is closed on $\{\ell+1,\dots,n\}$.

Proof. For any permutation σ on $\{1,\dots,n\}$, put $k := \max\{0 \leq \ell \leq n-1 : \sigma(\{\ell+1,\dots,n\}) = \{\ell+1,\dots,n\}\}$, $\tau := \sigma|_{\{1,\dots,k\}}$, and $\rho := \sigma|_{\{k+1,\dots,n\}}$, where τ is empty when $k = 0$. Then $\sigma = \tau \circ \rho$, and ρ is closed on $\{k+1,\dots,n\}$. But there is no $\ell > k$ such that ρ is closed on $\{\ell+1,\dots,n\}$. Hence

$$\begin{aligned} |A| &= \sum_{k=0}^{n-1} \sum_{\rho} \varepsilon(\rho) \left\{ \sum_{\tau} \varepsilon(\tau) a_{1\tau(1)} \cdots a_{k\tau(k)} \right\} a_{k+1\rho(k+1)} \cdots a_{n\rho(n)} \\ &= \sum_{k=0}^{n-1} \sum_{\rho} \varepsilon(\rho) a_{k+1\rho(k+1)} \cdots a_{n\rho(n)} |A_k|. \end{aligned}$$

This completes the proof. □

EXAMPLE 2.1. *When $A = (A_{ij})$ is a tridiagonal matrix*

$$\begin{pmatrix} a_1 & b_1 & & \\ b_1 & a_2 & \ddots & \\ & \ddots & \ddots & b_{n-1} \\ & & b_{n-1} & a_n \end{pmatrix}, \tag{2.2}$$

the expansion (2.1) *reduces to*

$$|A_k| = a_k |A_{k-1}| - b_{k-1}^2 |A_{k-2}|, \tag{2.3}$$

which coincides with (81) *in Gelfand and Fomin* [1, p. 127].

In this paper, we call the expansion (2.1) the *Jacobi equation* for (P). In the following, we show the reason. A connection between Jacobi's condition and the theory of quadratic forms in R^n was discussed in [1, p. 127] as follows. According to the classical conjugate point theory, the quadratic functional

$$\int_0^T (P\dot{y}^2 + Ry^2)dt, \tag{2.4}$$

where $P(t) > 0$, is positive for all $y(t)$ such that $y(0) = y(T) = 0$ if and only if $[0, T]$ contains no point conjugate to 0, where the corresponding Jacobi equation is

$$\frac{d}{dt}(P\dot{y}) = Ry$$

(see [1, p. 111]). By introducing the points $0 = t_0 < t_1 < \cdots < t_n < t_{n+1} = T$, we get $n + 1$ equal parts of length $\Delta t := T/(n + 1)$. Then the quadratic functional (2.4) is approximated by the quadratic form

$$\sum_{k=0}^{n} \left\{ P_k \left(\frac{y_{k+1} - y_k}{\Delta t} \right)^2 + R_k y_k^2 \right\} \Delta t, \tag{2.5}$$

where $P_k := P(t_k)$, $R_k := R(t_k)$, $y_k := y(t_k)$, and $y_0 = y_{n+1} = 0$. By putting $a_k := R_k\Delta t + (P_{k-1} + P_k)/\Delta t$, $b_k := -P_{k-1}/\Delta t$, $y = (y_1, \ldots, y_n)$, and A as (2.2), the quadratic form (2.5) is expressed as $y^T Ay$. Furthermore, by making the change of variables

$$Y_0 := 0, \quad Y_1 := \Delta t, \quad Y_{k+1} := \frac{(\Delta t)^{k+1}|A_k|}{P_1 \cdots P_k}, \quad k = 1, \ldots, n,$$

the recursion relation (2.3) reduces to

$$\frac{P_k \frac{Y_{i+1} - Y_k}{\Delta t} - P_{k-1} \frac{Y_k - Y_{k-1}}{\Delta t}}{\Delta t} = R_k Y_k. \tag{2.6}$$

Tending $\Delta t \to 0$ in (2.6), we get the Jacobi equation $d(P\dot{Y})/dt = RY$. Therefore, the expansion (2.1) can be regarded as the Jacobi equation for (P).

Additionally, when A_{k-1} is nonsingular, the expansion (2.1) is simplified as below.

LEMMA 2.2. *Divide A_k as follows:*

$$A_k = \begin{pmatrix} A_{k-1} & a \\ a^T & a_{kk} \end{pmatrix}. \tag{2.7}$$

If A_{k-1} is nonsingular, then it holds that $|A_k| = |A_{k-1}|(a_{kk} - a^T A_{k-1}^{-1} a)$. Furthermore, when all of $|A_1|, \ldots, |A_{k-1}|$ are positive, the necessary and sufficient condition for $|A_k|$ to be positive is that $a_{kk} - a^T A_{k-1}^{-1} a > 0$.

Proof. Our assertion follows from

$$A_k = \begin{pmatrix} I_{k-1} & 0 \\ a^T A_{k-1}^{-1} & 1 \end{pmatrix} \begin{pmatrix} A_{k-1} & 0 \\ 0 & a_{kk} - a^T A_{k-1}^{-1} a \end{pmatrix} \begin{pmatrix} I_{k-1} & A_{k-1}^{-1} a \\ 0 & 1 \end{pmatrix}, \tag{2.8}$$

where I_{k-1} denotes the $(k - 1) \times (k - 1)$ identity matrix. □

DEFINITION 2.3. *For any symmetric matrix A, we call the recursion relation on $\{y_i\}$*

$$y_k = \sum_{i=0}^{k-1} \sum_{\rho} \varepsilon(\rho) a_{i+1\rho(i+1)} a_{i+2\rho(i+2)} \cdots a_{k\rho(k)} y_i, \qquad k = 1, \ldots, n, \tag{2.9}$$

the Jacobi equation for A. We say that k is conjugate to 1 if the solution $\{y_i\}$ of the Jacobi equation with $y_0 > 0$ changes the sign from positive to nonpositive at $i = k$. Namely,

$$y_0 > 0, y_1 > 0, \ldots, y_{k-1} > 0, \text{ and } y_k \leq 0. \tag{2.10}$$

When A_{k-1} is nonsingular, $a_{kk} - a^T A_{k-1}^{-1} a$ is called the kth pivot of A for $k = 2, \ldots, n$. $|A_1| = a_{11}$ is called the first pivot.

THEOREM 2.4. *For any $n \times n$-symmetric matrix A, the following three conditions are equivalent.*

(a) *A is positive-definite.*

(b) *There is no point conjugate to 1.*

(c) *All the pivots ($k = 1, 2, \ldots, n$) are positive.*

Proof. The assertion follows from Sylvester's criterion, Lemma 2.2, and Definition 2.3. □

DEFINITION 2.5. *Let $\bar{x} \in R^n$ be any extremal for* (P)*; that is, it satisfies $f'(\bar{x}) = 0$. By taking $f''(\bar{x})$ as A in Definition* 2.3*, we define the Jacobi equation, conjugate points, and the* kth *pivot for* (P) *at $\bar{x}$.*

Combining Theorems 1.2, 2.4, and Definition 2.5, we readily get the following theorem.

THEOREM 2.6. *A sufficient condition for an extremal $\bar{x}$ to be a minimum for* (P) *is that there is no point conjugate to* 1.

REMARK 2.1. *Since each variable x_k ($1 \leq k \leq n$) plays the same role with one another in* (P)*, there is no reason to start with index* 1 *to define conjugate points. Indeed, let σ be an arbitrary permutation on $\{1, \ldots, n\}$, and denote by A^σ the matrix whose (i, j)-component is a $(\sigma(i), \sigma(j))$-component of A. Then we can define conjugate points as well as above, and Theorems* 2.4 *and* 2.6 *remain valid. However, concerning sufficient optimality conditions, such an extension is redundant, since it suffices to test the descending principal minors $|A_1|, \ldots, |A_n|$.*

3. Strict conjugate points: Necessity. In this section, we describe a necessary optimality condition for (P) in terms of conjugate points. Since the descending principal minors $|A_1|, \ldots, |A_n|$ are not enough to characterize $A \geq 0$, the situation is slightly different from the sufficiency case. Namely, though the implication $A \geq 0 \Rightarrow |A_k| \geq 0$ ($1 \leq k \leq n$) is true, its inverse is not true in general.

DEFINITION 3.1. *Let $A = (a_{ij})$ be an $n \times n$-symmetric matrix, and let $1 \leq i,\ j \leq n$ be two distinct integers. Then we say that j is strictly conjugate to i if there exist a permutation σ and $1 < k \leq n$ such that $\sigma(1) = i$, $\sigma(k) = j$, and if a solution $\{y_i\}$ of the Jacobi equation* (2.9) *for A^σ with $y_0 > 0$ changes the sign from nonnegative to negative at k; that is,*

$$y_0 > 0, y_1 \geq 0, \ldots, y_{k-1} \geq 0, \text{ and } y_k < 0, \tag{3.1}$$

where A^σ is defined as in Remark 2.1.

THEOREM 3.2. *Let A be a symmetric matrix. Then $A \geq 0$ if and only if there is no pair $1 \leq i,\ j \leq n$ of integers such that j is strictly conjugate to i.*

Proof. It is well known that $A \geq 0$ if and only if all of the principal minors of A are nonnegative; see, e.g., [5]. Hence our assertion is a direct consequence of Definition 3.1. □

DEFINITION 3.3. *Let $\bar{x} \in R^n$ be any extremal for* (P)*, and take $f''(\bar{x})$ as A. Then we say that j is conjugate to i at $\bar{x}$ if j is conjugate to i in the sense of Definition* 3.1.

Combining Theorems 1.2, 3.2, and Definition 3.3, we readily get the following theorem.

THEOREM 3.4. *A necessary condition for an extremal $\bar{x}$ to be a minimum for* (P) *is that there is no pair $1 \leq i,\ j \leq n$ of positive integers such that j is strictly conjugate to i.*

We close this section with showing the relationship between conjugate points and strict conjugate points.

PROPOSITION 3.5. *If j is strictly conjugate to i, then there exist a permutation σ and $m \geq 1$ such that m is conjugate to* 1 *concerning the matrix A^σ defined in Remark* 2.1.

Proof. By definition, there exist a permutation σ and $1 < k \leq n$ such that $\sigma(1) = i$, $\sigma(k) = j$, and a solution $\{y_i\}$ of the Jacobi equation (2.9) for A^σ with $y_0 > 0$ satisfies (3.1). Putting $m := \min\{1 \leq \ell \leq n\ :\ y_\ell \leq 0\}$, we have $y_0 > 0, \ldots, y_{m-1} > 0$, and $y_m \leq 0$. □

4. Examples. In this section, we give three examples and show that the preceding results work very well. Each example is derived from the classical shortest path problem on a surface S in R^3 by approximating the arc $X(t) = (x(t), y(t), z(t)) \in S$ by a polygonal curve. This approximation is done by the following procedure.

(i) Introduce the points $0 = t_0 < t_1 < \cdots < t_n < t_{n+1} = T$, and divide the interval $[0, T]$ into $n+1$ equal parts of length $\Delta t := T/(n+1)$.
(ii) For each feasible arc $X(t)$, put $X_k := X(t_k)$.
(iii) Approximate the length of the arc $X(t)$ by the length of the polygonal curve joining $n+2$ points $X_0, X_1, \ldots, X_{n+1}$.
(iv) Minimize the length of the polygonal curve.

EXAMPLE 4.1. *This example is same as Example* 1.1. *Here we compute conjugate points for the finite-dimensional analogue. By means of the spherical coordinates, any point on the* kth *longitude ℓ_k is expressed as*

$$X_k = (\sin\theta(t_k)\cos t_k, \sin\theta(t_k)\sin t_k, \cos\theta(t_k)).$$

Hence the minimization problem of the length of the polygonal arc joining $n+2$ points $A = X_0,\ X_1, \ldots, X_{n+1} = B$ is formulated as follows:

$$(\mathrm{P}_1) \quad \text{Min} \quad f(\theta_1, \ldots, \theta_n) := \sum_{k=0}^{n} \sqrt{2(1 - \cos\Delta t \sin\theta_{k+1}\sin\theta_k - \cos\theta_{k+1}\cos\theta_k)},$$

where $\theta_0 = \theta_{n+1} = \pi/2$. The variable $\theta \in R^n$ that corresponds to the equatorial arc $\bar{\theta}(t) \equiv \pi/2$ is $\bar{\theta} := (\pi/2, \ldots, \pi/2)$. Then the Hesse matrix of f at $\bar{\theta}$ is

$$f''(\bar{\theta}) = \frac{1}{\sqrt{2(1-c)}} \begin{pmatrix} 2c & -1 & & \\ -1 & 2c & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2c \end{pmatrix}, \tag{4.1}$$

where $c := \cos\Delta t$. *It follows from Lemma* 4.1 *below that the principle minor of size* k *of* (4.1) *is positive if* $(k+1)\Delta t < \pi$. *Since* $\Delta t = T/(n+1)$, *we conclude that*

(a) *when* $T < \pi$, *there is no point conjugate to* 1, *and*

(b) *when* $T \geq \pi$, *the first number* k *satisfying* $(k+1)/(n+1) \geq \pi/T$ *is conjugate to* 1,

which matches the classical result.

The following lemma is easily proved by induction.

LEMMA 4.1. *For* $c = \cos\Delta t$, *define* $k \times k$*-matrices*

$$A_k := \begin{pmatrix} 2c & -1 & & \\ -1 & 2c & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2c \end{pmatrix}, \qquad B_k := \begin{pmatrix} c & -1 & & \\ -1 & 2c & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2c \end{pmatrix}. \tag{4.2}$$

Then their determinants are given by $|A_k| = \sin(k+1)\Delta t/\sin\Delta t$ *and* $|B_k| = \cos k\Delta t$, *respectively.*

EXAMPLE 4.2. *This example is a finite-dimensional analogue to the initial-free shortest path problem on the unit sphere* S, *that is, finding a shortest path joining the initial longitude* $\ell_0 := \{(\sin\alpha, 0, \cos\alpha) : \ 0 \leq \alpha \leq \pi\}$ *and the point* $B = (\cos T, \sin T, 0)$.

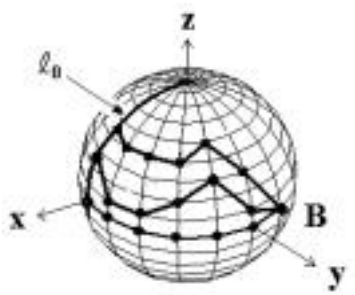


FIG. 4.1.

As well as in Example 4.1, *the finite-dimensional analogue is formulated as*

$$(\mathrm{P}_2) \quad \mathrm{Min} \quad f(\theta_0, \theta_1, \ldots, \theta_n) := \sum_{k=0}^{n} \sqrt{2(1 - \cos\Delta t \sin\theta_{k+1} \sin\theta_k - \cos\theta_{k+1}\cos\theta_k)},$$

where $\theta_{n+1} := \pi/2$. *The variable* $\theta \in R^{n+1}$ *that corresponds to the equatorial arc* $\bar\theta(t) \equiv \pi/2$ *is* $\bar\theta := (\pi/2, \ldots, \pi/2)$. *Then*

$$f''(\bar\theta) = \frac{1}{\sqrt{2(1-c)}} \left.\begin{pmatrix} c & -1 & & \\ -1 & 2c & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2c \end{pmatrix}\right\} n+1, \tag{4.3}$$

where $c := \cos\Delta t$. *It follows from Lemma* 4.1 *that the principle minor of size* k *of* (4.3) *is positive if* $k\Delta t < \pi/2$. *Since* $\Delta t = T/(n+1)$, *we conclude that*

(a) *when* $T < \pi/2$, *there is no point conjugate to* 1, *and*
(b) *when* $T \geq \pi/2$, *the first number* k *satisfying* $kT/(n+1) \geq \pi/2$ *is conjugate to* 1,

which matches the classical result.

EXAMPLE 4.3. *The original variational problem is finding a shortest path on the cylinder* $S := \{(x, y, z) : x^2 + y^2 = 1,\ z \in R\}$ *joining* $A = (1, 0, 0)$ *and* $B = (\cos T, \sin T, \gamma)$, *where* $\gamma \in R$ *is given.*

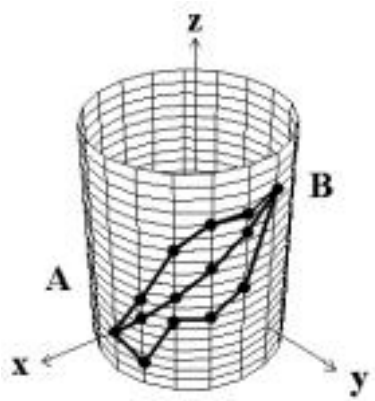


FIG. 4.2.

By applying the procedure (i)–(iv) *above to it, we get its finite-dimensional analogue*

$$(\mathrm{P}_3) \quad \text{Min} \quad f(z_1, \ldots, z_n) := \sum_{k=0}^{n} \sqrt{(z_{k+1} - z_k)^2 + 4 \sin^2 \frac{\Delta t}{2}},$$

where $z_0 := 0,\ z_{n+1} := \gamma$. *Furthermore, the variable* $z \in R^n$ *that corresponds to the spiral* $\bar{z}(t) = t\gamma/T$ *is* $\bar{z} := (\gamma/(n+1), \ldots, n\gamma/(n+1))$. *Then the Hesse matrix of* f *at* $\bar{z}$ *is*

$$f''(\bar{z}) = s(s+d)^{-3/2} \begin{pmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix}, \tag{4.4}$$

where $s := 4 \sin^2(\Delta t/2)$ *and* $d := \gamma^2/(n+1)^2$. *Since* $|A_k| = \{s(s+d)^{-3/2}\}^k (k+1) > 0$, *there is no point conjugate to* 1, *which matches the classical result.*

Acknowledgment. The author would like to thank the referees for their valuable comments and suggestions.

REFERENCES

[1] I. M. GELFAND AND S. V. FOMIN, *Calculus of Variations*, Prentice Hall, Englewood Cliffs, NJ, 1963.

[2] H. Kawasaki, *Conjugate points for a minimization problem in a finite dimensional space*, Kyushu University Preprint Series in Mathematics, 23 (1999), pp. 1–12.
[3] H. Kawasaki and V. Zeidan, *Conjugate points for variational problems with equality and inequality state constraints*, SIAM J. Control Optim., 39 (2000) pp. 433–456.
[4] P. D. Loewen and H. Zheng, *Generalized conjugate points for optimal control problems*, Nonlinear Anal., 22 (1994), pp. 771–791.
[5] G. Strang, *Linear Algebra and its Applications*, Academic Press, New York, 1976.
[6] J. Warga, *A second-order Lagrangian condition for restricted control problems*, J. Optim. Theory Appl., 24 (1978), pp. 475–483.
[7] V. Zeidan, *The Riccati equation for optimal control problems with mixed state-control constraints: Necessity and sufficiency*, SIAM J. Control Optim., 32 (1994), pp. 1297–1321.
[8] V. Zeidan, *Admissible directions and generalized coupled points for optimal control problems*, Nonlinear Anal., 26 (1996), pp. 479–507.
[9] V. Zeidan and P. Zezza, *The conjugate point condition for smooth control sets*, J. Math. Anal. Appl., 132 (1988), pp. 572–589.
[10] V. Zeidan and P. Zezza, *Coupled points in the calculus of variations and applications to periodic problems*, Trans. Amer. Math. Soc., 315 (1989), pp. 323–335.
[11] V. Zeidan and P. Zezza, *Coupled points in optimal control theory*, IEEE Trans. Automat. Control, 36 (1991), pp. 1276–1281.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 1, pp. 64–87

STOCK TRADING: AN OPTIMAL SELLING RULE*

Q. ZHANG†

Abstract. Trading in stock markets consists of three major steps: select a stock, purchase a number of shares, and eventually sell them to make a profit. The timing to buy and sell is extremely crucial. A selling rule can be specified by two preselected levels: a target price and a stop-loss limit. This paper is concerned with an optimal selling rule based on the model characterized by a number of geometric Brownian motions coupled by a finite-state Markov chain. Such a policy can be obtained by solving a set of two-point boundary value differential equations. Moreover, the corresponding expected target period and probability of making money and that of losing money are derived. Analytic solutions are obtained in one- and two-dimensional cases. Finally, a numerical example is considered to demonstrate the effectiveness of our method.

Key words. optimal selling rule, geometric Brownian motion, Markov switching, two-point boundary value problem

AMS subject classifications. 91B26, 91B28, 91B70

PII. S0363012999356325

1. Introduction.

Trading in stocks consists of three major steps: (a) select a stock based on certain criteria; (b) buy a number of shares at the right time (usually associated with the so-called pivot points; see Livermore [14] and O'Neil [18]); (c) hold the position for a period of time and then sell it to make a profit. A selling (liquidation) decision can be made when the price of the underlying stock reaches a target price or a stop-loss limit.

To analyze and study the performance of a stock, it is important to establish a mathematical model to characterize its price movement. In mathematical finance, the price of a stock is often modeled as a geometric Brownian motion (see Merton [15]) which is determined by two parameters: the expected return and volatility (see Elliott and Kopp [7], Karatzas [11], and Karatzas and Shreve [12] for analysis of the model and applications). Such parameters are usually assumed to be deterministic when analyzing option pricing; see Duffie [6] and Hull [10]. As a result, such a model is good only for a relatively short period because it wouldn't respond to random changes in these parameters. Some modifications with random parameters are available in the literature in which the volatility is dictated by additional stochastic differential equations; see Fouque, Papanicolaou, and Ronnie [9], Hull [10], and Musiela and Rutkowski [16], among others, for related results.

A major factor that dominates the movement of an individual stock is the trends of the general market. If the overall market moves up, most stocks go up; if the general market goes down, most follow. By and large, the movements of a market can be viewed as a composition of a primary movement and secondary movement. Such a classification is summarized in Table 1.

In order to incorporate the broad trend of a stock market, it is necessary to revise and modify the geometric Brownian motion model to allow the expected return and volatility parameters to depend on general market movements. In view of this, it is

*Received by the editors May 17, 1999; accepted for publication (in revised form) October 10, 2000; published electronically May 3, 2001. This work was supported in part by ONR grant N00014-96-1-0263.

http://www.siam.org/journals/sicon/40-1/35632.html

†Department of Mathematics, Boyd Graduate Studies Research Center, University of Georgia, Athens, GA 30602 (qingz@math.uga.edu).

TABLE 1
Major movements of a stock market.

Market movements	Market trends	Duration
Primary	Broad upward or downward	Several years
Secondary	Significant decline in a great bull market or strong recovery in a great bear market	Several weeks to a few months

natural to introduce a finite-state Markov chain $\alpha(\cdot)$ representing the general market direction. For example, one may consider $\alpha(t) = (\alpha_1(t), \alpha_2(t))$, where $\alpha_1(t) \in \{1, 2\}$ represents the primary market trend (here the state 1 represents up-trend and 2 down-trend) and $\alpha_2(t) \in \{1, 2\}$ represents the secondary market movement indicator. In this case, we may consider the Markov chain

$$\alpha(t) \in \{(2,2), (1,2), (2,1), (1,1)\} \tag{1}$$

with a generator given by

$$Q = \begin{pmatrix} -\lambda_1 & \lambda_1 & 0 & 0 \\ \lambda_2 & -\lambda_2 & 0 & 0 \\ 0 & 0 & -\lambda_1 & \lambda_1 \\ 0 & 0 & \lambda_2 & -\lambda_2 \end{pmatrix} + \begin{pmatrix} -\mu_1 & 0 & \mu_1 & 0 \\ 0 & -\mu_1 & 0 & \mu_1 \\ \mu_2 & 0 & -\mu_2 & 0 \\ 0 & \mu_2 & 0 & -\mu_2 \end{pmatrix},$$

where λ_1 and μ_1 are the transition rates of going up and λ_2 and μ_2 the rates of going down. If a great bull market lasts for 5 years and a bear market for 3 years, then $\lambda_1 = 1/3$ and $\lambda_2 = 1/5$ when time is measured in years. Similarly, we can take $\mu_1 = 4$ and $\mu_2 = 12$ representing a recovery in a bear market which lasts for a quarter of a year and a decline in a bull market that runs for about 4–5 weeks. Clearly, it is more realistic to include the random process $\alpha(\cdot)$ in the model. In this paper, we consider a model with a single stock and its price observes a switching geometric Brownian motion. Moreover, the stock pays no dividends. Given the current price of a stock, a selling rule in this paper consists of a target price and a stop-loss limit. A selling decision is made whenever the price reaches either the target price or the stop-loss limit. The timing to sell is as important as that of buying. The primary goal of investing is to make a profit. However, in reality, one often picks up the wrong stock or purchases it at the wrong time. In this case, it is necessary to sell it sooner to stop loss. In practice, a target price is typically around a gain of 20% to 50% and a stop-loss limit generally varies from 10% to 30% depending on the risk an investor is willing to take. Clearly, it is not a good idea to adopt uniform profit-taking or cut-loss rates. Each stock is different and has its own trait. It should be treated differently with different liquidation rules.

In this paper, we consider a set of target prices and stop-loss limits. Our goal is to choose a target price and a stop-loss limit in that set in order to maximize an expected reward function. We aim at deriving these price limits. In addition, we obtain the expected target period (holding duration) and the probability of making money and that of losing money.

In practice, an often used criterion for measuring the performance of a portfolio is that of percentage return per unit time. However, such a criterion leads to frequent transactions because it encourages small profit taking within short holding time τ_0

(say, a few minutes to a couple of days). Clearly, such a criterion is not suitable to many individual investors because of limited time available for trading and additional transaction costs. A discounted reward, on the other hand, rules out too frequent transactions because the time factor $(1/\tau_0)$ is replaced by a discount rate $e^{-\rho\tau_0}$ (see section 6 for discussions on selection of $\rho > 0$). Such a discounted reward function is natural in many financial problems. Moreover, the resulting selling rule involves only a set of ordinary differential equations (ODEs) rather than a number of partial differential equations (PDEs), which makes the corresponding computation much easier. Our optimal selling rule can be determined by solving a set of ODEs with two-point boundary conditions. In this paper, we prove the existence and uniqueness of the solution to these equations. Examples with one-dimensional (1-D) and two-dimensional (2-D) models are considered. Analytic solutions are obtained in these cases. Finally, daily closes of the Microsoft Corp. stock in 1999 are used to demonstrate the effectiveness of our results.

The paper is organized as follows. In the next section, we formulate the optimization problem under consideration. In sections 3 and 4, we derive an optimal selling policy under the model formulated and assumptions imposed in this paper. We also obtain expected exit time and profit and loss probabilities. In section 5, analytic solutions of these policies are obtained in both 1-D and 2-D cases. A numerical example is reported in section 6. All proofs of results are postponed and given in the appendix.

2. Problem formulation. Let $\mathcal{M} = \{1, 2, \ldots, m\}$ denote the state space of the Markov chain $\alpha(\cdot)$. Note that each element in $\mathcal{M}$ is an index and may be used to represent a vector as in (1). Let $Q = (q_{ij})_{m\times m}$ be the generator of $\alpha(\cdot)$ with $q_{ij} \geq 0$ for $i \neq j$ and $\sum_{j=1}^{m} q_{ij} = 0$ for each $i \in \mathcal{M}$. In this paper, the market-trend indicator process $\alpha(t)$ is not necessarily observable. Only the generator and its initial distribution $\{p_i = P(\alpha(0) = i) \text{ for } i \in \mathcal{M}\}$ are available.

Let $S(t)$ denote the price of a stock at time t. It satisfies the equation

$$\begin{cases} dS(t) = \mu(\alpha(t))S(t)dt + \sigma(\alpha(t))S(t)dw(t), \\ S(0) = S_0,\ t \geq 0, \end{cases} \tag{2}$$

where $S_0 > 0$ is the initial price; $\mu(i)$, $i \in \mathcal{M}$, is the expected return; $\sigma(i)$, $i \in \mathcal{M}$, represents the stock volatility; and $w(\cdot)$ is a standard Brownian motion. The processes $\alpha(\cdot)$ and $w(\cdot)$ are independent. In addition, we assume $\sigma^2(i) > 0$ for $i \in \mathcal{M}$.

This paper is concerned with a stock selling rule. A commonly used selling rule in practice is of the form

$$\tau_0 = \inf\Big\{t > 0 :\ S(t) \notin (A, B)\Big\}, \tag{3}$$

i.e., to sell a stock at time τ_0 for prespecified $A \leq S_0 \leq B$.

In this paper, we study only those strategies with $A > 0$ and $B < \infty$.

The condition $A > 0$ is mainly motivated by practical considerations. In fact, the lower bound A prescribes the maximum risk level of an investment. Cutting a loss short is crucial to preserve the capital. This is especially important during a sharp market downturn. An often used cut-loss level is 10–30% for active traders, which corresponds to $A = S_0(1\text{–}10\%)$ to $A = S_0(1\text{–}30\%)$. W. J. O'Neil, the founder of *Investor's Business Daily*, suggests an even more conservative limit of 8% in [18]. In addition, Dammon and Spatt [4] have shown that an optimal selling rule should include $A = x_L > 0$ when transaction costs (commissions) and capital gain taxes are added to the picture. Similar results are obtained in Cadenillas and Pliska [2] when

$\mu > \sigma^2$ with constant μ and σ. However, from a purely mathematical point of view, it is possible to have $A = 0$ as an optimal lower bound. This is demonstrated by Øksendal [17, p. 199] using an optimal stopping approach. This condition means that an investor needs to risk the entire capital to maximize a potential future return.

The condition $B < \infty$ is imposed mainly for theoretical convenience (used in the proof of Theorem 3.2). It is shown in [17] that the optimal upper bound $B = x_0 < \infty$. In addition, even in the presence of capital gain taxes and transactions costs, the optimal B should be finite when $\sigma^2/2 < \mu < \sigma^2$ (see [2]). Furthermore, our numerical study in this paper indicates that the optimal $B < \infty$; see Remark 5.1 and Tables 6 and 8 in which the optimal $B(= S_0 e^{z_2^*})$ is bounded with fixed discount factor ρ. However, when $\mu > \sigma^2$, it is proved in [2] that the optimal $B = \infty$. Infinite upper bound is also obtained in [4]. Intuitively, the condition $B = \infty$ suggests that one should never sell his position no matter how high its price goes. However, in reality, a stock's price often rests at a certain level (or goes sideways) after substantial advances during a period of time. In this case, a predetermined $B < \infty$ (say 20–50% annual return) would help an investor lock in real profit following these price advances and move the money elsewhere for other investment opportunities. In view of this, such a predetermined rule is more desirable for a relatively short-term (several months to a year) investment.

Clearly, an optimal time to sell depends on the reward (utility) function and system parameters. It may also depend on transaction costs and capital gain taxes if these factors are included in the model. By and large, including commission and tax factors in the model typically makes transactions less frequent, which corresponds to smaller A and larger B. We refer the reader to [2], [3], and [4], among others, for further discussions on models with these factors.

Assumption (A). In this paper, we consider the selling rule (3) with

$$A_1 \le A \le A_2 \text{ and } B_1 \le B \le B_2$$

for any given $0 < A_1 < A_2 < S_0 < B_1 < B_2 < \infty$.

Our goal is to find an optimal pair (A, B) under this assumption for a given reward function.

For convenience, we choose a_1, a_2, b_1, b_2 such that

$$A_1 = S_0 e^{-b_1},\ A_2 = S_0 e^{-a_1},\ B_1 = S_0 e^{a_2},\ B_2 = S_0 e^{b_2}.$$

It is easy to see that $0 < a_1 \le b_1 < \infty$ and $0 < a_2 \le b_2 < \infty$. Moreover, choose z_1 and z_2 such that

$$A = S_0 e^{-z_1} \text{ and } B = S_0 e^{z_2}$$

for some z_1 and z_2. Let $\mathcal{I} = [a_1, b_1] \times [a_2, b_2]$. We consider $(z_1, z_2) \in \mathcal{I}$.

Here the target price is given by $B = S_0 e^{z_2}$ and the stop-loss limit is $A = S_0 e^{-z_1}$. A selling rule is determined by (z_1, z_2). For example, taking $z_1 = -\log 0.9$ and $z_2 = \log 1.2$ corresponds to a target price of 20% gain and a stop limit of 10% loss.

Let

$$X(t) = \int_0^t \left(\mu(\alpha(s)) - \frac{\sigma^2(\alpha(s))}{2}\right) ds + \int_0^t \sigma(\alpha(s)) dw(s),$$

where the stochastic integral

$$\int_0^t \sigma(\alpha(s)) dw(s) = \sum_{n=0}^{\infty} \sigma(\alpha_n)(w(t_{n+1}) - w(t_n))$$

with random jump times $0 = t_0 < t_1 < \cdots < t_n < \cdots$ of $\alpha(\cdot)$ and $\alpha(t) = \alpha_n$ for $t \in [t_n, t_{n+1})$, $n = 0, 1, 2, \ldots$. Note that $\alpha(\cdot)$ is a Markov chain with generator Q. It follows that $\lim_{n\to\infty} t_n = \infty$ with probability one; see Davis [5, p. 60]. Using $X(t)$, we can write $S(t)$ as follows:

$$S(t) = S_0 \exp X(t).$$

Moreover, τ_0 can be defined in terms of $X(\cdot)$:

$$\tau_0 = \inf\Big\{t > 0 : \ X(t) \notin (-z_1, z_2)\Big\}.$$

The objective of the problem is to find $(z_1, z_2) \in \mathcal{I}$ to maximize

$$V = V(z_1, z_2) := \sum_{i=1}^{m} p_i E[\Phi(X(\tau_0))e^{-\rho\tau_0}|\alpha(0) = i],$$

where $\rho > 0$ is a discount factor and $\Phi(x)$ is a function of x. For example, we may consider a discounted reward

$$E\left[\left(\frac{S(\tau_0) - S_0}{S_0}\right)e^{-\rho\tau_0}\right], \tag{4}$$

which corresponds to $\Phi(x) = e^x - 1$. This reward function will be used in our numerical example in section 6.

3. An optimal policy. Given $x \in [-z_1, z_2] \in \mathcal{I}$, consider the switching diffusion

$$\begin{cases} d\xi(t) = r(\alpha(t))dt + \sigma(\alpha(t))dw(t), \\ \xi(0) = x, \end{cases} \tag{5}$$

where $r(i) = \mu(i) - \sigma^2(i)/2$ for $i \in \mathcal{M}$ represents a continuously compounded return rate. It follows that

$$\xi(t) = x + X(t).$$

For $x \in [-z_1, z_2]$, we define

$$\tau(x) = \inf\Big\{t \geq 0 : \ \xi(t) \notin (-z_1, z_2)\Big\}.$$

Given $\alpha(0) = i$ and $\xi(0) = x$, let $v(x, i)$ denote the value function

$$v(x, i) = E[\Phi(\xi(\tau(x)))e^{-\rho\tau(x)}].$$

Then $\tau_0 = \tau(0)$ and $E[\Phi(X(\tau_0))e^{-\rho\tau_0}|\alpha(0) = i] = v(0, i)$. The corresponding reward function

$$V = V(z_1, z_2) = \sum_{i=1}^{m} p_i v(0, i). \tag{6}$$

In order to evaluate V, one has only to find $v(x, i)$. Formally, $v(x, i)$ satisfies the following differential equations:

$$\begin{cases} \dfrac{\sigma^2(i)}{2}\dfrac{\partial^2 v(x,i)}{\partial x^2} + r(i)\dfrac{\partial v(x,i)}{\partial x} - \rho v(x, i) + Qv(x, \cdot)(i) = 0, \\ v(-z_1, i) = \Phi(-z_1), \ v(z_2, i) = \Phi(z_2) \ \text{ for } x \in (-z_1, z_2), \ i \in \mathcal{M}, \end{cases} \tag{7}$$

where $Qf(\cdot)(i) = \sum_{j \neq i} q_{ij}(f(j) - f(i))$ for any function f on $\mathcal{M}$. In fact, if (7) has a smooth solution $v(x, i)$, then using Dynkin's formula (see Fleming and Soner [8]), we can show that $v(x, i) = E[\Phi(\xi(\tau(x)))e^{-\rho\tau(x)}]$, which in turn proves the uniqueness of the solution.

Remark 3.1. Solving the differential equations in (7) consists of a two-point boundary value (TPBV) problem. In general, such a problem may not have a solution. For example, the equation $\ddot{y} + y = 0$ with $y(0) = 0$ and $y(\pi) = 1$ does not have a solution.

Let $C^2[-z_1, z_2]$ denote the space of functions that are twice continuously differentiable on $[-z_1, z_2]$. The next theorem is concerned with the existence of a C^2 solution and optimal (z_1, z_2). All proofs of results are given in the appendix.

THEOREM 3.2. *Under Assumption* (A)*, the following assertions hold:*

(a) *For each* $i \in \mathcal{M}$, $v(x, i) \in C^2[-z_1, z_2]$ *and is the unique solution to* (7).

(b) *For each fixed* (x, i), $v(x, i)$ *is a continuous function of* (z_1, z_2) *on* $\mathcal{I}$.

(c) *There exists an optimal selling policy determined by* (z_1^*, z_2^*) *with the target price* $S_0 e^{z_2^*}$ *and stop-loss limit* $S_0 e^{-z_1^*}$.

Remark 3.3. One may also consider a model involving a time variable in the differential equation. In this case, the class of ODEs in (7) becomes a set of PDEs which is much more difficult to deal with. From the computational point of view, it is easier to work with an ODE than with a PDE. In addition, the existence of a time dependent v typically requires $\Phi \in C^3$; see Krylov [13]. Such a condition is not suitable when evaluating exit probabilities in the next section because the corresponding $\Phi(x)$ is merely Borel measurable; see Remark 4.4.

4. Expected exit time and probabilities. In this section, we compute the expected holding time $E\tau_0$, the profit probability $P(S(\tau_0) \geq S_0 e^{z_2^*})$, and the loss probability $P(S(\tau_0) \leq S_0 e^{-z_1^*})$, where (z_1^*, z_2^*) is determined by the optimal policy given in Theorem 3.2.

Expected Holding Time $E\tau_0$. We first consider τ_0. Given z_1 and z_2, define

$$T(x, i) = E[\tau(x) | \xi(0) = x, \alpha(0) = i].$$

The next lemma is concerned with the uniform boundedness of T.

LEMMA 4.1. *Given* $(z_1, z_2) \in \mathcal{I}$, *there exists a constant* K *such that*

$$T(x, i) \leq K$$

for all $x \in [-z_1, z_2]$ *and* $i \in \mathcal{M}$.

The boundedness of T and Dynkin's formula lead to the following theorem.

THEOREM 4.2. *For each* $i \in \mathcal{M}$, $T(x, i) \in C^2[-z_1, z_2]$ *and is the unique solution to the following equation:*

$$\begin{cases} \dfrac{\sigma^2(i)}{2}\dfrac{\partial^2 T(x,i)}{\partial x^2} + r(i)\dfrac{\partial T(x,i)}{\partial x} + QT(x,\cdot)(i) + 1 = 0, \\ T(-z_1, i) = T(z_2, i) = 0 \text{ for } x \in (-z_1, z_2),\ i \in \mathcal{M}. \end{cases} \tag{8}$$

The expected exit time is given by

$$E\tau_0 = \sum_{i=1}^{m} p_i E[\tau_0 | \alpha(0) = i] = \sum_{i=1}^{m} p_i E[\tau(0) | \xi(0) = 0, \alpha(0) = i] = \sum_{i=1}^{m} p_i T(0, i).$$

Remark 4.3. Note that the ODEs in (8) corresponds to the HJB equation in a "control" problem with the cost function

$$J(x,i) = E\left[\int_0^{\tau(x)} dt \Big| \xi(0) = x, \alpha(0) = i\right].$$

Of course, there is no control variable involved.

Profit Probability $P(S(\tau_0) \geq S_0 e^{z_2})$. Let

$$P_1(x,i) = P(\xi(\tau(x)) \geq z_2 | \xi(0) = x, \alpha(0) = i).$$

Then it is easy to see that

$$P(S(\tau_0) \geq S_0 e^{z_2} | \alpha(0) = i) = P_1(0,i).$$

Moreover, P_1 should satisfy the following equations:

$$\text{(9)} \qquad \begin{cases} \dfrac{\sigma^2(i)}{2}\dfrac{\partial^2 P_1(x,i)}{\partial x^2} + r(i)\dfrac{\partial P_1(x,i)}{\partial x} + QP_1(x,\cdot)(i) = 0, \\ P_1(-z_1,i) = 0,\ P_1(z_2,i) = 1 \text{ for } x \in (-z_1, z_2),\ i \in \mathcal{M}. \end{cases}$$

Remark 4.4. Similarly, the ODEs in (9) correspond to the HJB equation of a control problem with the cost function

$$J(x,i) = E\left[e^{-\rho\tau}\Phi(\xi(\tau)) \Big| \xi(0) = x, \alpha(0) = i\right]$$

with no (or fixed) control variable. Here $\Phi(x) = I_{\{x=z_2\}}$, which is *not* differentiable.

THEOREM 4.5. *For each $i \in \mathcal{M}$, $P_1(x,i) \in C^2[-z_1, z_2]$ and is the unique solution to* (9).

The profit probability is

$$P_1^* := P(S(\tau_0) \geq S_0 e^{z_2}) = \sum_{i=1}^m p_i P_1(0,i).$$

Loss Probability $P(S(\tau_0) \leq S_0 e^{-z_1})$. Similarly, let

$$P_2(x,i) = P(\xi(\tau(x)) \leq -z_1 | x(0) = x, \alpha(0) = i).$$

Then we have

$$P(S(\tau_0) \leq S_0 e^{-z_1} | \alpha(0) = i) = P_2(0,i).$$

Moreover, P_2 satisfies

$$\text{(10)} \qquad \begin{cases} \dfrac{\sigma^2(i)}{2}\dfrac{\partial^2 P_2(x,i)}{\partial x^2} + r(i)\dfrac{\partial P_2(x,i)}{\partial x} + QP_2(x,\cdot)(i) = 0, \\ P_2(-z_1,i) = 1,\ P_2(z_2,i) = 0 \text{ for } x \in (-z_1, z_2),\ i \in \mathcal{M}. \end{cases}$$

THEOREM 4.6. *For each $i \in \mathcal{M}$, $P_2(x,i) \in C^2[-z_1, z_2]$ and is the unique solution to* (10).

Then the loss probability is

$$P_2^* := P(S(\tau_0) \leq S_0 e^{-z_1}) = \sum_{i=1}^m p_i P_2(0,i).$$

5. Analytic solutions with $m = 1, 2$. In this section, we consider the simple cases with $m = 1$ and $m = 2$. We aim at deriving analytic solutions.

1-D Case ($m = 1$). We consider the 1-D case, i.e., $m = 1$. Let $\sigma = \sigma(1)$ and $r = r(1)$. We assume $r \neq 0$. Then, the differential equations in (7) become

$$\begin{cases} \dfrac{\sigma^2}{2}\ddot{v}(x) + r\dot{v}(x) - \rho v(x) = 0, \ x \in (-z_1, z_2), \\ v(-z_1) = \Phi(-z_1), \ v(z_2) = \Phi(z_2), \end{cases}$$

where $\dot{f}(x) = df(x)/dx$ and $\ddot{f}(x) = d^2 f(x)/dx^2$. Solve this equation to obtain

$$\begin{aligned} v(x) = & \left(\frac{e^{\eta_1 z_1 - \eta_2(z_1+z_2)}\Phi(z_2) - e^{\eta_1 z_1}\Phi(-z_1)}{e^{(\eta_1-\eta_2)(z_1+z_2)} - 1} \right) e^{\eta_1 x} \\ & + \left(\frac{-e^{-\eta_2 z_2}\Phi(z_2) + e^{\eta_1 z_1 + (\eta_1-\eta_2)z_2}\Phi(-z_1)}{e^{(\eta_1-\eta_2)(z_1+z_2)} - 1} \right) e^{\eta_2 x}, \end{aligned}$$

where

$$\begin{cases} \eta_1 = \dfrac{-r + \sqrt{r^2 + 2\rho\sigma^2}}{\sigma^2}, \\ \eta_2 = \dfrac{-r - \sqrt{r^2 + 2\rho\sigma^2}}{\sigma^2}. \end{cases}$$

The objective is to choose $(z_1, z_2) \in \mathcal{I}$ to maximize

$$V = v(0) = \frac{(e^{\eta_1 z_1 + (\eta_1-\eta_2)z_2} - e^{\eta_1 z_1})\Phi(-z_1) + (e^{(\eta_1-\eta_2)z_1 - \eta_2 z_2} - e^{-\eta_2 z_2})\Phi(z_2)}{e^{(\eta_1-\eta_2)(z_1+z_2)} - 1}.$$

We next compute $E\tau_0$. The corresponding differential equation is

$$\begin{cases} \dfrac{\sigma^2}{2}\ddot{T}(x) + r\dot{T}(x) + 1 = 0, \\ T(-z_1) = T(z_2) = 0. \end{cases}$$

Let $\eta_0 = -2r/\sigma^2$. Then

$$T(x) = \frac{z_1 + z_2}{r(e^{\eta_0 z_2} - e^{-\eta_0 z_1})} e^{\eta_0 x} - \frac{x}{r} - \frac{z_1 e^{\eta_0 z_2} + z_2 e^{-\eta_0 z_1}}{r(e^{\eta_0 z_2} - e^{-\eta_0 z_1})}.$$

It follows that

$$E\tau_0 = T(0) = \frac{z_1 + z_2}{r(e^{\eta_0 z_2} - e^{-\eta_0 z_1})} - \frac{z_1 e^{\eta_0 z_2} + z_2 e^{-\eta_0 z_1}}{r(e^{\eta_0 z_2} - e^{-\eta_0 z_1})}.$$

To compute $P(S(\tau_0) \geq S_0 e^{z_2})$ and $P(S(\tau_0) \leq S_0 e^{-z_1})$, we solve the corresponding equations (9) and (10) and obtain

$$\begin{cases} P_1(x) = \dfrac{e^{-\eta_0 z_1} - e^{\eta_0 x}}{e^{-\eta_0 z_1} - e^{\eta_0 z_2}}, \\ P_2(x) = \dfrac{e^{\eta_0 x} - e^{\eta_0 z_2}}{e^{-\eta_0 z_1} - e^{\eta_0 z_2}}. \end{cases}$$

TABLE 2
Cases when $\rho = 4$ and $\sigma = 0.3$.

r	0.1	0.2	0.3	0.4	0.5
$(-z_1^*, z_2^*)$	(−0.36,0.13)	(−0.36,0.14)	(−0.36,0.16)	(−0.36,0.19)	(−0.36,0.21)
$E\tau_0$	0.46	0.43	0.43	0.42	0.40
P_1^*	0.83	0.90	0.94	0.97	0.98
P_2^*	0.17	0.10	0.06	0.03	0.02

TABLE 3
Cases when $\rho = 4$ and $r = 0.15$.

σ	0.1	0.2	0.3	0.4	0.5
$(-z_1^*, z_2^*)$	(−0.36,0.06)	(−0.36,0.09)	(−0.36,0.13)	(−0.36,0.16)	(−0.36,0.20)
$E\tau_0$	0.39	0.51	0.43	0.34	0.27
P_1^*	1.00	0.97	0.87	0.79	0.72
P_2^*	0.00	0.03	0.12	0.21	0.28

TABLE 4
Dependence of (z_1^*, z_2^*) on b_1.

b_1	0.5	1	2	3	4	5
$(-z_1^*, z_2^*)$	$(-0.5, 0.30)$	$(-1, 0.30)$	$(-2, 0.30)$	$(-2.81, 0.30)$	$(-3.77, 0.30)$	$(-4.06, 0.31)$

Setting $x = 0$, we have

$$\begin{cases} P_1^* = P_1(0) = \dfrac{e^{-\eta_0 z_1} - 1}{e^{-\eta_0 z_1} - e^{\eta_0 z_2}}, \\ P_2^* = P_2(0) = \dfrac{1 - e^{\eta_0 z_2}}{e^{-\eta_0 z_1} - e^{\eta_0 z_2}}. \end{cases}$$

If we choose $\mathcal{I} = [0.01, 0.36] \times [0.01, 2.3]$, then $b_1 = 0.36$ which limits the maximum risk to 30% and $b_2 = 2.3$ which indicates a maximal 900% return.

Intuitively, stocks with larger expected return rate r correspond to bigger z_2^*, with shorter holding time and higher probability of making profit. This can be seen from Table 2 with fixed $\rho = 4$ and $\sigma = 0.3$.

On the other hand, for stocks with higher volatility, the corresponding z_2^* should be bigger to create more room for higher returns. In the meantime, this brings with it higher risk and therefore smaller probability of making money. This can be seen from Table 3.

Note that $z_1^* = 0.36$ in all of these cases. This means in general an optimal cut-loss level (30%) should be the maximum risk one is willing to take.

Remark 5.1. In addition, we also examine the dependence of (z_1^*, z_2^*) on the choice of b_1 and b_2. For this purpose, we choose $\rho = 4$, $r = 0.75$, and $\sigma = 0.38$ (which will be used in section 6). According to our numerical tests, if $b_2 \leq 0.30$, then $z_2^* = b_2$. For $b_2 > 0.30$, $z_2^* = 0.30$. Therefore, the optimal level z_2^* is mainly determined by other parameters, such as $r(i)$ and $\sigma(i)$, which is not sensitive to the choice of b_2 when it is greater than 0.3. Fixing $b_2 = 5$, we next examine the dependence of (z_1^*, z_2^*) on b_1. The results are summarized in Table 4.

These results tell us that $z_1^* \sim b_1$; it means the larger the b_1, the higher the risk is allowed, and that usually yields better expected return.

2-D Case ($m = 2$). The 2-D case ($m = 2$) corresponds to the model with Markov switching. Let the generator of $\alpha(\cdot)$ have the form

$$Q = \begin{pmatrix} -\lambda_1 & \lambda_1 \\ \lambda_2 & -\lambda_2 \end{pmatrix}$$

for $\lambda_1 > 0$ and $\lambda_2 > 0$. The corresponding stationary distribution is

$$(\nu_1, \nu_2) = \left(\frac{\lambda_2}{\lambda_1 + \lambda_2}, \frac{\lambda_1}{\lambda_1 + \lambda_2} \right).$$

In this example, we consider the market with two trends: up and down. Here $r(1) > 0$ corresponds to the up-trend return rate and $r(2) < 0$ the down-trend return rate. Let $\tilde{r} = r(1)\lambda_2 + r(2)\lambda_1$. Then the long-term average rate is given by

$$r(1)\nu_1 + r(2)\nu_2 = (\lambda_1 + \lambda_2)^{-1}\tilde{r} \tag{11}$$

which is typically greater than 0 in practice.

The corresponding differential equations are given by

$$\begin{cases} \dfrac{\sigma^2(1)}{2}\dfrac{\partial^2 v(x,1)}{\partial x^2} + r(1)\dfrac{\partial v(x,1)}{\partial x} - \rho v(x,1) + \lambda_1(v(x,2) - v(x,1)) = 0, \\ \dfrac{\sigma^2(2)}{2}\dfrac{\partial^2 v(x,2)}{\partial x^2} + r(2)\dfrac{\partial v(x,2)}{\partial x} - \rho v(x,2) + \lambda_2(v(x,1) - v(x,2)) = 0, \\ v(-z_1, 1) = \Phi(-z_1),\ v(-z_1, 2) = \Phi(-z_1), \\ v(z_2, 1) = \Phi(z_2),\ v(z_2, 2) = \Phi(z_2) \end{cases} \tag{12}$$

for $-z_1 < x < z_2$. Using the first equation, we obtain

$$v(x,2) = \frac{1}{\lambda_1}\left(-\frac{\sigma^2(1)}{2}\frac{\partial^2 v(x,1)}{\partial x^2} - r(1)\frac{\partial v(x,1)}{\partial x} + (\rho + \lambda_1)v(x,1) \right).$$

Substituting this into the second equation leads to

$$\begin{aligned} &\left(\frac{\sigma^2(1)\sigma^2(2)}{4}\right)\frac{\partial^4 v(x,1)}{\partial x^4} + \left(\frac{\sigma^2(1)r(2) + \sigma^2(2)r(1)}{2}\right)\frac{\partial^3 v(x,1)}{\partial x^3} \\ &\quad + \left(r(1)r(2) - \frac{\sigma^2(1)(\rho + \lambda_2) + \sigma^2(2)(\rho + \lambda_1)}{2} \right)\frac{\partial^2 v(x,1)}{\partial x^2} \\ &\quad - \Big(r(1)(\rho + \lambda_2) + r(2)(\rho + \lambda_1) \Big)\frac{\partial v(x,1)}{\partial x} + (\rho^2 + \rho(\lambda_1 + \lambda_2))v(x,1) = 0. \end{aligned}$$

The corresponding characteristic equation is given by

$$\begin{aligned} \psi(\eta) = \ & \frac{\sigma^2(1)\sigma^2(2)}{4}\eta^4 + \frac{\sigma^2(1)r(2) + \sigma^2(2)r(1)}{2}\eta^3 \\ & + \left(r(1)r(2) - \frac{\sigma^2(1)(\rho + \lambda_2) + \sigma^2(2)(\rho + \lambda_1)}{2} \right)\eta^2 \\ & - \Big(r(1)(\rho + \lambda_2) + r(2)(\rho + \lambda_1) \Big)\eta + \rho^2 + \rho(\lambda_1 + \lambda_2) = 0. \end{aligned} \tag{13}$$

It is easy to show that ψ can be written as

$$\begin{aligned} \psi(\eta) = \frac{\sigma^2(1)\sigma^2(2)}{4}\Bigg\{ & \left(\eta^2 + \frac{2r(1)}{\sigma^2(1)}\eta - \frac{2(\rho + \lambda_1)}{\sigma^2(1)} \right) \\ & \times \left(\eta^2 + \frac{2r(2)}{\sigma^2(2)}\eta - \frac{2(\rho + \lambda_2)}{\sigma^2(2)} \right) - \frac{4\lambda_1\lambda_2}{\sigma^2(1)\sigma^2(2)} \Bigg\}. \end{aligned}$$

Note that $\psi(\infty) = \psi(-\infty) = \infty$ and $\psi(0) > 0$. Moreover, let $\eta^+ > 0$ and $\eta^- < 0$ denote the zeros of

$$\eta^2 + \frac{2r(1)}{\sigma^2(1)}\eta - \frac{2(\rho+\lambda_1)}{\sigma^2(1)}.$$

One may also take $\eta^+ > 0$ and $\eta^- < 0$ to be the zeros of

$$\eta^2 + \frac{2r(2)}{\sigma^2(2)}\eta - \frac{2(\rho+\lambda_2)}{\sigma^2(2)}.$$

Then both $\psi(\eta^+)$ and $\psi(\eta^-)$ are less than zero. Therefore, in view of the intermediate value property, $\psi(\eta)$ should have four real zeros denoted by η_i, $i = 1, 2, 3, 4$. Hence, there are constants c_i, $i = 1, 2, 3, 4$, such that

$$\begin{cases} v(x,1) = \sum_{i=1}^{4} c_i e^{\eta_i(x+z_1)}, \\ v(x,2) = \sum_{i=1}^{4} c_i \kappa_i e^{\eta_i(x+z_1)}, \end{cases}$$

where

$$\kappa_i = \frac{1}{\lambda_1}\left(-\frac{\sigma^2(1)}{2}\eta_i^2 - r(1)\eta_i + \rho + \lambda_1\right).$$

Using the initial conditions, we have

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ \kappa_1 & \kappa_2 & \kappa_3 & \kappa_4 \\ e^{\eta_1(z_1+z_2)} & e^{\eta_2(z_1+z_2)} & e^{\eta_3(z_1+z_2)} & e^{\eta_4(z_1+z_2)} \\ \kappa_1 e^{\eta_1(z_1+z_2)} & \kappa_2 e^{\eta_2(z_1+z_2)} & \kappa_3 e^{\eta_3(z_1+z_2)} & \kappa_4 e^{\eta_4(z_1+z_2)} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{pmatrix} = \begin{pmatrix} \Phi(-z_1) \\ \Phi(-z_1) \\ \Phi(z_2) \\ \Phi(z_2) \end{pmatrix}.$$

In view of Theorem 3.2, the above equation has a unique solution (c_1, c_2, c_3, c_4).

The objective is to choose (z_1, z_2) to maximize

$$V := p_1 v(0,1) + p_2 v(0,2) = \sum_{i=1}^{4} c_i(p_1 + \kappa_i p_2) e^{\eta_i z_1}.$$

We next compute $T(x,i)$. The corresponding differential equations are given by

$$\begin{cases} \frac{\sigma^2(1)}{2}\frac{\partial^2 T(x,1)}{\partial x^2} + r(1)\frac{\partial T(x,1)}{\partial x} + \lambda_1(T(x,2) - T(x,1)) + 1 = 0, \\ \frac{\sigma^2(2)}{2}\frac{\partial^2 T(x,2)}{\partial x^2} + r(2)\frac{\partial T(x,2)}{\partial x} + \lambda_2(T(x,1) - T(x,2)) + 1 = 0, \\ T(-z_1,1) = T(-z_1,2) = T(z_2,1) = T(z_2,2) = 0 \text{ for } -z_1 < x < z_2. \end{cases} \tag{14}$$

Following a similar procedure as in solving (12), we can show that the corresponding characteristic function $\psi^0(\eta)$ is identical to the one in (13) with $\rho = 0$. In view of (11), we consider only the case when $\tilde{r} := r(1)\lambda_2 + r(2)\lambda_1 \neq 0$. Note that $\dot{\psi}^0(0) = -\tilde{r}$. Therefore, ψ^0 has zeros

$$\begin{array}{l} \eta_1^0 < 0,\ \eta_2^0 < 0,\ \eta_3^0 > 0,\ \eta_4^0 = 0 \ \text{ if } \tilde{r} > 0, \\ \eta_1^0 < 0,\ \eta_2^0 > 0,\ \eta_3^0 > 0,\ \eta_4^0 = 0 \ \text{ if } \tilde{r} < 0. \end{array}$$

Thus we have

$$\begin{cases} T(x,1) = \sum_{i=1}^{3} c_i e^{\eta_i^0(x+z_1)} + c_4 - \dfrac{(\lambda_1+\lambda_2)x}{\tilde{r}}, \\ T(x,2) = \sum_{i=1}^{3} c_i \kappa_i^0 e^{\eta_i^0(x+z_1)} + c_4 - \dfrac{(\lambda_1+\lambda_2)x - (r(1)-r(2))}{\tilde{r}}, \end{cases}$$

where

$$\kappa_i^0 = \frac{1}{\lambda_1}\left(-\frac{\sigma^2(1)}{2}(\eta_i^0)^2 - r(1)\eta_i^0 + \lambda_1\right),\ i = 1,2,3.$$

In view of the initial conditions in (14), the following equation has a unique solution:

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ \kappa_1^0 & \kappa_2^0 & \kappa_3^0 & 1 \\ e^{\eta_1^0(z_1+z_2)} & e^{\eta_2^0(z_1+z_2)} & e^{\eta_3^0(z_1+z_2)} & 1 \\ \kappa_1^0 e^{\eta_1^0(z_1+z_2)} & \kappa_2^0 e^{\eta_2^0(z_1+z_2)} & \kappa_3^0 e^{\eta_3^0(z_1+z_2)} & 1 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{pmatrix} = \frac{1}{\tilde{r}} \begin{pmatrix} -(\lambda_1+\lambda_2)z_1 \\ -(\lambda_1+\lambda_2)z_1 - (r(1)-r(2)) \\ (\lambda_1+\lambda_2)z_2 \\ (\lambda_1+\lambda_2)z_2 - (r(1)-r(2)) \end{pmatrix}.$$

Hence the expected exit time is given by

$$E\tau_0 = p_1 T(0,1) + p_2 T(0,2) = \sum_{i=1}^{3} c_i(p_1 + \kappa_i^0 p_2)e^{\eta_i^0 z_1} + c_4 + \frac{p_2(r(1)-r(2))}{\tilde{r}}.$$

We now compute $P_1(x,i)$. The corresponding differential equation is

$$\begin{cases} \dfrac{\sigma^2(1)}{2}\dfrac{\partial^2 P_1(x,1)}{\partial x^2} + r(1)\dfrac{\partial P_1(x,1)}{\partial x} + \lambda_1(P_1(x,2) - P_1(x,1)) = 0, \\ \dfrac{\sigma^2(2)}{2}\dfrac{\partial^2 P_1(x,2)}{\partial x^2} + r(2)\dfrac{\partial P_1(x,2)}{\partial x} + \lambda_2(P_1(x,1) - P_1(x,2)) = 0, \\ P_1(-z_1,1) = P_1(-z_1,2) = 0,\ P_1(z_2,1) = P_1(z_2,2) = 1 \end{cases} \tag{15}$$

for $-z_1 < x < z_2$.

Similarly, we can show

$$\begin{cases} P_1(x,1) = \sum_{i=1}^{3} c_i e^{\eta_i^0(x+z_1)} + c_4, \\ P_1(x,2) = \sum_{i=1}^{3} c_i \kappa_i^0 e^{\eta_i^0(x+z_1)} + c_4, \end{cases}$$

where, using the initial conditions in (15), (c_1,c_2,c_3,c_4) is determined by

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ \kappa_1^0 & \kappa_2^0 & \kappa_3^0 & 1 \\ e^{\eta_1^0(z_1+z_2)} & e^{\eta_2^0(z_1+z_2)} & e^{\eta_3^0(z_1+z_2)} & 1 \\ \kappa_1^0 e^{\eta_1^0(z_1+z_2)} & \kappa_2^0 e^{\eta_2^0(z_1+z_2)} & \kappa_3^0 e^{\eta_3^0(z_1+z_2)} & 1 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}.$$

Therefore, we obtain

$$P_1^* = P(S(\tau_0) \geq S_0 e^{z_2}) = p_1 P_1(0,1) + p_2 P_1(0,2) = \sum_{i=1}^{3} c_i (p_1 + \kappa_i^0 p_2) e^{\eta_i^0 z_1} + c_4.$$

Finally, we compute $P_2(x,i)$. The corresponding differential equation is identical to that in (15) except the boundary conditions become

$$P_1(-z_1,1) = P_1(-z_1,2) = 1,\ P_1(z_2,1) = P_1(z_2,2) = 0.$$

Similarly, the loss probabilities given $\alpha = 1,2$ can be written as

$$\begin{cases} P_2(x,1) = \sum_{i=1}^{3} c_i e^{\eta_i^0 (x+z_1)} + c_4, \\ P_2(x,2) = \sum_{i=1}^{3} c_i \kappa_i^0 e^{\eta_i^0 (x+z_1)} + c_4, \end{cases}$$

where the constants c_i, $i = 1,2,3,4$, are determined by

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ \kappa_1^0 & \kappa_2^0 & \kappa_3^0 & 1 \\ e^{\eta_1^0(z_1+z_2)} & e^{\eta_2^0(z_1+z_2)} & e^{\eta_3^0(z_1+z_2)} & 1 \\ \kappa_1^0 e^{\eta_1^0(z_1+z_2)} & \kappa_2^0 e^{\eta_2^0(z_1+z_2)} & \kappa_3^0 e^{\eta_3^0(z_1+z_2)} & 1 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}.$$

The loss probability is

$$P_2^* = P(S(\tau_0) \leq S_0 e^{-z_1}) = p_1 P_2(0,1) + p_2 P_2(0,2) = \sum_{i=1}^{3} c_i (p_1 + \kappa_i^0 p_2) e^{\eta_i^0 z_1} + c_4.$$

Remark 5.2. It would be interesting to study the sensitivity (or robustness) of the solution with respect to these parameters. Our numerical experiments indicate that the solution (z_1^*, z_2^*) is not sensitive with respect to small changes in $(\lambda_i, r(i), \sigma(i))$ in both the 1-D and 2-D cases. However, there is no straightforward way to prove this even in these simple cases. The main difficulty seems due to the fact that a maximizer of a function is not necessarily continuous with respect to these parameters.

6. A numerical example. In this section, we consider a numerical example with the reward function $\Phi(x) = e^x - 1$ given in (4). We begin with related computational issues. In this paper, the time is measured in years.

Estimation of $r(i)$ and $\sigma(i)$. To use the results in this paper, one needs to estimate $r(i)$ and $\sigma(i)$. We first consider a 1-D case. In this case, $r = r(1)$ and $\sigma = \sigma(1)$. The following procedure is standard for estimating the so-called historical volatility; see Hull [10].

Let S_i, $i = 0,1,\ldots,n$, denote the daily closing price of a stock and let

$$\zeta_i = \log S_i - \log S_{i-1},\ i = 1,2,\ldots,n.$$

Let $t_i = i/N_0$, where $N_0 = 252$ equals the number of trading days per annum. Then

$$\zeta_i = r(t_i - t_{i-1}) + \sigma(w(t_i) - w(t_{i-1})) \sim N\left(\frac{r}{N_0}, \frac{\sigma^2}{N_0}\right).$$

Let

$$\overline{\zeta} = \frac{\zeta_1 + \cdots + \zeta_n}{n}.$$

Then in view of the law of large numbers, the mean $r = N_0\overline{\zeta}$. Moreover, the standard deviation

$$\frac{\sigma}{\sqrt{N_0}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\zeta_i - \overline{\zeta})^2}.$$

Therefore, the volatility rate

$$\sigma = \sqrt{N_0} \cdot \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\zeta_i - \overline{\zeta})^2}. \tag{16}$$

In the 2-D case, we start from a major market index such as the DJIA, NASDAQ, or S&P 500 to determine the market trends (up- or down-trends during a period of time) in the past 52 weeks. The jump rates λ_1 and λ_2 can be determined using one of these indices. Next we split the entire historical stock price data into two parts: the up-trend part consists of the price during the market up-trend periods and the down-trend part includes the price during the down-trend periods. Then we treat the up-trend and down-trend parts separately as in the 1-D case to obtain the up-trend volatility $\sigma(1)$ and down-trend volatility $\sigma(2)$.

To estimate the expected return rates, in the up-trend part let S_0^{up} denote the initial price, N^{up} the total points gained during the entire combined periods, and n^{up} the total number of trading days of the up-trend periods. Similarly, let S_0^{down}, N^{down}, and n^{down} denote the initial price, total points declined, and number of trading days in the down-trend periods, respectively. We choose

$$r(1) = N_0\left(\frac{\log(S_0^{\mathrm{up}} + N^{\mathrm{up}}) - \log S_0^{\mathrm{up}}}{n^{\mathrm{up}}}\right),$$
$$r(2) = N_0\left(\frac{\log(S_0^{\mathrm{down}} - N^{\mathrm{down}}) - \log S_0^{\mathrm{down}}}{n^{\mathrm{down}}}\right).$$

Remark 6.1. There are several other approaches available for estimating the parameters $r(i)$ and $\sigma(i)$ with constant $\alpha(t)$ (no switching). For example, linear filtering theory [17, p. 93] can be used to estimate the expected return $r(i)$. Another volatility estimation method uses daily price spreads (daily highs vs. daily lows). The larger the averaged spread the more volatile the stock. Interestingly, such an estimate is quite consistent with the estimate in (16); see Tompkins [19].

We next apply our method to daily closes of Microsoft stock (NASDAQ-MSFT) in 1999 (see Figure 1) based on the closes in the year 1998.

We use the NASDAQ Composite Index from 1998 to determine the general market movement which is given as follows:

Up-trend	1/2–4/22	6/16–7/20	9/11–9/23	10/9–12/31
Down-trend	4/23–6/15	7/21–9/10	9/24–10/8	

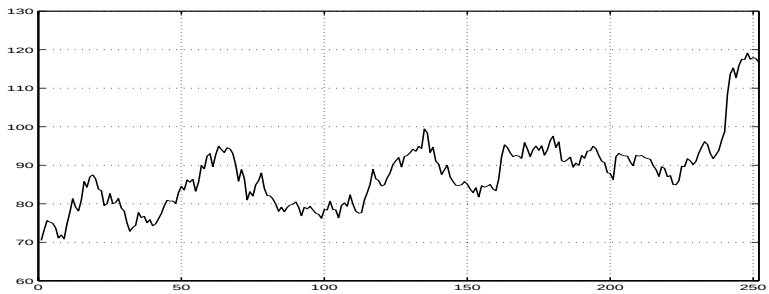


FIG. 1. *Adjusted closing price of MSFT (Jan. 4–Dec. 31, 1999).*

Moreover, the total number of "up" trading days is 167 and that of "down" days is 85. Thus, the average up (and average down, respectively) duration time is

$$\frac{1}{\lambda_1} = \frac{167}{252} \cdot \frac{1}{4}$$

and

$$\left(\frac{1}{\lambda_2} = \frac{85}{252} \cdot \frac{1}{3}, \text{ respectively} \right).$$

This gives the jump rates of $\alpha(\cdot)$

$$\lambda_1 = 6.04, \ \lambda_2 = 8.90.$$

Remark 6.2. We would like to point out that the switching $\alpha(t)$ can also be used to characterize the trends of an industrial group index or even that of an individual stock.

In addition, we obtain

$$r(1) = 1.50, \quad r(2) = -1.61, \quad \sigma(1) = 0.44, \quad \sigma(2) = 0.63.$$

We choose $p_1 = p_2 = 0.5$ because the market had rallied long enough since September 1998 that a market correction was possible. We choose $\mathcal{I} = [0.01, 0.36] \times [0.01, 2.3]$. Moreover, we take the initial value $S_0 = 70.5$, which is the closing price on January 4, 1999.

Selection of ρ. Note that the expected holding time $E\tau_0$ and the probability ratio P_1^*/P_2^* depend on the discount factor ρ. Typically, $E\tau_0$ provides a time period that capital has to be tied to the investment. On the other hand, P_1^*/P_2^* gives the likelihood of profitability. Clearly, these are two important factors in trading decision making. Intuitively, larger ρ discounts more on a future return, which leads to shorter holding time and larger probability of making a profit (see Table 5). In view of this, the discount factor ρ should be chosen according to how long investment capital is available and/or the outlook of profit probability. For example, if one plans to invest the capital for half a year, then one should choose ρ such that $E\tau_0 \sim 0.5$. Or one may choose ρ so that the probability ratio

$$P^* := \frac{P_1^*}{P_2^*} \geq k_0$$

TABLE 5
Dependence of $E\tau_0$ and P_1^*/P_2^* on ρ in 2-D model.

ρ	1	2	3	4	5	6	7	8	9	10
$E\tau_0$	2.07	1.23	1.00	0.51	0.44	0.39	0.35	0.33	0.30	0.29
P_1^*/P_2^*	1.58	2.59	3.51	5.21	6.42	7.76	9.10	10.24	11.84	13.0

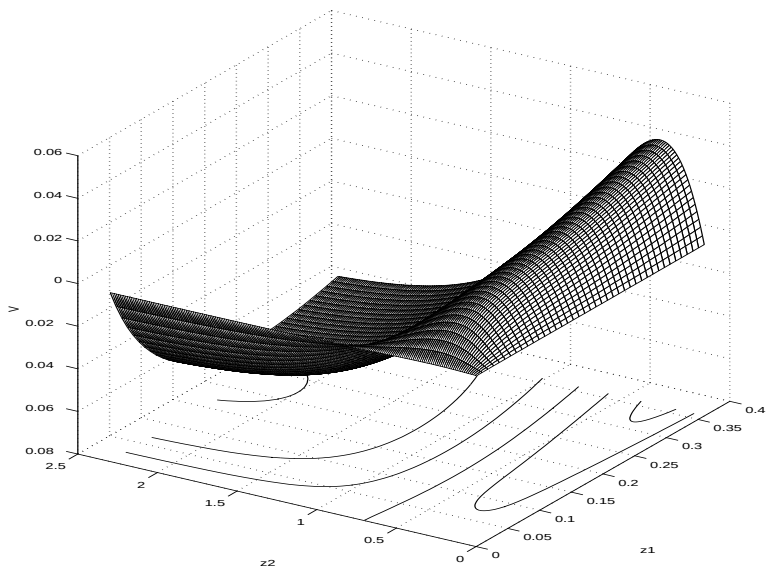


FIG. 2. *2-D reward function: MSFT (Jan. 2–Dec. 30, 1998).*

for a predetermined $k_0 > 0$. These two criteria can also be used jointly if so desired. To illustrate this idea, we compute $E\tau_0$ and P_1^*/P_2^* for $\rho = 1, 2, \ldots, 10$. Table 5 gives the dependence of these quantities on ρ.

Suppose, for example, we plan to invest capital for half a year and want the corresponding profit probability greater than 0.8 (or $P_1^*/P_2^* \geq 4$). Then we should choose $\rho = 4$. Using $\rho = 4$, the graph of the reward function V is given in Figure 2, which indicates the maximum occurs at $(z_1^*, z_2^*) = (0.36, 0.28)$.

If one decides to observe our selling rule, then, according to Figure 1, sell on January 20, 1999 at 81 if choosing $\rho = 10$, sell on January 26, 1999 at 81.48 if $\rho = 9$, and so on. These selling dates are listed in Table 6, in which $\overline{\tau}_0$ is the actual holding time if sold at or above the target price $S_0 e^{z_2^*}$ at closing. In Table 6, "*" means the target price has not yet been reached. Since the target period $E\tau_0 = 2.07$ is a little more than two years, one expects the corresponding target 195.18 to be reached around the beginning of year 2001.

One may also use the 1-D formula for similar calculations. In this case, $r = 0.75$ and $\sigma = 0.38$. The corresponding 1-D reward function (with $\rho = 4$) is given in Figure 3. As in Tables 5 and 6, Tables 7 and 8 provide corresponding results in the 1-D setting.

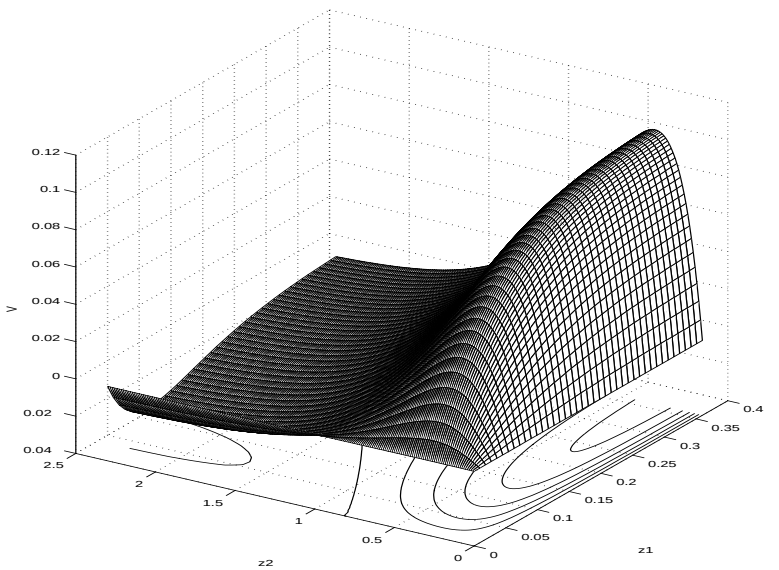


FIG. 3. 1-*D reward function: MSFT* (*Jan.* 2–*Dec.* 30, 1998).

TABLE 6
2-*D model: Comparisons with real data.*

ρ	1	2	3	4	5	6	7	8	9	10
$S_0 e^{z_2^*}$	195.18	116.35	100.20	92.99	88.81	85.80	83.85	82.89	81.48	81.00
sold on	*	12/22	12/15	4/5	3/25	1/26	1/26	1/26	1/26	1/20
$\overline{\tau}_0$	*	0.98	0.96	0.25	0.23	0.064	0.064	0.064	0.064	0.048
$\lvert E\tau_0 - \overline{\tau}_0 \rvert$	*	0.25	0.04	0.26	0.21	0.33	0.29	0.27	0.24	0.24

TABLE 7
Dependence of $E\tau_0$ *and* P_1^*/P_2^* *on* ρ *in* 1-*D model.*

ρ	1	2	3	4	5	6	7	8	9	10
$E\tau_0$	2.34	0.78	0.50	0.38	0.31	0.26	0.23	0.20	0.19	0.17
P_1^*/P_2^*	41.08	41.16	41.80	43.01	44.72	46.49	48.20	50.57	52.10	53.93

Looking at Tables 6 and 8, it can be seen that the 2-D model provides more realistic target prices. As for the expected holding times, the 1-D model provides better estimates for larger $\rho(\geq 5)$, which corresponds to smaller $E\tau_0$. On the other hand, the 2-D model yields better estimates for smaller $\rho(\leq 4)$. In addition, the 1-D model in Table 7 produces a much greater probability ratio estimate P_1^*/P_2^*.

7. Concluding remarks. Selling a stock is a crucial step to nail down real profit or to cut loss short. But emotions may come into play when selling. A rigorous selling rule would help an investor to control human emotion and consistently make profits. The results obtained in this paper, including the predetermined target price, stop-loss limit, target period, and the corresponding probabilities, could be used as a

TABLE 8
1-D model: comparisons with real data.

ρ	1	2	3	4	5	6	7	8	9	10
$S_0e^{z_2^*}$	428.86	129.70	104.24	95.08	89.77	86.72	84.75	82.82	81.88	80.94
sold on	*	*	12/15	7/16	3/25	3/25	1/26	1/26	1/26	1/20
$\overline{\tau}_0$	*	*	0.96	0.54	0.23	0.23	0.064	0.064	0.064	0.048
$\lvert E\tau_0 - \overline{\tau}_0\rvert$	*	*	0.46	0.16	0.08	0.03	0.17	0.17	0.13	0.12

guide to actual trading.

Appendix. In this appendix, we provide proofs of the results in this paper. We also give two technical lemmas required in the proofs.

Proof of Theorem 3.2. To show (a), it suffices to show the equation has a C^2 solution. Then following from Dynkin's formula, we have

$$v(x,i) = E[e^{-\rho\tau(x)}\Phi(\xi(\tau(x)))|\alpha(0) = i],$$

which implies the uniqueness of the solution.

Let $\phi(x) = (v(x,1),\dots,v(x,m))' \in \mathbb{R}^{1\times m}$. Then the differential equations (7) can be written as

$$\frac{1}{2}\mathrm{diag}(\sigma^2(1),\dots,\sigma^2(m))\ddot{\phi} + \mathrm{diag}(r(1),\dots,r(m))\dot{\phi} - \rho\phi + Q\phi = 0. \tag{17}$$

Let

$$\Gamma = \mathrm{diag}\left(\frac{r(1)}{\sigma^2(1)},\dots,\frac{r(m)}{\sigma^2(m)}\right),$$

$$\Sigma = \mathrm{diag}\left(\frac{1}{\sigma^2(1)},\dots,\frac{1}{\sigma^2(m)}\right).$$

Then (17) can be written as

$$\ddot{\phi} + 2\Gamma\dot{\phi} - 2\rho\Sigma\phi + 2\Sigma Q\phi = 0.$$

Let $\theta \in [0,1]$, $a, b \in \mathbb{R}^m$. Consider the differential equation

$$\begin{cases} \ddot{\phi} + 2\Gamma\dot{\phi} - 2\rho\Sigma\phi + 2\theta\Sigma Q\phi = 0, \\ \phi(-z_1) = a,\ \phi(z_2) = b. \end{cases} \tag{18}$$

Let

$$y(x) = (\phi'(x), \dot{\phi}'(x))' = \left(v(x,1),\dots,v(x,m), \frac{\partial v(x,1)}{\partial x},\dots,\frac{\partial v(x,m)}{\partial x}\right)'$$

and

$$A^\theta = \begin{pmatrix} 0_{m\times m} & I_{m\times m} \\ 2\Sigma(\rho I - \theta Q) & -2\Gamma \end{pmatrix}.$$

Then the differential equation (18) is equivalent to

$$\dot{y} = A^\theta y,\ \phi(-z_1) = a,\ \phi(z_2) = b. \tag{19}$$

Let

$$\Theta = \Big\{\theta \in [0,1] : \text{ such that (18) (or (19)) has a solution for all } a, b \in \mathbb{R}^m\Big\}.$$

Using the condition $\phi(-z_1) = a$, we can write

$$y(x) = \exp(A^\theta(x+z_1)) \begin{pmatrix} a \\ c \end{pmatrix},$$

where $c \in \mathbb{R}^m$ needs to be determined by the condition $\phi(z_2) = b$. Write

$$\exp(A^\theta(x+z_1)) = \begin{pmatrix} F_{11}^\theta(x+z_1) & F_{12}^\theta(x+z_1) \\ F_{21}^\theta(x+z_1) & F_{22}^\theta(x+z_1) \end{pmatrix},$$

such that $F_{ij}^\theta \in \mathbb{R}^{m\times m}$ for $i,j = 1,2$. Then (19) has a solution for all $b \in \mathbb{R}^m$ if and only if $F_{12}^\theta(z_1+z_2)$ is invertible.

We first show that Θ is not empty. In fact, $0 \in \Theta$. To show this, we note that when $\theta = 0$

$$\exp(A^0 x) = \begin{pmatrix} \operatorname{diag}(e^{\eta_1^1 x}, \dots, e^{\eta_m^1 x}) & \operatorname{diag}(e^{\eta_1^2 x}, \dots, e^{\eta_m^2 x}) \\ \operatorname{diag}(\eta_1^1 e^{\eta_1^1 x}, \dots, \eta_m^1 e^{\eta_m^1 x}) & \operatorname{diag}(\eta_1^2 e^{\eta_1^2 x}, \dots, \eta_m^2 e^{\eta_m^2 x}) \end{pmatrix} \times \begin{pmatrix} I_{m\times m} & I_{m\times m} \\ \operatorname{diag}(\eta_1^1, \dots, \eta_m^1) & \operatorname{diag}(\eta_1^2, \dots, \eta_m^2) \end{pmatrix}^{-1},$$

where η_i^1 and η_i^2 are the eigenvalues of A^0 given by

$$\begin{cases} \eta_i^1 = -\dfrac{r(i)}{\sigma^2(i)} - \sqrt{\left(\dfrac{\mu(i)}{\sigma^2(i)}\right)^2 + \dfrac{2\rho}{\sigma^2(i)}}, \\ \eta_i^2 = -\dfrac{r(i)}{\sigma^2(i)} + \sqrt{\left(\dfrac{\mu(i)}{\sigma^2(i)}\right)^2 + \dfrac{2\rho}{\sigma^2(i)}}, \end{cases}$$

$i = 1, \dots, m$.

$$F_{12}^0(z_1+z_2) = \operatorname{diag}\left(\frac{e^{\eta_1^2(z_1+z_2)} - e^{\eta_1^1(z_1+z_2)}}{\eta_1^2 - \eta_1^1}, \dots, \frac{e^{\eta_m^2(z_1+z_2)} - e^{\eta_m^1(z_1+z_2)}}{\eta_m^2 - \eta_m^1}\right),$$

which is invertible.

Next, let

$$\theta^* = \sup\Big\{\theta : \ [0,\theta] \subset \Theta\Big\}.$$

Clearly, $\theta^* \geq 0$. We show that $\theta^* = 1$. Note that if we can show $\theta^* \in \Theta$, then $F_{12}^{\theta^*}(z_1+z_2)$ is invertible, which implies, for any $\delta > 0$,

$$\exp(A^{\theta^*+\delta}(z_1+z_2)) = \exp(A^{\theta^*}(z_1+z_2)) + O(\delta),$$

because

$$A^{\theta^*+\delta}(z_1+z_2) = A^{\theta^*}(z_1+z_2) + \delta \begin{pmatrix} 0_{m\times m} & 0_{m\times m} \\ -2\Sigma Q & 0_{m\times m} \end{pmatrix}.$$

Thus for δ small enough, $F_{12}^{\theta}(z_1+z_2))$ is invertible. Following a similar argument, it is easy to see that $[0,\delta)\subset\Theta$. Therefore, $\theta^*>0$. If $\theta^*<1$, then it follows $(\theta^*+\delta)\in\Theta$, which contradicts the definition of θ^*. It remains to show $\theta^*\in\Theta$. By the definition of θ^*, there exists a sequence $\{\theta_n\}\subset\Theta$ such that $\theta_n\to\theta^*$ as $n\to\infty$. For any $a=(a_1,\ldots,a_m)'$ and $b=(b_1,\ldots,b_m)'$, let $\phi_n(x)\in C^2[-z_1,z_2]$ be the corresponding solutions with terminal reward function

$$\Phi(x,i)=\frac{b_i-a_i}{z_1+z_2}(x-z_2)+b_i.$$

Then it can be shown that ϕ_n is uniformly bounded by applying Dynkin's formula to

$$e^{-\rho t}\phi_n(\xi_n(t),\alpha_n(t)),$$

which leads to

$$\phi_n(x)=\Big(E[e^{-\rho\tau(x)}\Phi(\xi_n(\tau(x)))|\alpha_n(0)=1],\ldots,E[e^{-\rho\tau(x)}\Phi(\xi_n(\tau(x)))|\alpha_n(0)=m]\Big)',$$

where $\alpha_n(t)$ is a Markov chain generated by θ_nQ and $\xi_n(t)$ is the corresponding diffusion given in (5).

Integrating twice the differential equation (18) from $-z_1$ to x yields

$$\begin{cases}\phi_n(x)=a+(x+z_1)\dot{\phi}_n(-z_1)+g_n(x),\\ b=a+(z_1+z_2)\dot{\phi}_n(-z_1)+g_n(z_2),\end{cases}$$

where

$$g_n(x)=-2\Gamma\int_{-z_1}^{x}(\phi_n(r)-a)dr+2\rho\Sigma\int_{-z_1}^{x}\int_{-z_1}^{u}\phi_n(r)drdu-2\theta_n\Sigma Q\int_{-z_1}^{x}\int_{-z_1}^{u}\phi_n(r)drdu.$$

In view of the uniform boundedness of ϕ_n, there exists a subsequence $\{n_k\}$ such that $\int_{-z_1}^{x}\phi_{n_k}(r)dr$ converges in sup-norm. Thus $g_{n_k}(-z_1)$ converges which in turn implies $\dot{\phi}_{n_k}(-z_1)$ converges. Therefore, $\phi_{n_k}(x)$ converges in sup-norm. Hence, for all $a,b\in\mathbb{R}^m$, (18) has a solution when $\theta=\theta^*$. It follows that $\theta^*\in\Theta$.

We next show that $\phi(x,i)=\phi_{z_1,z_2}(x,i)$ is continuous. This can be seen from

$$\begin{aligned}\phi(x)=F_{11}^{1}(x+z_1)\Phi(-z_1)\mathbb{1}+F_{12}^{1}(x+z_1)(F_{12}^{1}(z_1+z_2))^{-1}\\ \times\left(\Phi(z_2)\mathbb{1}-F_{11}^{1}(z_1+z_2)\Phi(-z_1)\mathbb{1}\right),\end{aligned}$$

where $\mathbb{1}=(1,\ldots,1)'\in\mathbb{R}^m$.

Given (x,i), the continuity of $v(x,i)$ in (z_1,z_2) implies the continuity of V in (z_1,z_2). The existence of (z_1^*,z_2^*) maximizing V follows from the compactness of $\mathcal{I}$.

This completes the proof. □

Remark A.1. Note that a key condition that guarantees the existence is the uniform boundedness of $\phi(x)$. The proof does not go through without such a condition. For example, consider

$$\ddot{y}+\eta^2y=0,$$

with $0\le\eta\le1$, $y(0)=a$, and $y(\pi)=b$. Then

$$y(x)=\left(\frac{b-a\cos\eta\pi}{\sin\eta\pi}\right)\sin\eta x+a\cos\eta x,$$

which is not uniformly bounded as $\eta \to 1$. Consequently,

$$e^{Ax} = \begin{pmatrix} \cos \eta x & \dfrac{\sin \eta x}{\eta} \\ -\eta \sin \eta x & \cos \eta x \end{pmatrix}.$$

$F_{12} = (\sin \eta x)/\eta$ is not invertible at $x = \pi$ when $\eta = 1$.

Recall the independence condition of $\alpha(\cdot)$ and $w(\cdot)$. We have the following lemma.

LEMMA A.2. *Let $\mathcal{D}$ denote the sigma-algebra generated by $\{\alpha(s) : s \le T\}$. Then*

$$E\left[\int_0^T \sigma(\alpha(s))dw(s)\Big|\mathcal{D}\right] = 0.$$

Proof. Given a positive integer n, let $t_k = kT/n$ for $k = 0, 1, \ldots, n$. Define

$$\alpha^n(t) = \alpha(t_k) \text{ if } t \in [t_k, t_{k+1}),\ k = 0, 1, \ldots, n,$$

and $\alpha^n(T) = \alpha(T)$. Then it follows that, as $n \to \infty$,

$$\int_0^T \Big(\sigma(\alpha(s)) - \sigma(\alpha^n(s))\Big)^2 ds \to 0 \text{ a.s.}$$

Since $\mathcal{M}$ is finite, $\sigma(i)$ is bounded. The Lebesgue dominated convergence theorem implies

$$E\int_0^T \Big(\sigma(\alpha(s)) - \sigma(\alpha^n(s))\Big)^2 ds \to 0.$$

We next show that, for each fixed n,

$$E\left[\int_0^T \sigma(\alpha^n(s))dw(s)\Big|\mathcal{D}\right] = 0 \text{ a.s.}$$

In fact, let $y^n(t) = y_k$ if $t \in [t_k, t_{k+1})$ for given $(y_0, y_1, \ldots, y_n) \in \mathbb{R}^{n+1}$. Define

$$f(y_0, y_1, \ldots, y_n, \omega) = \int_0^T \sigma(y^n(s))dw(s)(\omega).$$

We have

$$g(y_0, y_1, \ldots, y_n) := E[f(y_0, y_1, \ldots, y_n, \omega)|\mathcal{D}] = 0.$$

It follows that, in view of the martingale property,

$$g(\alpha(t_0), \alpha(t_1), \ldots, \alpha(t_n)) = Ef(\alpha(t_0), \alpha(t_1), \ldots, \alpha(t_n), \omega) = 0.$$

Moreover, using Jensen's inequality, we have

$$\left(E\left[\int_0^T \Big(\sigma(\alpha(s)) - \sigma(\alpha^n(s))\Big)dw(s)\Big|\mathcal{D}\right]\right)^2$$
$$\le E\left[\left(\int_0^T \Big(\sigma(\alpha(s)) - \sigma(\alpha^n(s))\Big)dw(s)\right)^2 \Big|\mathcal{D}\right].$$

Therefore, we obtain

$$E\left(E\left[\int_0^T \sigma(\alpha(s))dw(s)\Big|\mathcal{D}\right] - E\left[\int_0^T \sigma(\alpha^n(s))dw(s)\Big|\mathcal{D}\right]\right)^2$$
$$\leq E\int_0^T \Big(\sigma(\alpha(s)) - \sigma(\alpha^n(s))\Big)^2 ds \to 0.$$

Hence we have

$$E\left[\int_0^T \sigma(\alpha(s))dw(s)\Big|\mathcal{D}\right] = 0 \text{ a.s.} \qquad \square$$

The next lemma will be needed in the proof of Lemma 4.1.

LEMMA A.3. *Given $z > 0$, there exist $s_0 > 0$ and $\delta_0 > 0$ such that*

$$P(|\xi(s_0)| \geq z | \xi(0) = x, \alpha(0) = i) \geq \delta_0 \tag{20}$$

for all $|x| \leq z$ and $i \in \mathcal{M}$.

Proof. If (20) does not hold, then, for some $z_0 > 0$, there exist $|x_0| \leq z_0$ and $i_0 \in \mathcal{M}$ such that, for all s_0 and δ_0,

$$P(|\xi(s_0)| \geq z_0 | \xi(0) = x_0, \alpha(0) = i_0) < \delta_0,$$

which implies $P(|\xi(s_0)| \geq z_0 | \xi(0) = x_0, \alpha(0) = i_0) = 0$ and hence $|\xi(s_0)| < z_0$ a.s. Therefore,

$$|\xi(s_0) - x_0| \leq z_0 + |x_0| \text{ a.s.} \tag{21}$$

In view of Lemma A.2, we have

$$E\left[\int_0^{s_0} r(\alpha(s))ds \int_0^{s_0} \sigma(\alpha(s))dw(s)\right]$$
$$= E\left\{\int_0^{s_0} r(\alpha(s))ds E\left[\int_0^{s_0} \sigma(\alpha(s))dw(s)\Big|\mathcal{D}\right]\right\} = 0.$$

It follows that

$$E(\xi(s_0) - x_0)^2 = E\left(\int_0^{s_0} r(\alpha(s))ds\right)^2 + E\left(\int_0^{s_0} \sigma(\alpha(s))dw(s)\right)^2$$
$$\geq E\int_0^{s_0} (\sigma(\alpha(s)))^2 ds.$$

Moreover, it is well known that for $i_1 \in \mathcal{M}$,

$$\nu_i := \lim_{s\to\infty} P(\alpha(s) = i | \alpha(0) = i_1) \text{ exists.}$$

It is easy to see that $\nu_i \geq 0$ and $\nu_1 + \cdots + \nu_m = 1$. Therefore,

$$\frac{1}{s_0} E\int_0^{s_0} (\sigma(\alpha(s)))^2 ds = \frac{1}{s_0}\int_0^{s_0} \sum_{i=1}^m \sigma^2(i) P(\alpha(s) = i | \alpha(0) = i_1) ds$$
$$\to \sum_{i=1}^m \sigma^2(i)\nu_i ds > 0.$$

This contradicts (21) because

$$\frac{E(\xi(s_0)-x_0)^2}{s_0} \le \frac{(z_0+|x_0|)^2}{s_0} \to 0 \text{ as } s_0 \to \infty. \qquad \square$$

Proof of 4.1. Let $z = \max\{z_1, z_2\}$. Then in view of Lemma A.3, there exist s_0 and δ_0 such that

$$P(|\xi(s_0)| \ge z | \xi(0) = x, \alpha(0) = i) \ge \delta_0$$

for all $|x| \le z$ and $i \in \mathcal{M}$.

Using this inequality, we have

$$\begin{aligned} P(\tau_0 > s_0) \le & \quad P(|\xi(s_0)| < z | \xi(0) = x, \alpha(0) = i) \\ \le & \quad 1 - P(|\xi(s_0)| \ge z | \xi(0) = x, \alpha(0) = i) \\ \le & \quad 1 - \delta_0. \end{aligned}$$

Similarly, we have

$$\begin{aligned} P(\tau_0 > 2s_0) \le & \quad P(|\xi(2s_0)| < z, |\xi(s_0)| < z | \xi(0) = x, \alpha(0) = i) \\ = & \quad \sum_{j=1}^{m} \int_{-z}^{z} P(|\xi(2s_0)| < z | \xi(s_0) = y, \alpha(s_0) = j) \\ & \qquad \times P(\xi(s_0) \in dy, \alpha(s_0) = j | \xi(0) = x, \alpha(0) = i) \\ \le & \quad (1-\delta_0) P(|\xi(s_0)| < z | \xi(0) = x, \alpha(0) = i) \\ \le & \quad (1-\delta_0)^2. \end{aligned}$$

In general, we can show, for each $n = 1, 2, \ldots$,

$$P(\tau_0 > n s_0) \le (1-\delta_0)^n.$$

Following the proof of Chow and Teicher [1, Corollary 4.1.3], we have

$$E\tau(x) \le s_0 \sum_{n=0}^{\infty} P(\tau_0 > n s_0) \le s_0 \sum_{n=0}^{\infty} (1-\delta_0)^n = \frac{s_0}{\delta_0}. \qquad \square$$

Proof of Theorem 4.2. The proof is similar to that of Theorem 3.2 except that if $T(x,i)$ is a solution to (8), then by Dynkin's formula,

$$T(x,i) = E[\tau(x) | \xi(0) = x, \alpha(0) = i] \le K. \qquad \square$$

Proof of Theorem 4.5. The proof follows that of Theorem 3.2 except by taking $\Phi(x) = I_{\{x=z_2\}}$, $\rho = 0$, and noting that

$$P_1(x,i) = E[\Phi(\xi(\tau(x)))| \alpha(0) = i]. \qquad \square$$

Proof of Theorem 4.6. This is similar to the proof of Theorem 4.5. $\square$

Acknowledgments. The author would like to thank the referees and the associate editor for their many valuable comments and suggestions which led to much improvement of the paper.

REFERENCES

[1] Y. S. Chow and H. Teicher, *Probability Theory*, Springer-Verlag, New York, 1978.

[2] A. Cadenillas and S. R. Pliska, *Optimal trading of a security when there are taxes and transaction costs*, Finance Stoch., 3 (1999), pp. 137–165.

[3] G. M. Constantinides, *Capital market equilibrium with personal tax*, Econometrica, 51 (1983), pp. 611-636.

[4] R. M. Dammon and C. S. Spatt, *The optimal trading and pricing of securities with asymetric capital gains taxes and transaction costs*, Rev. Financial Studies, 9 (1996), pp. 921–952.

[5] M. H. A. Davis, *Markov Model and Optimization*, Chapman and Hall, London, 1993.

[6] D. Duffie, *Dynamic Asset Pricing Theory*, 2nd ed., Princeton University Press, Princeton, NJ, 1996.

[7] R. J. Elliott and P. E. Kopp, *Mathematics of Financial Markets*, Springer-Verlag, New York, 1998.

[8] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1992.

[9] J. P. Fouque, G. Papanicolaou, and S. K. Ronnie, *Derivatives in Financial Markets with Stochastic Volatility*, Cambridge University Press, London, 2000.

[10] J. C. Hull, *Options, Futures, and Other Derivatives*, 3rd ed., Prentice-Hall, Upper Saddle River, NJ, 1997.

[11] I. Karatzas, *Lectures on the Mathematics of Finance*, AMS, Providence, RI, 1997.

[12] I. Karatzas and S. E. Shreve, *Methods of Mathematical Finance*, Springer-Verlag, New York, 1998.

[13] N. V. Krylov, *Nonlinear Elliptic and Parabolic Equations of the Second Order*, D. Reidel, Boston, 1987.

[14] J. L. Livermore, *How to Trade in Stocks*, Duell, Sloan and Pearce, New York, 1940.

[15] R. C. Merton, *Lifetime portfolio selection under uncertainty: The continuous-time case*, Rev. Econom. Statist., 51 (1969), pp. 247-257.

[16] M. Musiela and M. Rutkowski, *Martingale Methods in Financial Modeling*, Springer-Verlag, New York, 1997.

[17] B. Øksendal, *Stochastic Differential Equations*, 4th ed., Springer-Verlag, New York, 1995.

[18] W. J. O'Neil, *How to Make Money in Stocks*, 2nd ed., McGraw-Hill, New York, 1995.

[19] R. G. Tompkins, *Options Analysis: A State-of-the-Art Guide to Options Pricing, Trading & Portfolio Applications*, Probus, New York, 1995.

[20] G. Yin and Q. Zhang, *Continuous-Time Markov Chains and Applications: A Singular Perturbation Approach*, Springer-Verlag, New York, 1998.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 1, pp. 88–106

ON THE BOUNDEDNESS AND CONTINUITY OF THE SPECTRAL FACTORIZATION MAPPING*

BIRGIT JACOB† AND JONATHAN R. PARTINGTON†

Abstract. It is known that the spectral factorization mapping is bounded, but not continuous, on L_∞, the space of essentially bounded measurable functions on the unit circle. In this article we study the spectral factorization mapping on decomposing Banach algebras. The most important example of a decomposing Banach algebra is the Wiener algebra, the space of all absolutely convergent Fourier series. It is shown that the spectral factorization mapping is locally Lipschitz continuous, but not bounded, on all decomposing Banach algebras in consideration. An application is given to the construction of approximate normalized coprime factorization.

Key words. spectral factorization, continuity of the factorization, boundedness of the factorization, normalized coprime factorization, decomposing Banach algebras

AMS subject classifications. 47A68, 46J10, 46J15

PII. S036301299935184X

1. Introduction.

Spectral factorization is the process by which a (possibly matrix-valued) function G is written as $G = W^*W$; that is, $G(e^{i\theta}) = W(e^{i\theta})^*W(e^{i\theta})$ for $\theta \in [0, 2\pi]$, or $G(i\omega) = W(i\omega)^*W(i\omega)$ for $\omega \in \mathbb{R}$. It is a procedure arising in many system-theoretic applications: for example, it is linked to H_∞ robust control, linear quadratic optimal control, and linear quadratic Gaussian (LQG) optimal control by the solution of Riccati equations [6, 8, 9, 19, 27, 29, 30]. Again, it is also seen in robust control via the construction of normalized coprime factorizations [30, 8].

For finite dimensional linear systems, the required spectral factors are generally rational functions, and there are well-established algebraic techniques for their solution. In the infinite dimensional case, it is common to proceed by approximation procedures (cf. [20, 11]), in which case it is important to know whether the operation of taking spectral factors is continuous and to be able to provide error estimates. The answer to this question depends crucially on which normed spaces one is using, as we shall see later. We shall build on work of Clancey and Gohberg [7], which we relate to work of Peller and Khrushchev [23], in order to give a systematic approach to this problem.

It is also important to know under what circumstances the norms of the spectral factors can be bounded in terms of the norm of the function being factorized. One application that we mention here is in the work of Patil on recovery of Hardy class functions from their values on a limited bandwidth [22, 21], where a sequence of spectral factorizations is employed. Similar calculations arise in the context of band-limited identification [2]. The boundedness problem turns out to be independent of the question of the continuity of the factorization. We shall see that, in most of the function spaces considered in systems theory (excepting the obvious case of the L_∞ norm), spectral factorization is actually unbounded.

In this paper we study only the scalar case. However, all the results obtained in this paper can be generalized without any problems to the matrix case.

*Received by the editors February 10, 1999; accepted for publication (in revised form) January 17, 2001; published electronically May 31, 2001.
http://www.siam.org/journals/sicon/40-1/35184.html
†School of Mathematics, University of Leeds, Leeds LS2 9JT, UK (birgit@amsta.leeds.ac.uk, J.R.Partington@leeds.ac.uk). The research of the first author was supported by the EPSRC.

The outline of this paper is as follows. In section 2 we review the notion of a decomposing Banach algebra and present some important examples. Spectral factorization is discussed in section 3, and its continuity in decomposing Banach algebras is the subject of section 4. The more difficult issue of boundedness is analyzed in section 5. Most of the results are presented originally for discrete-time systems (spectral factorization on the unit circle): a brief discussion of the half-plane case is given in section 6.

2. Decomposing Banach algebras. In this section we recall the notion of decomposing Banach algebras and we give many examples of these Banach algebras. We start by recalling some facts about Banach algebras which are frequently used in this paper.

By $\mathbb{T}$ we denote the unit circle $\{z \in \mathbb{C} \mid |z| = 1\}$ and by $\mathbb{D}$ the disc $\{z \in \mathbb{C} \mid |z| < 1\}$. Let $L_p(\mathbb{T})$, $1 < p < \infty$, be the Banach space of p-integrable functions on $\mathbb{T}$.

Let $\mathcal{B}$ be a commutative Banach algebra satisfying $\mathcal{B} \subset L_2(\mathbb{T})$. We will denote the norm of $\mathcal{B}$ by $\|\cdot\|_{\mathcal{B}}$ and let $\mathcal{GB}$ be the set of all invertible elements of $\mathcal{B}$. If $f \in \mathcal{B}$, we define the exponential by $e^f := \sum_{n=0}^{\infty} \frac{f^n}{n!}$. Then $e^f \in \mathcal{GB}$. By $\exp \mathcal{B}$ we denote the set $\{f \in \mathcal{B} \mid f = e^g, g \in \mathcal{B}\}$. Note that $\exp \mathcal{B}$ coincides with the component of $\mathcal{GB}$ which contains the identity, which is denoted by 1. Another important subset of $\mathcal{B}$ is

$$\mathcal{B}_{\text{pos}} := \{f \in \mathcal{B} \mid f(t) > 0 \text{ for every } t \in \mathbb{T}\}.$$

For more information on Banach algebras we refer the reader to Rickart [24], Larsen [18], and Bonsall and Duncan [5].

Next we will recall the notion of decomposing Banach algebras. For every $f \in L_2(\mathbb{T})$ we can define a function $Pf \in L_2(\mathbb{T})$ by

$$(Pf)(z) = \frac{1}{2\pi i} \int_{\mathbb{T}} \frac{f(\omega)}{\omega - z}\, d\omega.$$

P is the projection from $L_2(\mathbb{T})$ to the Hardy class $H_2(\mathbb{D})$, the space of all functions in $L_2(\mathbb{T})$ which are holomorphic within the unit disc; see, for example, Hoffman [13, p. 151]. Moreover, P is the projection from $L_p(\mathbb{T})$ onto $H_p(\mathbb{D})$ if $p \in (1, \infty)$; see Hoffman [13, p. 151].

DEFINITION 2.1. *A commutative Banach algebra $\mathcal{B} \subset L_2(\mathbb{T})$ is called a* decomposing Banach algebra *if*

- (A1) *$f \in \mathcal{B}$ implies $\bar{f} \in \mathcal{B}$ and $Pf \in \mathcal{B}$;*
- (A2) *$\mathcal{B}$ is a Banach algebra with respect to pointwise multiplication on $\mathbb{T}$;*
- (A3) *the set of trigonometric polynomials is dense in $\mathcal{B}$;*
- (A4) *every nonzero multiplicative functional on $\mathcal{B}$ coincides with a functional $f \mapsto f(t)$ defined as the value at some point $t \in \mathbb{T}$ (note that $\varphi : \mathcal{B} \to \mathbb{C}$ is a multiplicative functional if φ is a linear functional and $\varphi(ab) = \varphi(a)\varphi(b)$, $a, b \in \mathcal{B}$).*

This class of Banach algebras coincides with a class of Banach algebras introduced by Peller and Khrushchev [23] in order to study the problem of best approximation by analytic functions. Clancey and Gohberg [7] gave a more general definition of a decomposing Banach algebra, although they then imposed extra conditions in order to guarantee the existence of factorizations. The following proposition has been proved by Peller and Khrushchev [23].

PROPOSITION 2.2. *Let $\mathcal{B}$ be a decomposing Banach algebra. Then $\mathcal{B}$ is continuously embedded in $C(\mathbb{T})$, the space of maximal ideals of the algebra $\mathcal{B}$ coincides with $\mathbb{T}$, and $P \in \mathcal{L}(\mathcal{B})$.*

From now on, the symbol $\mathcal{B}$ will invariably denote a decomposing Banach algebra. The following proposition follows directly from the Beurling–Gel'fand theorem (see, for example, Larsen [18, p. 79 ff]).

PROPOSITION 2.3. *Every $f \in \mathcal{B}$, such that $f(t) \neq 0$ for all $t \in \mathbb{T}$, is an element of $\mathcal{GB}$.*

Using Proposition 2.3, we get $\mathcal{B}_{\rm pos} \subseteq \exp \mathcal{B}$, because an element $f \in \mathcal{B}_{\rm pos}$ can be joined to the identity via the path $\lambda + (1-\lambda)f$. As usual, for a function $f \in L_1(\mathbb{T})$ and $n \in \mathbb{Z}$, we write $\hat{f}(n)$ for the nth Fourier coefficient of f. Let $\mathcal{B}_+$ be defined by

$$\mathcal{B}_+ := \{f \in \mathcal{B} \mid \hat{f}(n) = 0 \text{ for all } n < 0\}.$$

For a decomposing Banach algebra $\mathcal{B}$, the set $\mathcal{B}_+$ is the set of all functions f in $\mathcal{B}$ that can be written as $f = Pg$ for some $g \in \mathcal{B}$; that is, P is the projection from $\mathcal{B}$ onto $\mathcal{B}_+$. By $Q \in \mathcal{L}(\mathcal{B})$ we denote the operator $Q := I - P$. Note that by this definition $1(t) = 1$ is an element of $\mathcal{B}_+$. An important property of $\mathcal{B}_+$ is that $\mathcal{B}_+$ is a Banach algebra.

Besides decomposing Banach algebras $\mathcal{B}$, we consider the Banach algebras $L_\infty(\mathbb{T})$ and $C(\mathbb{T})$. By $L_\infty(\mathbb{T})$ we denote the space of measurable and essentially bounded functions on $\mathbb{T}$. Clearly $L_\infty(\mathbb{T})$ satisfies (A2) and $L_\infty(\mathbb{T})_+$ is given by $H_\infty(\mathbb{D})$, the space of holomorphic and bounded functions within the unit disc $\mathbb{D}$. However, $L_\infty(\mathbb{T})$ does not satisfy the second part of (A1); see, for example, Hoffman [13, p. 155]. Let $C(\mathbb{T})$ be the Banach space of continuous functions on $\mathbb{T}$. The algebra $C(\mathbb{T})$ satisfies (A2)–(A4), but the second part of (A1) does not hold. Note that the space $C(\mathbb{T})_+$ coincides with the disc algebra $A(\mathbb{D})$, the Banach space of all continuous functions on $\bar{\mathbb{D}}$ which are holomorphic within the unit disc. Finally, we define

$$\begin{aligned}(L_\infty(\mathbb{T}))_{\rm pos} &:= \{f \in L_\infty(\mathbb{T}) \mid f^{-1} \in L_\infty(\mathbb{T}) \text{ and } f(t) > 0 \text{ for almost every } t \in \mathbb{T}\},\\ (C(\mathbb{T}))_{\rm pos} &:= \{f \in C(\mathbb{T}) \mid f(t) > 0 \text{ for } t \in \mathbb{T}\}.\end{aligned}$$

Thus $L_\infty(\mathbb{T})$ and $C(\mathbb{T})$ are Banach algebras, but not decomposing Banach algebras. Moreover, Proposition 2.2 implies that a decomposing Banach algebra cannot be a C*-algebra. Otherwise, the Gel'fand–Naimark theorem [18, p. 277] implies $\mathcal{B} = C(\mathbb{T})$, which cannot hold since $C(\mathbb{T})$ is not a decomposing Banach algebra.

All Banach algebras presented in the following are decomposing Banach algebras. The proofs can be found in Peller and Khrushchev [23]. One of the most important examples is the Wiener algebra.

Example 2.4. The *Wiener algebra* W consists of all absolutely convergent Fourier series

$$f(e^{it}) = \sum_{n\in\mathbb{Z}} a_n e^{int} \tag{2.1}$$

and is equipped with the norm

$$\|f\|_W := \sum_{n\in\mathbb{Z}} |a_n| < \infty.$$

The next example shows that it is also possible to introduce a weighted version of the Wiener algebra.

Example 2.5. Let $\omega = \{\omega_n\}_{n=0}^\infty$ be a sequence of numbers satisfying

$$\omega_{|n|} > 0 \quad \text{and} \quad \omega_{|n|} \leq \omega_{|n-k|}\omega_{|k|}, \qquad n, k \in \mathbb{Z}, \tag{2.2}$$

and

$$\overline{\lim}_{n\to+\infty}\omega_n^{1/n} = 1. \tag{2.3}$$

Then $W(\omega)$ denotes the *weighted Wiener algebra* with respect to the weights ω, which consists of all Fourier series f, given by (2.1), with

$$\|f\|_{W(\omega)} := \sum_{n\in\mathbb{Z}} \omega_{|n|}|a_n| < \infty.$$

Note that

$$\omega_{|n|} \le \omega_{|n-n|}\omega_{|n|}$$

implies $\omega_0 \ge 1$, and thus

$$\omega_{|0|} \le \omega_{|0-k|}\omega_{|k|} = \omega_{|k|}^2$$

shows $\inf_{k\ge0}\omega_k \ge 1 > 0$.

Choosing the sequence $\omega_n := 1$, we get $W(\omega) = W$.

Example 2.6. Let $p \in (1,\infty)$ and $\omega = \{\omega_n\}_{n\in\mathbb{N}_0}$ be a sequence of numbers satisfying

$$\sup_{m\in\mathbb{Z}}\left(\sum_{k+j=m}\left(\frac{\omega_{|m|}}{\omega_{|k|}\omega_{|j|}}\right)^{p'}\right)^{1/p'} < \infty \tag{2.4}$$

and (2.3), where $1/p + 1/p' = 1$. Then $\mathcal{F}\ell_p(\omega)$ denotes the space of all Fourier series f, given by (2.1), which satisfy

$$\|f\|_{\mathcal{F}\ell_p(\omega)} := \left(\sum_{n\in\mathbb{Z}}\omega_{|n|}^p|a_n|^p\right)^{1/p} < \infty.$$

Using $\omega_0 = \|e^{nit}e^{-nit}\|_{\mathcal{F}\ell_p(\omega)} \le \|e^{nit}\|_{\mathcal{F}\ell_p(\omega)}\|e^{-nit}\|_{\mathcal{F}\ell_p(\omega)} = \omega_n^2$, we get $\inf_{n\ge0}\omega_n > 0$. Finally, using (2.4), we see $\sup_{n\ge0}\omega_n = \infty$.

Another quite important example is the Hölder continuous functions of order α.

Example 2.7. By λ_α, $\alpha > 0$, we denote the Banach algebra of *Hölder continuous functions of order* α. If $\alpha \in (0,1]$, then λ_α is the closure of the set of trigonometric polynomials under the norm

$$\|f\|_{\lambda_\alpha} := \|f\|_\infty + \sup_{s,t\in\mathbb{T},s\ne t}\frac{|f(s)-f(t)|}{|s-t|^\alpha}.$$

λ_1 is also known as the *Lipschitz class*. If $\alpha > 1$, then λ_α is the Banach algebra given by the closure of the set of trigonometric polynomials under the norm

$$\|f\|_{\lambda_\alpha} := \sum_{j=0}^{n}\frac{1}{j!}\|f^{(j)}\|_\infty + \|f^{(n)}\|_{\lambda_{\alpha-n}},$$

where n is the integer for which $n < \alpha \le n+1$.

Example 2.8. For $q \in (1,\infty)$, $p \in [1,\infty]$, and $\alpha \in (1/p,\infty)$, or $q = 1$, $p \in [1,\infty]$, and $\alpha = 1/p$, the *Besov class* B^α_{pq} is defined by

$$B^\alpha_{pq} := \left\{ f \in L_p \mid \|f\|_{B^\alpha_{pq}} := \int_0^{2\pi} \frac{\|\Delta_t^n f\|^q_{L_p(\mathbb{T})}}{|t|^{1+\alpha q}}\, dt < \infty \right\},$$

where n is an integer such that $\alpha < n$, and $\Delta_t^n := \Delta_t \Delta_t^{n-1}$ with $(\Delta_t f)(e^{is}) := f(e^{i(s+t)}) - f(e^{is})$. This definition does not depend on the choice of n, $n > \alpha$.

Example 2.9. Let $\mathcal{L}^s_p$, $p \in [1,\infty)$, and $s > 1/p$, denote the *Sobolev spaces* of order s, which denotes the space of all Fourier series f, given by (2.1), which satisfy

$$\|f\|_{\mathcal{L}^s_p} := \left(\sum_{n\in\mathbb{Z}} (1+|n|)^{sp} |a_n|^p \right)^{1/p} < \infty.$$

If s is an integer, then an equivalent norm on $\mathcal{L}^s_2$ is given by

$$\|f\|_{\mathcal{L}^s_2} := \left(\sum_{j=0}^n \|f^{(j)}\|^2_{L_2(\mathbb{T})} \right)^{1/2}, \qquad f \in \mathcal{L}^s_2,$$

and $(\mathcal{L}^s_2)_+$ coincides with the Hardy–Sobolev space on the disc; see, for example, Baratchart and Zerner [3].

Example 2.10. Let Z be one of the spaces VMO, $A(\mathbb{D}) + \overline{A(\mathbb{D})}$, $H_1(\mathbb{D}) + \overline{H_1(\mathbb{D})}$, or $L_1(\mathbb{T}) + \widetilde{L_1(\mathbb{T})}$. Here VMO is the space of functions of vanishing mean oscillation, i.e.,

$$VMO := \left\{ f \in L_1(\mathbb{T}) \mid \lim_{a \searrow 0} \sup_{\lambda(I)\le a} \frac{1}{\lambda(I)} \int_I |f - f_I| d\lambda = 0 \right\},$$

where I is an arc of $\mathbb{T}$, λ is the normalized Lebesgue measure on $\mathbb{T}$, and $f_I := \frac{1}{\lambda(I)} \int_I f d\lambda$ is the mean value on I. A suitable norm on VMO is given by

$$\|f\|_* = |\hat{f}(0)| + \sup_I \frac{1}{\lambda(I)} \int_I |f - f_I| d\lambda.$$

If f is a function on $\mathbb{T}$, then $\bar{f}$ denotes the function given by $\bar{f}(t) := \overline{f(t)}$, and $\tilde{f}$ denotes the harmonic conjugate of f. Note that for every $f \in L_1(\mathbb{T})$ there exists a harmonic conjugate $\tilde{f}$; see Zygmund [31]. By $Z^{(n)}$, $n \in \mathbb{N}$, we denote the space

$$Z^{(n)} := \{ f \text{ a distribution on } \mathbb{T} \mid f^{(n)} \in Z \}$$

provided with the norm

$$\|f\|_{Z^{(n)}} := \sum_{j=0}^n \frac{1}{j!} \|f^{(j)}\|_Z.$$

If Z is one of the spaces VMO, $A(\mathbb{D}) + \overline{A(\mathbb{D})}$, $H_1(\mathbb{D}) + \overline{H_1(\mathbb{D})}$, then $Z^{(n)}$, $n \in \mathbb{N}$, is a decomposing Banach algebra. Moreover, if $n \ge 2$, then $(L_1(\mathbb{T}) + \widetilde{L_1(\mathbb{T})})^{(n)}$ is a decomposing Banach algebra.

Example 2.11. Let $\mathcal{C}$ be a decomposing Banach algebra, and let Q be a positive constant. Furthermore, let $\{M_n\}_{n\geq 0}$ be an increasing and logarithmically convex sequence with $M_0 = 1$ for which there exists a constant $k > 0$ such that

$$\liminf_{n\to\infty} \frac{M_n}{nM_{n-k}} > 0.$$

Then the *Carleman class* $C_Q(\mathcal{C}, \{M_n\})$, the space of all infinitely differentiable functions f on $\mathbb{T}$ with

$$\lim_{n\to\infty} \frac{\|f^{(n)}\|_{\mathcal{C}}}{Q^n n! M_n} = 0,$$

equipped with the norm

$$\|f\|_{C_Q(\mathcal{C},\{M_n\})} := \sup_{n\geq 0} \frac{\|f^{(n)}\|_{\mathcal{C}}}{Q^n n! M_n},$$

is a decomposing Banach algebra.

3. Spectral factorization. Next we recall the definitions of spectral densities and spectral factors, and we show that every spectral density that is positive and bounded away from zero admits a spectral factor. Let $\mathcal{A}$ denote $L_\infty(\mathbb{T})$, $C(\mathbb{T})$, or a decomposing Banach algebra $\mathcal{B}$.

Definition 3.1. *We call $f \in \mathcal{A}$ a* spectral density *if there exists a function $h \in \mathcal{GA}_+ := \mathcal{G}(\mathcal{A}_+)$ such that*

$$f(e^{it}) = |h(e^{it})|^2.$$

h is called a spectral factor *of f.*

If $f \in \mathcal{A}$ is a spectral density, then the spectral factor is unique up to a constant of modulus 1. Next we consider whether every $f \in \mathcal{A}_{\text{pos}}$ is a spectral density.

Proposition 3.2. *Every $f \in (L_\infty(\mathbb{T}))_{\text{pos}}$ is a spectral density, and one spectral factor is given by*

$$h(z) := \exp\left\{\frac{1}{2}\frac{1}{2\pi}\int_0^{2\pi} \frac{e^{it}+z}{e^{it}-z} \log f(e^{it})\, dt\right\}. \tag{3.1}$$

Proof. h and h^{-1} are holomorphic within the disc [13, p. 61], and $f(e^{it}) = |h(e^{it})|^2$ [13, pp. 30, 32], which proves the statement. ∎

However, Treil [28] showed that there exist functions $f \in (C(\mathbb{T}))_{\text{pos}}$ for which there is no continuous spectral factor.

Proposition 3.3. *Every $f \in \mathcal{B}_{\text{pos}}$ is a spectral density, and one spectral factor is given by* (3.1).

Proof. We have $f(e^{it}) = |h(e^{it})|^2$ [13, pp. 30, 32]. We write $f = e^g$ with $g \in \mathcal{B}$. Using

$$\frac{e^{it}+z}{e^{it}-z} = \frac{2e^{it}}{e^{it}-z} - 1,$$

we get

$$\begin{aligned} h(z) &= \exp\left\{-\frac{1}{2}(P(\log f))(0)\right\} \exp\{(P(\log f))(z)\} \\ &= \exp\left\{-\frac{1}{2}(Pg)(0)\right\} \exp\{(Pg)(z)\}. \end{aligned} \tag{3.2}$$

Since $Pg \in \mathcal{B}_+$ and $\mathcal{B}_+$ is a Banach algebra, this shows $h \in \mathcal{GB}_+$. Thus the proof is completed. □

Note that it is easy to see that, for $\mathcal{A} = L_\infty$ or $\mathcal{B}$, this sufficient condition is also a necessary condition, i.e., $f \in \mathcal{A}$ is a spectral density if and only if $f \in \mathcal{A}_{\rm pos}$. We now consider the mapping $\Phi : \mathcal{A}_{\rm pos} \to \mathcal{A}_+$, given by

$$\Phi(f)(z) := \exp\left\{\frac{1}{2}\frac{1}{2\pi}\int_0^{2\pi}\frac{e^{it}+z}{e^{it}-z}\log f(e^{it})\,dt\right\}.$$

Thus Φf is a spectral factor of f, and we have

$$\Phi(f)\Phi(f) = f, \quad f \in \mathcal{A}_{\rm pos}. \tag{3.3}$$

We call the mapping Φ the *spectral factorization mapping*. We also could define the spectral factorization mapping for $\mathcal{A} = C(\mathbb{T})$. However, in this case Φ only maps from $C(\mathbb{T})_{\rm pos}$ to $H_\infty(\mathbb{D}) = L_\infty(\mathbb{T})_+$ and not to $A(\mathbb{D}) = C(\mathbb{T})_+$.

4. Continuity of the spectral factorization. In this section we discuss the continuity of the spectral factorization mapping.

DEFINITION 4.1. *We say the spectral factorization mapping is* continuous *if for every converging sequence $\{f_n\}_n \in \mathcal{A}_{\rm pos}$ with limit in $\mathcal{A}_{\rm pos}$ the sequence of corresponding spectral factors converges.*

Anderson [1] showed that the spectral factorization mapping is not continuous in $L_\infty(\mathbb{T})$ and in $C(\mathbb{T})$. This result also follows from the example given by Treil [28] of a continuous spectral density with discontinuous spectral factor. However, we shall show that the spectral factorization mapping is locally Lipschitz continuous in decomposing Banach algebra. The next lemma is taken from Clancey and Gohberg [7, Theorem 1.1, p. 35].

LEMMA 4.2. *If $f \in \mathcal{B}_{\rm pos}$, then the operators $T_f, R_f \in \mathcal{L}(\mathcal{B})$, given by*

$$T_f(x) := P(fx) + Q(x) \qquad \text{and} \qquad R_f(x) := P(x) + Q(fx), \quad x \in \mathcal{B},$$

are invertible. Moreover, $e^{-Pg} = T_f^{-1}1$ and $e^{-Qg} = R_f^{-1}1$, where $f = e^g$.

A further useful result is the following.

LEMMA 4.3. *There exists a constant $m > 0$ such that*

$$\|\overline{g}\|_{\mathcal{B}} \le m\|g\|_{\mathcal{B}}, \qquad g \in \mathcal{B}. \tag{4.1}$$

Proof. The supremum-norm $\|\cdot\|_\infty$ is also a norm on $\mathcal{B}$, and we have $\|\bar{f}f\|_\infty = \|f\|_\infty^2$. Thus $\mathcal{B}$ is a A^*-algebra. For more information on A^*-algebras we refer the reader to Rickart [24]. Since the involution is continuous on A^*-algebras (see Rickart [24, p. 187]), we get (4.1). Thus the lemma is proved. □

Our next result is a continuity result for spectral factorization. It is a stronger form of the results of Clancey and Gohberg [7, p. 205], and we provide more complete arguments than they did.

THEOREM 4.4. *The spectral factorization mapping is locally Lipschitz continuous on $\mathcal{B}$. More precisely, for every $f \in \mathcal{B}_{\rm pos}$ there exist constants $\rho, c > 0$ such that for all $f_1, f_2 \in B_\rho(f) := \{g \in \mathcal{B}_{\rm pos} \mid \|g - f\|_{\mathcal{B}} < \rho\}$ we have that f_1, f_2 are spectral densities and that*

$$\|\Phi(f_1) - \Phi(f_2)\|_{\mathcal{B}} \le c\|f_1 - f_2\|_{\mathcal{B}}, \|\Phi(f_1)^{-1} - \Phi(f_2)^{-1}\|_{\mathcal{B}} \le c\|f_1 - f_2\|_{\mathcal{B}},$$

$$\|\overline{\Phi(f_1)} - \overline{\Phi(f_2)}\|_{\mathcal{B}} \le c\|f_1 - f_2\|_{\mathcal{B}}, \|\overline{\Phi(f_1)}^{-1} - \overline{\Phi(f_2)}^{-1}\|_{\mathcal{B}} \le c\|f_1 - f_2\|_{\mathcal{B}}.$$

Proof. Let $f \in \mathcal{B}_{\text{pos}}$ be arbitrary. Using a standard result about the continuity of inversion (which, for example, can be found in Hille and Phillips [12, Theorem 2.2, p. 118]), there exist constants $\rho_1, m > 0$ such that for every $f_1, f_2 \in B_{\rho_1}(f)$, we have that f_1, f_2 are spectral densities, and $\|f_1^{-1}\|_{\mathcal{B}}, \|f_2^{-1}\|_{\mathcal{B}} \leq m$. Let $f_1, f_2 \in B_{\rho_1}(f)$ be arbitrary, and choose $g_1, g_2 \in \mathcal{B}$ such that $f_1 = e^{g_1}$ and $f_2 = e^{g_2}$. Since f_1 and f_2 are spectral densities, Lemma 4.2 shows that the operators T_{f_1} and T_{f_2} are invertible. Moreover, we have $\|T_{f_1} - T_{f_2}\| \leq \|P\|\|f_1 - f_2\|_{\mathcal{B}}$.

Using again the theorem of inversion, there exist constants $\rho_2 \in (0, \rho_1]$ and $d_0, d_1 > 0$ such that for every $f_1, f_2 \in B_{\rho_2}(f)$, we have $\|T_{f_1}^{-1}\|_{\mathcal{B}}, \|T_{f_2}^{-1}\|_{\mathcal{B}} \leq d_0$ and $\|T_{f_1}^{-1} - T_{f_2}^{-1}\| \leq d_1\|f_1 - f_2\|_{\mathcal{B}}$. Especially, $\|T_{f_1}^{-1}(1) - T_{f_2}^{-1}(1)\|_{\mathcal{B}} \leq d_1\|1\|_{\mathcal{B}}\|f_1 - f_2\|_{\mathcal{B}}$, and thus by Lemma 4.2 we get

$$\|e^{-Pg_1} - e^{-Pg_2}\|_{\mathcal{B}} \leq d_1\|1\|_{\mathcal{B}}\|f_1 - f_2\|_{\mathcal{B}}.$$

In a similar manner it can be proved that there exist constants $\rho \in (0, \rho_2]$ and $d_2, d_3 > 0$ such that for every $f_1, f_2 \in B_\rho(f)$, we have $\|e^{-Qg_1}\|_{\mathcal{B}}, \|e^{-Qg_2}\|_{\mathcal{B}} \leq d_2$ and $\|e^{-Qg_1} - e^{-Qg_2}\| \leq d_3\|f_1 - f_2\|_{\mathcal{B}}$.

Using $e^{Pg_1} = f_1 e^{-Qg_1}$ and $e^{Pg_2} = f_2 e^{-Qg_2}$, there now exists a constant $d_4 > 0$ such that for every $f_1, f_2 \in B_\rho(f)$ we have

$$\|e^{Pg_1} - e^{Pg_2}\|_{\mathcal{B}} \leq d_4\|f_1 - f_2\|_{\mathcal{B}}.$$

Since $\mathcal{B}$ is continuously embedded into $C(\mathbb{T})$, we get

$$|e^{Pg_1}(0) - e^{Pg_2}(0)| \leq \|e^{Pg_1} - e^{Pg_2}\|_{C(\mathbb{T})} \leq c\|e^{Pg_1} - e^{Pg_2}\|_{\mathcal{B}} \leq cd_4\|f_1 - f_2\|_{\mathcal{B}},$$

and thus $\inf\{|e^{-Pg_1}(0)| \mid f_1 \in B_\rho(f)\} > 0$. In a similar way it can be shown that $\inf\{|e^{Pg_1}(0)| \mid f_1 \in B_\rho(f)\} > 0$. Thus there exists a constant $d_5 > 0$, such that for every $f_1, f_2 \in B_\rho(f)$, we have

$$\left|\sqrt{e^{Pg_1}(0)} - \sqrt{e^{Pg_2}(0)}\right| \leq d_5\|f_1 - f_2\|_{\mathcal{B}}$$

and

$$\left|\sqrt{e^{-Pg_1}(0)} - \sqrt{e^{-Pg_2}(0)}\right| \leq d_5\|f_1 - f_2\|_{\mathcal{B}}.$$

Finally, using (3.2) and Lemma 4.3, the theorem is proved. □

We conclude this section with the following corollary concerning normalized coprime factorization. In the special situation that the decomposing Banach algebra $\mathcal{B}$ equals the Wiener algebra W, this result can already be found in Mäkilä and Partington [20]. We say the functions $(N, D) \in \mathcal{B}_+ \times \mathcal{B}_+$ are *coprime factors* of $G : \mathbb{D} \to \mathbb{C} \cup \{\infty\}$ if D is not identically 0, $G = N/D$, and there exist $U, V \in \mathcal{B}_+$ such that

$$UN + VD = 1 \qquad \text{on } \mathbb{D}.$$

Moreover, the functions $(N, D) \in \mathcal{B}_+ \times \mathcal{B}_+$ are called *normalized coprime factors* of $G : \mathbb{D} \to \mathbb{C} \cup \{\infty\}$ if D is not identically 0, $G = N/D$, and

$$|N|^2 + |D|^2 = 1 \qquad \text{on } \mathbb{T}.$$

COROLLARY 4.5. *Let $(N_k, D_k) \in \mathcal{B}_+ \times \mathcal{B}_+$, $k \in \mathbb{N}_0$, be coprime factors. If $N_k \to N_0$ and $D_k \to D_0$ as k tends to ∞, then there exist normalized coprime factors $(\tilde{N}_k, \tilde{D}_k) \in \mathcal{B}_+ \times \mathcal{B}_+$ of N_k/D_k such that $\tilde{N}_k \to \tilde{N}_0$ and $\tilde{D}_k \to \tilde{D}_0$ as k tends to ∞. Moreover, the convergence is locally Lipschitz continuous, i.e., for every coprime factor N, D there exist constants $\rho, c > 0$ such that for all $N_1, N_2 \in B_\rho(N) := \{g \in \mathcal{B}_+ \mid \|g - N\| < \rho\}$ and $D_1, D_2 \in B_\rho(D)$ with (N_i, D_i), $i = 1, 2$, are coprime factors; there exist normalized coprime factors $(\tilde{N}_i, \tilde{D}_i)$ of N_i/D_i, $i = 1, 2$, such that*

$$\|\tilde{N}_1 - \tilde{N}_2\|_{\mathcal{B}} \le c \max\{\|N_1 - N_2\|_{\mathcal{B}}, \|D_1 - D_2\|_{\mathcal{B}}\},$$
$$\|\tilde{D}_1 - \tilde{D}_2\|_{\mathcal{B}} \le c \max\{\|N_1 - N_2\|_{\mathcal{B}}, \|D_1 - D_2\|_{\mathcal{B}}\}.$$

Proof. Let N, D be coprime factors. Then there exist constants $\rho, c_1 > 0$ such that $\|n\|, \|d\| < c_1$ if $n \in B_\rho(N)$ and $d \in B_\rho(D)$. We now choose $N_1, N_2 \in B_\rho(N)$ and $D_1, D_2 \in B_\rho(D)$ such that (N_i, D_i), $i = 1, 2$, are coprime factors and define $F_i(e^{it}) := |N_i(e^{it})|^2 + |D_i(e^{it})|^2$, $t \in [0, 2\pi)$ and $i = 1, 2$. Thus there exists a constant $c_2 > 0$ such that

$$\|F_1 - F_2\| \le c_2 \max\{\|N_1 - N_2\|, \|D_1 - D_2\|\}. \tag{4.2}$$

The coprimeness of N_i and D_i implies that $F_i \in \mathcal{B}_{\rm pos}$. Thus F_i is a spectral density, and so there exist functions $H_i \in \mathcal{GB}_+$ with $F_i(e^{it}) = |H_i(e^{it})|^2$, $t \in [0, 2\pi)$. Defining $\tilde{N}_i := N_i H_i^{-1}$ and $\tilde{D}_i := D_i H_i^{-1}$, we get $|\tilde{N}_i(e^{it})|^2 + |\tilde{D}_i(e^{it})|^2 = 1$, and $(\tilde{N}_i, \tilde{D}_i)$ are normalized coprime factors of N_i/D_i, $i = 1, 2$. Now (4.2) together with the previous theorem completes the proof. ∎

5. Boundedness of the spectral factorization. In this section we are concerned with the question whether or not the spectral factorization mapping is bounded. Boundedness of the spectral factorization mapping guarantees that the norm of the spectral factor is small if the spectral densities are small in norm. Thus the spectral factorization mapping is bounded if and only if it is continuous at the point 0. Note that boundedness is not included in the definition of continuity, since 0 is not an element of $\mathcal{A}_{\rm pos}$. We will see in this section that in general boundedness does not imply continuity and continuity does not imply boundedness.

DEFINITION 5.1. *We say that the spectral factorization mapping is* bounded *if for every bounded sequence $\{f_n\}_n \in \mathcal{A}_{\rm pos}$ the sequence of the corresponding spectral factors $\{\Phi f_n\}_n$ is bounded.*

Clearly, the spectral factorization mapping is bounded on $L_\infty(\mathbb{T})$ and on $C(\mathbb{T})$. However, we show in this section that on almost every decomposing Banach algebra the spectral factorization mapping is unbounded. In the next proposition we give a simple equivalent condition for the spectral factorization mapping to be bounded.

PROPOSITION 5.2. *The spectral factorization mapping is bounded on the decomposing Banach algebra $\mathcal{B}$ if and only if there exists a constant $M > 0$ such that*

$$\|f\|_{\mathcal{B}}^2 \le M \|f \bar{f}\|_{\mathcal{B}}, \quad f \in \mathcal{GB}_+. \tag{5.1}$$

Proof. Since $\mathcal{B}$ is separable, the spectral factorization mapping is bounded if and only if there is a constant $M_1 > 0$ such that

$$\|\Phi f\|_{\mathcal{B}}^2 \le M_1 \|f\|_{\mathcal{B}}, \qquad f \in \mathcal{B}_{\rm pos}. \tag{5.2}$$

Thus we need to show that (5.1) is equivalent to (5.2). Assume (5.1) holds. Then for $f \in \mathcal{B}_{\rm pos}$ we have $\Phi f \in \mathcal{GB}_+$. Using (3.3), we thus get (5.2).

We now assume that (5.2) holds. Let $f \in \mathcal{GB}_+$. Defining $g := f\bar{f}$, we get $g \in \mathcal{B}_{\mathrm{pos}}$. Using (5.2) for g and the definition of a spectral factor, we get (5.1). ∎

Note that the condition (5.1) is required only for $f \in \mathcal{GB}_+$ and not for every f in $\mathcal{B}$. In a C*-algebra, condition (5.1) is even satisfied for every $f \in \mathcal{B}$ and with $M = 1$ (see, for example, Rickart [24, p. 190]). However, Rickart [24, p. 190] shows that in a decomposing Banach algebra the condition (5.1) cannot be satisfied for every $f \in \mathcal{B}$, since it would imply that $\mathcal{B}$ is isomorphic to $C(\mathbb{T})$.

Recalling the formula $\Phi(f)\overline{\Phi(f)} = f$, $f \in \mathcal{B}_{\mathrm{pos}}$ (see (3.3)), a definition of the boundedness of the spectral factorization mapping should also include that the sequence $\{\overline{\Phi f_n}\}_n$ is bounded if $\{f_n\}_n \in \mathcal{B}_{\mathrm{pos}}$ is bounded. However, this is already included in our definition, as the following proposition shows.

PROPOSITION 5.3. *Assuming that the spectral factorization mapping is bounded on $\mathcal{B}$, for every bounded sequence $\{f_n\}_n \in \mathcal{B}_{\mathrm{pos}}$ the sequence $\{\overline{\Phi f_n}\}_n$ is bounded.*

Proof. The proof follows from Lemma 4.3. ∎

The next theorem provides us with an easy, checkable sufficient condition for the spectral factorization mapping to be unbounded on a decomposing Banach algebra $\mathcal{B}$.

THEOREM 5.4. *If the spectral factorization mapping is bounded on $\mathcal{B}$, then the Wiener algebra W can be continuously embedded into $\mathcal{B}$, i.e., $W \subset \mathcal{B}$ and*

$$\|f\|_{\mathcal{B}} \le c\|f\|_W, \qquad f \in W,$$

for some constant $c > 0$.

Proof. Since the spectral factorization mapping is bounded on $\mathcal{B}$, there exists a constant $M > 0$ such that (5.1) holds. Equation (4.1) shows that there exists a constant $m > 0$ such that

$$\|e^{-in\cdot}\|_{\mathcal{B}} \le m\|e^{in\cdot}\|_{\mathcal{B}}, \qquad n > 0. \tag{5.3}$$

Defining $f_n \in \mathcal{GB}_+$ by

$$f_n(e^{it}) := 2 + e^{int}, \qquad t \in [0, 2\pi],$$

we get

$$(\|e^{in\cdot}\|_{\mathcal{B}} - 2) \le \|f_n\|_{\mathcal{B}} \le M^{1/2}\|f_n\bar{f}_n\|_{\mathcal{B}}^{1/2} \le M^{1/2}(5 + 2(1+m)\|e^{in\cdot}\|_{\mathcal{B}})^{1/2}, \quad n > 0.$$

However, this can only hold if $\sup_{n\ge 0}\|e^{in\cdot}\|_{\mathcal{B}} < \infty$. This, together with (5.3), implies

$$c := \sup_{n\in\mathbb{Z}}\|e^{in\cdot}\|_{\mathcal{B}} < \infty.$$

Choosing an arbitrary trigonometric polynomial p, i.e.,

$$p(e^{it}) = \sum_{n=-N}^{N} a_n e^{int}, \quad t \in [0, 2\pi],$$

we get

$$\|p\|_{\mathcal{B}} \le \sum_{n=-N}^{N} |a_n|\|e^{in\cdot}\|_{\mathcal{B}} \le c\|p\|_W.$$

Since the trigonometric polynomials are dense in W and $\mathcal{B}$, the theorem is proved. ∎

COROLLARY 5.5. *If* $\lim_{n\to+\infty}\|e^{in\cdot}\|_{\mathcal{B}}=\infty$, *then the spectral factorization mapping is unbounded on* $\mathcal{B}$.

COROLLARY 5.6. *The spectral factorization mapping is not bounded on the following decomposing Banach algebras:*

1. $W(\omega)$ *if* $\sup_{n\in\mathbb{N}_0}\omega_n=\infty$.
2. $\mathcal{F}\ell_p(\omega)$, $p\in(1,\infty)$.
3. λ_α, $\alpha\in(0,\infty)$.
4. B^α_{pq}, $q\in(1,\infty)$, $p\in[1,\infty]$, *and* $\alpha\in(1/p,\infty)$.
5. $\mathcal{L}^s_p$, $p\in[1,\infty)$, *and* $s>1/p$.
6. $VMO^{(n)}$, $(A(\mathbb{D})+\overline{A(\mathbb{D})})^{(n)}$, $(H_1(\mathbb{D})+\overline{H_1(\mathbb{D})})^{(n)}$, *and* $(L_1(\mathbb{T})+\widetilde{L_1(\mathbb{T})})^{(k)}$, *if* $n\geq 1$ *and* $k\geq 2$.
7. $C_Q(\mathcal{C},\{M_n\})$.

Proof. Let us first consider the space λ_α for $\alpha\in(0,1]$. In order to show that the spectral factorization mapping is not bounded on λ_α, it remains to show

$$\lim_{n\to+\infty}\|e^{in\cdot}\|_{\lambda_\alpha}=\infty;$$

see the previous corollary. However, this follows from the calculation

$$\|e^{in\cdot}\|_{\lambda_\alpha}\geq\sup_{s,t\in[0,2\pi],s\neq t}\frac{|e^{ins}-e^{int}|}{|e^{is}-e^{it}|}|e^{is}-e^{it}|^{1-\alpha}=2^{1-\alpha}n.$$

Since $\lambda_\alpha\subset\lambda_1$, $\alpha\in(1,\infty)$, part 3 holds. In Peller and Khrushchev [23] it is proved that $\mathcal{L}^s_p\subset\lambda_\beta$, $\beta\in(0,s-1/p)$, and $B^\alpha_{pq}\subset\lambda_\beta$, $\beta\in(0,\alpha-1/p)$. Thus parts 4 and 5 hold. Finally, parts 1 and 2 follow from Corollary 5.5, and parts 6 and 7 are easy to see. □

In order to show that on $B^{1/p}_{p1}$, $p\in[1,\infty]$, the spectral factorization mapping is not bounded, we give another easy, checkable sufficient condition.

PROPOSITION 5.7. *If there exists a constant* $m>0$ *such that*

$$\|\bar f f\|_{\mathcal{B}}\leq m\|f^2\|_{\mathcal{B}},\qquad f\in\mathcal{GB}_+, \tag{5.4}$$

then the spectral factorization mapping is unbounded on $\mathcal{B}$.

Proof. We assume that the spectral factorization mapping is bounded, i.e., there exists a constant $M>0$ (Proposition 5.2) such that

$$\|f\|^2_{\mathcal{B}}\leq M\|f\bar f\|_{\mathcal{B}},\qquad f\in\mathcal{GB}_+.$$

This, together with (5.4), implies

$$\|f\|^2_{\mathcal{B}}\leq c\|f^2\|_{\mathcal{B}},\qquad f\in\mathcal{GB}_+,$$

where $c:=mM$. Applying this estimate repeatedly, we get for $f\in\mathcal{GB}_+$

$$\|f\|_{\mathcal{B}}\leq c^{1/2}\|f^2\|^{1/2}_{\mathcal{B}}\leq c^{1/2}c^{1/4}\|f^4\|^{1/4}_{\mathcal{B}}\leq\cdots\leq c\|f^{2^n}\|^{1/2^n}_{\mathcal{B}},$$

and thus by the spectral radius formula, that is, $\|f\|_\infty=\lim_{n\to\infty}\|f^n\|^{1/n}_{\mathcal{B}}$, we get $\|f\|_{\mathcal{B}}\leq c\|f\|_\infty$ for $f\in\mathcal{GB}_+$. Next we extend this estimate to all functions $f\in\mathcal{B}_+$. Let $f\in\mathcal{B}_+$. Then for $\delta\in(\|f\|_\infty,\|f\|_{\mathcal{B}})$ we get that δ is not in the spectrum of f, and thus $f+\delta\in\mathcal{GB}_+$. This implies

$$\|f\|_{\mathcal{B}}-\delta\leq\|f+\delta\|_{\mathcal{B}}\leq c\|f+\delta\|_\infty\leq c(\|f\|_\infty+\delta).$$

Thus we get $\|f\|_{\mathcal{B}} - \delta \le c(\|f\|_\infty + \delta)$, and, taking the limit $\delta \to \|f\|_\infty$, we get

$$\|f\|_{\mathcal{B}} \le (2c+1)\|f\|_\infty.$$

Thus $\|\cdot\|_{\mathcal{B}}$ and $\|\cdot\|_\infty$ are equivalent norms on $\mathcal{B}_+$. Next we show that $\mathcal{B} = C(\mathbb{T})$. Since we have assumed that the spectral factorization mapping is bounded on $\mathcal{B}$, Corollary 5.5 and Lemma 4.3 show

$$d := \sup_{n\in\mathbb{Z}} \|e^{in\cdot}\|_{\mathcal{B}} < \infty.$$

For an arbitrary trigonometric polynomial p, i.e.,

$$p(e^{it}) := \sum_{n=-N}^{N} a_n e^{int}, \qquad t \in [0, 2\pi],$$

we get

$$\begin{aligned}\|p\|_{\mathcal{B}} &= \|e^{-iN\cdot} e^{iN\cdot} p\|_{\mathcal{B}} \le d\|e^{iN\cdot} p\|_{\mathcal{B}} \\ &\le d(2c+1)\|e^{iN\cdot} p\|_\infty = d(2c+1)\|p\|_\infty.\end{aligned}$$

The trigonometric polynomials are dense in $\mathcal{B}$ and in $C(\mathbb{T})$, and $\mathcal{B} \subset C(\mathbb{T})$, and so we get $\mathcal{B} = C(\mathbb{T})$. Since $C(\mathbb{T})$ is not a decomposing Banach algebra, our assumption can no longer hold, and so the spectral factorization mapping is not bounded on $\mathcal{B}$. □

COROLLARY 5.8. *The spectral factorization mapping is unbounded on $B_{p,1}^{1/p}$, where $p \in [1, \infty]$.*

Proof. Using the triangle inequality

$$||a|^2 - |b|^2| \le |a^2 - b^2|, \qquad a, b \in \mathbb{C},$$

it is easy to see that the Banach algebras $B_{p,1}^{1/p}$ satisfy (5.4), and so, using Theorem 5.7, the spectral factorization mapping is not bounded on these decomposing Banach algebras. □

However, Theorem 5.7 does not imply that the spectral factorization mapping is unbounded on any decomposing Banach algebra, since (5.4) is not satisfied by every decomposing Banach algebra. In the next example we show that the Wiener algebra does not satisfy (5.4).

Example 5.9. Let the sequence $(p_n)_{n\in\mathbb{N}_0}$ of trigonometric polynomials be given by

$$\begin{aligned}p_0(e^{it}) &:= 1, \\ p_{n+1}(e^{it}) &:= p_n(e^{it})\big(8 - 4e^{i2\cdot 7^n t} - 2e^{i4\cdot 7^n t} - e^{i6\cdot 7^n t}\big).\end{aligned}$$

It is easy to see that $p_n \in \mathcal{G}W_+$. Moreover, we get

$$\begin{aligned}(p_{n+1}(e^{it}))^2 = (p_n(e^{it}))^2\big(&64 - 64e^{i2\cdot 7^n t} - 16e^{i4\cdot 7^n t} \\ &+12e^{i8\cdot 7^n t} + 4e^{i10\cdot 7^n t} + e^{i12\cdot 7^n t}\big)\end{aligned}$$

and

$$\begin{aligned}|p_{n+1}(e^{it})|^2 = |p_n(e^{it})|^2(&-8e^{-i6\cdot 7^n t} - 12e^{-i4\cdot 7^n t} - 22e^{-i2\cdot 7^n t} \\ &+85 - 22e^{i2\cdot 7^n t} - 12e^{i4\cdot 7^n t} - 8e^{i6\cdot 7^n t}).\end{aligned}$$

Noting that p_n is a polynomial of degree $7^n - 1$, it is easy to see that

$$\|(p_n(e^{it}))^2\|_W = 161\|(p_{n-1}(e^{it}))^2\|_W = 161^n$$

and

$$\| \, |p_n(e^{it})|^2\|_W = 169\| \, |p_{n-1}(e^{it})|^2\|_W = 169^n,$$

and so

$$\||p_n(e^{it})|^2\|_W = \left(\frac{169}{161}\right)^n \| \, (p_n(e^{it}))^2\|_W.$$

This shows that (5.4) does not hold for the Wiener class.

The only decomposing Banach algebra considered in section 2 which can be embedded into W and which does not satisfy (5.4) is the weighted Wiener algebra $W(\omega)$ with $\sup_{n\in\mathbb{N}_0} \omega_n < \infty$. However, the next theorem shows that the spectral factorization mapping is not bounded on $W(\omega)$.

THEOREM 5.10. *The spectral factorization mapping is not bounded on $W(\omega)$ with $\sup_{n\in\mathbb{N}_0} \omega_n < \infty$. In particular, the spectral factorization mapping is not bounded on W.*

Proof. $\sup_{n\in\mathbb{N}_0} \omega_n < \infty$ implies $W = W(\omega)$ and

$$k\|f\|_W \le \|f\|_{W(\omega)} \le K\|f\|_W, \qquad f \in W,$$

for some constants $k, K > 0$, and so it remains to show that the spectral factorization mapping is not bounded on W.

Rudin [25] and Shapiro [26] showed independently of each other that there exist polynomials

$$\tilde{h}_n(e^{it}) = \sum_{j=1}^{n} a_j e^{ijt}, \qquad t \in [0, 2\pi], n \in \mathbb{N},$$

such that $\{a_j\}_{j\in\mathbb{N}} \subseteq \{-1, 1\}$ and $\|\tilde{h}_n\|_{H_\infty} < 5n^{1/2}$. These polynomials are also called Rudin–Shapiro polynomials [17]. Clearly, $\tilde{h}_n \in W_+$ and $\|\tilde{h}_n\|_W = n$. We now define

$$h_n := 10n^{-1/4} + \frac{1}{n^{3/4}}\tilde{h}_n.$$

By this definition, we get $h_n \in W_+$, $\|h_n\|_{H_\infty} < 15n^{-1/4}$, and $\|\tilde{h}_n\|_W = 10n^{-1/4} + n^{1/4}$. Moreover, $|h_n(t)| \ge 5n^{-1/4}$, and so $h_n \in \mathcal{G}W_+$.

We now define $f_n \in W$ by $f_n(e^{it}) = |h_n(e^{it})|^2 = \overline{h_n(e^{it})}h_n(e^{it})$. This definition implies $f_n(t) > 0$, $t \in \mathbb{T}$, and so f_n is a spectral density and $\Phi f_n = b_n h_n$ for some constant b_n of modulus 1. We get that $\|f_n\|_{L_\infty(\mathbb{T})} < 225n^{-1/2}$ and $\|f_n\|_{L_2(\mathbb{T})} < \sqrt{2\pi}225n^{-1/2}$. This implies

$$\|f_n\|_W \le \|f_n\|_{L_2(\mathbb{T})} \left(\sum_{j=-n}^{n} 1^2\right)^{1/2} \le 225\sqrt{2\pi}\left(2 + \frac{1}{n}\right)^{1/2}.$$

This proves that the sequence $\{f_n\}_n$ is bounded in W, whereas the corresponding sequence of spectral factors is not bounded in W. Thus the spectral factorization mapping is not bounded in W. ∎

6. Spectral factorization on the right half-plane. Up until now we have studied the spectral factorization mapping on the disc. In what follows we will see that similar results are available on the right half-plane. Some of the results below can be obtained by transforming to the disc by a Möbius mapping such as $z = (1-s)/(1+s)$; for example, it was shown [4] that for $W(i\mathbb{R})$, the Wiener algebra on the imaginary axis, the transformed algebra is a decomposing Banach algebra on the circle. (This was used in order to prove various properties of the Nehari extension.) However, it is possible to proceed directly, and we do this now.

In order to define decomposing Banach algebras on the imaginary line, we need to introduce some definitions. By $\mathbb{C}_+$ we denote the right half-plane $\{z \in \mathbb{C} \mid \operatorname{Re} z > 0\}$, and by $BMO(\mathbb{R})$ we denote the set of all locally integrable functions f defined on $\mathbb{R}$ such that

$$\sup_I \frac{1}{|I|} \int_I |f(\omega) - f_I| \, d\omega < \infty,$$

where the supremum is taken over all intervals I of finite positive measure, and

$$f_I = \frac{1}{|I|} \int_I f(\omega) \, d\omega.$$

Moreover, by $BMOA(\mathbb{C}_+)$ we denote the set of all holomorphic functions $h : \mathbb{C}_+ \to \mathbb{C}$ such that $h(z)/(1+z)$ is an element of $H_2(\mathbb{C}_+)$ and the boundary function $h(i\omega)$ is in $BMO(\mathbb{R})$. For every function $f \in L_\infty(i\mathbb{R})$ there exists a function $h \in BMOA(\mathbb{C}_+)$ such that

$$h(s) - h(\beta) = \frac{1}{2\pi} \int_{\mathbb{R}} f(i\omega) \left[\frac{1}{i\omega - s} - \frac{1}{i\omega - \beta} \right] d\omega. \tag{6.1}$$

h is determined to within a constant. More information on the spaces $BMO(\mathbb{R})$ and $BMOA(\mathbb{C}_+)$ can be found in the books of Garnett [10] and Koosis [16].

DEFINITION 6.1. *A commutative Banach algebra $\mathcal{H} \subset L_\infty(i\mathbb{R})$ is called a* decomposing Banach algebra on the imaginary line *if the following hold.*

(H1) *$f \in \mathcal{H}$ implies $\bar{f} \in \mathcal{H}$ and $h \in \mathcal{H}$, where h is given by* (6.1).

(H2) *$\mathcal{H}$ is a unital Banach algebra with respect to pointwise multiplication on $i\mathbb{R}$.*

(H3) *The set of rational functions in $C(i\mathbb{R} \cap \{\infty\})$ is dense in $\mathcal{H}$.*

(H4) *Every multiplicative functional on $\mathcal{H}$ coincides with a functional $f \mapsto f(i\omega)$ defined as the value at some point $\omega \in \mathbb{R} \cup \{\infty\}$.*

Let $P \in \mathcal{L}(\mathcal{H})$ be given by $Pf := h$, where h is given by (6.1) with $h(i\infty) = f(i\infty)$. Similar to Proposition 2.2 and Proposition 2.3, we get the following proposition.

PROPOSITION 6.2. *$\mathcal{H}$ is continuously embedded in $C(i\mathbb{R} \cup \{i\infty\})$, the space of maximal ideals of the algebra $\mathcal{H}$ coincides with $i\mathbb{R} \cup \{i\infty\}$, and every $f \in \mathcal{H}$ with $f(i\omega) > 0$ for $\omega \in \mathbb{R} \cup \{\infty\}$ is an element of $\mathcal{GH}$.*

As in the case on the circle, there are many examples for decomposing Banach algebras on the imaginary line available. Here we introduce only one of the most important examples: the Wiener algebra on the imaginary line. For $g \in L_1(\mathbb{R})$ we denote by $\hat{g}$ the *two-sided Laplace transform* of g, which is given by

$$\hat{g}(s) := \int_{-\infty}^{\infty} e^{-st} g(t) \, dt, \qquad s \in i\mathbb{R}.$$

The two-sided Laplace transform of a function $g \in L_1(\mathbb{R})$ is an element of $C(i\mathbb{R} \cup \{i\infty\})$.

Example 6.3. The *Wiener algebra* $W(i\mathbb{R})$ *(on the imaginary line)* is the algebra of functions $f \in C(i\mathbb{R} \cup \{i\infty\})$ of the form

$$f = \hat{g} + c,$$

where c is a constant and $g \in L_1(\mathbb{R})$, and the norm of f is given by

$$\|f\|_{W(i\mathbb{R})} := |c| + \|g\|_{L_1(\mathbb{R})}.$$

Clancey and Gohberg [7] showed that $W(i\mathbb{R})$ is indeed a decomposing Banach algebra on the imaginary line.

Besides decomposing Banach algebras on the imaginary line, the spectral factorization can be defined on $L_\infty(i\mathbb{R})$ and $C(i\mathbb{R})$. Let $\mathcal{K}$ denote $L_\infty(i\mathbb{R})$, $C(i\mathbb{R})$, or a decomposing Banach algebra $\mathcal{H}$. In a similar way as on the circle, we define $\mathcal{K}_+$ and $\mathcal{K}_{\rm pos}$.

DEFINITION 6.4. *We call* $f \in \mathcal{K}$ *a* spectral density *if there exists a function* $h \in G\mathcal{K}_+$ *such that*

$$f(i\omega) = |h(i\omega)|^2.$$

h is called a spectral factor *of f.*

If $f \in \mathcal{K}$ is a spectral density, then the spectral factor is unique up to a constant of modulus 1. The following proposition can be proved in a similar way as Theorem 3.5 and Corollary 3.6 in Jacob, Winkin, Zwart [15].

PROPOSITION 6.5. *Let* $\mathcal{K} = L_\infty(i\mathbb{R})$ *or* $\mathcal{H}$. *Then every* $f \in \mathcal{K}_{pos}$ *is a spectral density, and one spectral factor is given by*

$$h(s) = \exp\left\{\frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{i\omega s - 1}{i\omega - s}\,\frac{1}{1+\omega^2}\log f(i\omega)\,d\omega\right\}, \quad \operatorname{Re} s > 0. \tag{6.2}$$

As on the circle, there exist functions $C(i\mathbb{R})_{\rm pos}$ for which there is no continuous spectral factor.

For $\mathcal{K} = L_\infty(i\mathbb{R})$ or $\mathcal{H}$, we define the spectral factorization mapping $\Phi : \mathcal{K}_{pos} \to \mathcal{K}_+$ on the right half-plane by

$$\Phi(f) := h,$$

where h is given by (6.2). We also could define the spectral factorization mapping for $\mathcal{K} = C(i\mathbb{R})$. However, in this case, Φ maps only from $C(\mathbb{T})_{\rm pos}$ to $L_\infty(\mathbb{T})_+$.

The continuity and boundedness of the spectral factorization mapping on the right half-plane is defined in a similar way as on the circle.

In Jacob, Winkin, and Zwart [14] it is shown that the spectral factorization mapping is not continuous on $L_\infty(i\mathbb{R})$ and on $C(i\mathbb{R})$, and in Jacob, Winkin, and Zwart [15] it is shown that the spectral factorization mapping is continuous on $W(i\mathbb{R})$. Following the line of Theorem 4.5 in [15], we see that the spectral factorization mapping is continuous on decomposing Banach algebras on the imaginary axis.

As on the circle, it is easy to see that the spectral factorization mapping is bounded on $L_\infty(i\mathbb{R})$ and on $C(i\mathbb{R})$. In the remaining part of this section we study the boundedness of the spectral factorization mapping on decomposing Banach algebras.

PROPOSITION 6.6. *If*

$$\sup_{n\in\mathbb{N}} \left\| \frac{1}{(\cdot+1)^n} \right\|_{\mathcal{H}} = \infty,$$

then the spectral factorization mapping is not bounded on $\mathcal{H}$.

Proof. We assume that the spectral factorization mapping is bounded on $\mathcal{H}$, i.e., there exists a constant $M > 0$ such that

$$\|f\|_{\mathcal{H}}^2 \le M\|f\bar{f}\|_{\mathcal{H}}, \qquad f \in \mathcal{GH}_+.$$

Similar to Lemma 4.3, it can be proved that there exists a constant $m > 0$ such that

$$\left\|\frac{1}{(\cdot+1)^n}\right\|_{\mathcal{H}} \le m\left\|\frac{1}{(-\cdot+1)^n}\right\|_{\mathcal{H}}, \qquad n > 0.$$

Defining $f_n \in \mathcal{GH}_+$ by

$$f_n(i\omega) := 2 + \frac{1}{(i\omega+1)^n}, \qquad \omega \in \mathbb{R},$$

we get

$$\begin{aligned}\left(\left\|\frac{1}{(\cdot+1)^n}\right\|_{\mathcal{H}} - 2\right) &\le \|f_n\|_{\mathcal{H}} \le M^{1/2}\|f_n\bar{f}_n\|_{\mathcal{H}}^{1/2} \\ &\le M^{1/2}\left(5 + 2(1+m)\left\|\frac{1}{(\cdot+1)^n}\right\|_{\mathcal{H}}\right)^{1/2}, \quad n > 0.\end{aligned}$$

However, this can only hold if $\sup_{n\ge 0}\|\frac{1}{(\cdot+1)^n}\|_{\mathcal{H}} < \infty$. Thus the proposition is proved. □

As on the circle, this sufficient condition excludes most of the examples of a decomposing Banach algebra on the imaginary line from having a bounded spectral factorization mapping, as the Hölder classes, the Sobolev classes, the Carleman classes, etc. We close this section by showing that the spectral factorization mapping is not bounded on the Wiener algebra $W(i\mathbb{R})$. The proof is based on an idea used in Rudin [25].

PROPOSITION 6.7. *The spectral factorization mapping is unbounded on $W(i\mathbb{R})$.*

Proof. First, we consider the sequences $\{q_n\}_{n\in\mathbb{N}_0}, \{p_n\}_{n\in\mathbb{N}_0} \subset L_1(\mathbb{R})$ defined by

$$\begin{aligned} q_0(t) := p_0(t) &:= \begin{cases} 1 \ , & t \in [0,1], \\ 0 \ , & \text{otherwise}, \end{cases} \\ p_{n+1} &:= p_n + q_n(\cdot - 2^n), \\ q_{n+1} &:= p_n - q_n(\cdot - 2^n). \end{aligned}$$

By this definition $\|p_n\|_{L_1(\mathbb{R})} = \|q_n\|_{L_1(\mathbb{R})} = 2^n$. Moreover, we define the sequences $\{k_n\}_{n\in\mathbb{N}_0}, \{l_n\}_{n\in\mathbb{N}_0} \subset W(i\mathbb{R})_+$ by $k_n := \hat{p}_n$ and $l_n := \hat{q}_n$. Thus $\|k_n\|_{W(i\mathbb{R})} = 2^n$, and the parallelogram law

$$|a+b|^2 + |a-b|^2 = 2|a|^2 + 2|b|^2$$

shows

$$\begin{aligned} &|k_{n+1}(i\omega)|^2 + |l_{n+1}(i\omega)|^2 \\ &= |\hat{p}_n(i\omega) + q_n\widehat{(\cdot - 2^n)}(i\omega)|^2 + |\hat{p}_n(i\omega) - q_n\widehat{(\cdot - 2^n)}(i\omega)|^2 \\ &= 2|\hat{p}_n(i\omega)|^2 + 2|q_n\widehat{(\cdot - 2^n)}(i\omega)|^2 = 2|k_n(i\omega)|^2 + 2|l_n(i\omega)|^2. \end{aligned}$$

Thus

$$|k_n(i\omega)|^2 \le |k_n(i\omega)|^2 + |l_n(i\omega)|^2 = 2^n(|k_0(i\omega)|^2 + |l_0(i\omega)|^2) = 2^{n+1}|k_0(i\omega)|^2.$$

An easy calculation shows

$$|k_0(i\omega)| = \frac{2}{|\omega|}\left|\sin\frac{\omega}{2}\right|,$$

and thus $\|k_0\bar{k}_0\|_{L_2(i\mathbb{R})} = \sqrt{\frac{2\pi}{3}}$, and

$$\|k_n\bar{k}_n\|_{L_2(i\mathbb{R})} \le \sqrt{\frac{2\pi}{3}}2^{n+1}.$$

Moreover,

$$\|k_n\|_{L_\infty(i\mathbb{R})} \le 2^{(n+1)/2}\|k_0\|_{L_\infty(i\mathbb{R})} \le 2^{(n+1)/2}\|k_0\|_{W(i\mathbb{R})} \le 2^{(n+1)/2}.$$

We now define

$$h_n := 2\cdot 2^{-n/4} + 2^{-(3n)/4}k_n.$$

By this definition, we get $h_n \in W(i\mathbb{R})_+$, and $\|h_n\|_{W(i\mathbb{R})} = 2\cdot 2^{-n/4} + 2^{n/4}$. Moreover, $|h_n(i\omega)| \ge (2-\sqrt{2})\cdot 2^{-n/4}$, and so $h_n \in \mathcal{G}W(i\mathbb{R})_+$.

We now define $f_n \in W(i\mathbb{R})$ by $f_n = |h_n|^2 = \overline{h_n}h_n$. This definition implies $f_n(t) > 0$, $t \in \mathbb{T}$, and so f_n is a spectral density, and $\Phi f_n = b_n h_n$ for some constant b_n of modulus 1. This implies

$$\begin{aligned}
\|f_n\|_{W(i\mathbb{R})} &= \|4\cdot 2^{-n/2} + 2\cdot 2^{-n}(k_n + \bar{k}_n) + 2^{-(3n)/2}k_n\bar{k}_n\|_{W(i\mathbb{R})}\\
&\le 2^{2-n/2} + 2^{1-n}\|k_n + \bar{k}_n\|_{W(i\mathbb{R})} + 2^{-(3n)/2}\|k_n\bar{k}_n\|_{W(i\mathbb{R})}\\
&\le 2^{2-n/2} + 4 + 2^{-(3n)/2}\|p_n * p_n(-\cdot)\|_{L_1(\mathbb{R})}\\
&\le 2^{2-n/2} + 4 + 2^{-(3n)/2}\left(\int_{-2^n}^{2^n} 1\,dt\right)^{1/2}\|p_n * p_n(-\cdot)\|_{L_2(\mathbb{R})}\\
&\le 2^{2-n/2} + 4 + 2^{1/2-n}\|p_n * p_n(-\cdot)\|_{L_2(\mathbb{R})}\\
&= 2^{2-n/2} + 4 + \frac{2^{1/2-n}}{\sqrt{2\pi}}\|k_n\bar{k}_n\|_{L_2(i\mathbb{R})}\\
&\le 2^{2-n/2} + 4 + \frac{2^{-n}}{\sqrt{\pi}}\sqrt{\frac{2\pi}{3}}2^{n+1} \le 10.
\end{aligned}$$

This proves that the sequence $\{f_n\}_n$ is bounded in $W(i\mathbb{R})$, whereas the corresponding sequence of spectral factors is not bounded in $W(i\mathbb{R})$. Thus the spectral factorization mapping is not bounded in $W(i\mathbb{R})$. □

7. Conclusions. We presented in this paper a detailed study of two properties of the spectral factorization mapping on L_∞ and C as well as on decomposing Banach algebras: continuity and boundedness. We showed that these two properties are quite different. On L_∞ and C the spectral factorization mapping is bounded but not

continuous, whereas on all decomposing Banach algebras under consideration this mapping is continuous and not bounded.

One of our main results concerning the boundedness of the spectral factorization mapping on decomposing Banach algebras was that the mapping can be bounded only if the Wiener algebra W can be continuously embedded into the decomposing Banach algebra $\mathcal{B}$ and $\mathcal{B} \neq W$. However, it is an open problem whether there exists a decomposing Banach algebra $\mathcal{B} \neq W$ such that W can be continuously embedded into $\mathcal{B}$.

REFERENCES

[1] B. D. O. Anderson, *Continuity of the matrix spectral factorization operation*, Math. Appl. Comput., 4 (1985), pp. 139–156.

[2] L. Baratchart, J. Leblond, J. R. Partington, and N. Torkhani, *Robust identification from band-limited data*, IEEE Trans. Automat. Control, 42 (1997), pp. 1318–1325.

[3] L. Baratchart and M. Zerner, *On the recovery of functions from pointwise boundary values in a Hardy–Sobolev class on the disc*, J. Comput. Appl. Math., 46 (1993), pp. 255–269.

[4] C. Bonnet and J. R. Partington, *Robust stabilization in the BIBO gap topology*, Internat. J. Robust Nonlinear Control, 7 (1997), pp. 429–447.

[5] F. F. Bonsall and J. Duncan, *Complete Normed Algebras*, Springer-Verlag, New York, 1973.

[6] F. M. Callier and J. Winkin, *Spectral factorization and LQ-optimal regulation for multivariable distributed systems*, Internat. J. Control, 52 (1990), pp. 55–75.

[7] K. Clancey and I. Gohberg, *Factorization of Matrix Functions and Singular Integral Operators*, Oper. Theory Adv. Appl. 3, Birkhäuser Verlag, Basel, 1981.

[8] R. F. Curtain and H. J. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1995.

[9] B. A. Francis, *A Course In H_∞ Control Theory*, Springer-Verlag, New York, 1987.

[10] J. B. Garnett, *Bounded Analytic Functions*, Academic Press, New York, 1981.

[11] M. Green and M. C. Smith, *Continuity properties of LQG optimal controllers*, Systems Control Lett., 26 (1995), pp. 33–39.

[12] E. Hille and R. S. Phillips, *Functional Analysis and Semi–Groups*, Amer. Math. Soc. Colloq. Publ. 31, AMS, Providence, RI, 1957.

[13] K. Hoffman, *Banach Spaces of Analytic Functions*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

[14] B. Jacob, J. Winkin, and H. Zwart, *Does the spectral factor depend continuously on the spectral density?*, in Mathematical Theory of Networks and Systems, A. Beghi, I. Finesso, and G. Picci, eds., Il Poligrafo, Padova, Italy, 1998, pp. 217–220.

[15] B. Jacob, J. Winkin, and H. Zwart, *Continuity of the spectral factorization on a vertical strip*, Systems Control Lett., 37 (1999), pp. 183–192.

[16] P. Koosis, *Introduction to H_p Spaces*, Cambridge University Press, Cambridge, UK, 1980.

[17] T. W. Körner, *Exercises for Fourier Analysis*, Cambridge University Press, Cambridge, UK, 1993.

[18] R. Larsen, *Banach Algebras*, Marcel Dekker, New York, 1973.

[19] D. Mustafa and K. Glover, *Minimum Entropy H_∞-Control*, Lecture Notes in Control and Inform. Sci. 146, Springer-Verlag, New York, 1990.

[20] P. M. Mäkilä and J. R. Partington, *Robust stabilization—BIBO stability, distance notions and robustness optimization*, Automatica J. IFAC, 23 (1993), pp. 681–693.

[21] J. R. Partington, *Interpolation, Identification, and Sampling*, Oxford University Press, Oxford, UK, 1997.

[22] D. J. Patil, *Representation of H^p-functions*, Bull. Amer. Math. Soc. (N.S.), 78 (1972), pp. 617–620.

[23] V. V. Peller and S. V. Khrushchev, *Hankel operators, best approximation, and stationary Gaussian processes*, Russian Math. Surveys, 37 (1982), pp. 61–144.

[24] C. E. Rickart, *Banach Algebras*, D. Van Nostrand Company, Toronto, Canada, 1960.

[25] W. Rudin, *Some theorems on Fourier coefficients*, Proc. Amer. Math. Soc., 10 (1959), pp. 855–859.

[26] H. S. Shapiro, *Extremal Problems for Polynomials and Power Series*, M.Sc. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1951.

[27] O. J. Staffans, *Quadratic optimal control of stable systems through spectral factorization*, Math. Control Signals Systems, 8 (1995), pp. 167–197.

[28] S. Treil, *A counterexample on continuous coprime factors*, IEEE Trans. Automat. Control, 39 (1994), pp. 1262–1263.
[29] M. Weiss and G. Weiss, *Optimal control of stable weakly regular linear systems*, Math. Control Signals Systems, 10 (1997), pp. 287–330.
[30] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control*, Prentice-Hall, Upper Saddle River, NJ, 1996.
[31] A. Zygmund, *Trigonometric Series*, Cambridge University Press, Cambridge, UK, 1959.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 1, pp. 107–134

TRAJECTORY CONTROL AND INTERCONNECTION OF 1D AND nD SYSTEMS*

P. ROCHA† AND J. WOOD‡

Abstract. In this paper we examine the relationship between control viewed as concatenation of trajectories and control viewed as interconnection of systems. We show that, for one-dimensional linear time-invariant systems, the ability to obtain a given subsystem by regular interconnection (a prerequisite for any feedback-type structure) is equivalent to the ability to drive any trajectory into that subsystem. However, in the case of multidimensional systems, the former is a stronger property than the latter. Trajectory controllability can, however, be expressed as a regular interconnection of behaviors in an extended variable space by introducing latent or auxiliary variables. This leads as a by-product to the notion of controlling a system by means of latent variables.

Key words. behavioral approach, multidimensional systems, controllability, regular interconnection, set-controllability

AMS subject classifications. 93A99, 93B25, 93B55

PII. S0363012999362797

1. Introduction.

The notion of feedback is ubiquitous in modern control theory. However, feedback interconnections of systems are based (albeit implicitly) on the still more fundamental concept of regular interconnection, formally introduced by Willems in the context of the behavioral approach [32]. As we will see in section 3, an interconnection is regular if the additional restrictions imposed on the plant by the controller are independent of the restrictions already present (e.g., only system inputs are restricted, as in a feedback loop). Given a plant and a desired subsystem which is a regular interconnection of the plant and some controller, we will say that the given subsystem is "achievable from the plant by regular interconnection."

Another fundamental concept in control systems theory is that of controllability. According to the behavioral definition, or the related state space definition of Kalman, controllability is the ability to drive any system trajectory into any other system trajectory (in finite time). Thus "controllability" is concerned with the driving of trajectories, whereas feedback "control" is concerned with the interconnection of systems. Clearly the two must be related. Indeed, for a one-dimensional (1D) linear state space model, controllability à la Kalman is necessary and sufficient to place a system's poles arbitrarily using state feedback. In the 1D behavioral context, Willems has shown that controllability is equivalent to achievability of any subsystem by regular interconnection.

Since many of the results in this paper answer open questions for standard 1D behaviors, our contribution is not only at the level of generalization of existing results in the 1D behavioral approach. However, most of our attention is given to the multidimensional (nD) case, in which every system trajectory is a function of two or more independent variables (rather than just one, usually time, in the standard 1D case). Such systems have classical applications in areas such as paper rolling, seismology,

*Received by the editors October 11, 1999; accepted for publication (in revised form) January 17, 2001; published electronically May 31, 2001.
http://www.siam.org/journals/sicon/40-1/36279.html
†Department of Mathematics, University of Aveiro, 3810 Aveiro, Portugal (procha@mat.ua.pt).
‡Department of Electrical and Computer Science, University of Southampton, Southampton SO17 1BJ, UK (jjw@ecs.soton.ac.uk). This author is a Royal Society University Research Fellow.

iterative learning control, and image processing, or, more generally, in the analysis of any system described by PDEs or difference equations. The theory of control for general linear nD systems is to date not very far developed, and we hope here to provide a helpful foundation for future work.

The structure of the paper is as follows. We begin by introducing some necessary background from the field of 1D/nD behavioral theory. Most of this material is standard, centering around concepts such as autonomy, controllability and input/output structures. Also, in section 2.4 we look at the structure and representation of factor behaviors, which will be crucial in what follows. Section 3 is devoted to an exposition of the theory of regular interconnection for 1D and nD systems, including the recently discovered equivalence to feedback interconnection. This section concludes with an algorithm, based on recent work by Bisiacco and Valcher, for testing achievability by regular interconnection and constructing a controller.

Then in section 4 we introduce the notion of set-controllability, which describes the ability to concatenate any trajectory of a given behavior with some trajectory of a prescribed subset. Thus this definition captures the "trajectory driving" aspect of control. We look at the characterization of set-controllability and the relationship between it and regular interconnection. In particular, we show that, for a 1D linear time-invariant system, a behavior $\mathcal{B}$ is set-controllable to a subbehavior $\mathcal{B}'$ if and only if $\mathcal{B}'$ is achievable from $\mathcal{B}$ by regular interconnection. In the nD case, however, achievability by regular interconnection is strictly stronger and is generally a very strong requirement.

We unify these two concepts in section 5 by introducing "regular extended interconnection," which is regular interconnection of extended behaviors obtained by introducing latent or auxiliary variables (i.e., through an autoregressive moving average (ARMA)–type representation of the system). This use of variables other than those which we wish to control is also a basic idea in recent work of Polderman and Mareels [16] and Trentelman and Willems [23, 24]. One of our main results is that achievability of a given subbehavior by regular extended interconnection is equivalent to set-controllability to that subbehavior in both the 1D and nD cases.

A preliminary version of this work was presented in [22].

2. Differential/difference behaviors. In this section we introduce the necessary background material on behavioral theory. Except where explicitly indicated, all of this material has appeared previously in the literature.

2.1. Behaviors, representations, and admissible signal spaces. We define a *system* to be a triple $(\mathcal{A}, q, \mathcal{B})$, where $\mathcal{A}$ is a set called the *signal space*, $q \in \mathbb{Z}_+$, and $\mathcal{B} \subseteq \mathcal{A}^q$ is called the system *behavior* [29, 30]. The elements of $\mathcal{B}$ or, more generally, of $\mathcal{A}^q$ are called *trajectories*. Throughout this paper, $\mathcal{A}$ will be some space of functions or distributions on some *signal domain* T and taking values in a field k taken to be $\mathbb{R}$ or $\mathbb{C}$. In the 1D case, $T = \mathbb{N}, \mathbb{Z}$, or $\mathbb{R}$; in the nD case, $T = \mathbb{N}^n, \mathbb{Z}^n$, or $\mathbb{R}^n$.

Our behaviors will be specified by sets of linear differential equations (or difference equations) with constant coefficients. Denote by $\mathcal{D}$ the polynomial ring $k[z_1, \ldots, z_n]$ in n indeterminates, and let $R \in \mathcal{D}^{g,q}$ for some $g, q \in \mathbb{Z}_+$ be a polynomial matrix. Let $\mathcal{A} = \mathcal{C}^\infty(\mathbb{R}^n, k)$ or $\mathcal{D}'(\mathbb{R}^n, k)$. Then the *differential behavior* defined by R is given by

$$\mathcal{B} \;=\; \ker R \;:=\; \left\{ w \in \mathcal{A}^q \;\middle|\; R\left(\frac{\partial}{\partial t_1}, \ldots, \frac{\partial}{\partial t_n}\right) w = 0 \right\}, \tag{2.1}$$

and R is said to be a *kernel representation* of $\mathcal{B}$.

In case of behaviors specified by difference equations, which we will consider simultaneously, $\mathcal{A}$ is equal to $k^{\mathbb{Z}^n}$ or $k^{\mathbb{N}^n}$, and the *difference behavior* defined by R is given by

$$\mathcal{B} \;=\; \ker R \;:=\; \{w \in \mathcal{A}^q \mid R(\sigma_1, \ldots, \sigma_n)\, w = 0\}\,, \tag{2.2}$$

where each σ_i is a backward shift operator, defined on a given trajectory w by

$$(\sigma_i w)(t_1, \ldots, t_n) \;:=\; w(t_1, \ldots, t_i + 1, \ldots, t_n). \tag{2.3}$$

We again say that R is a *kernel representation* of $\mathcal{B}$. In order to consider both differential and difference behaviors simultaneously, we will drop the operator notation and simply write $Rw = R(z_1, \ldots, z_n)w$ for a given polynomial matrix R applied to a given trajectory w, where the meaning is understood to be given by substitution in R of partial derivative or shift operators according to $\mathcal{A}$.

As an example, the two-dimensional (2D) differential behavior with three variables described by the single PDE

$$\frac{\partial w_1}{\partial t_1}(t_1, t_2) + \frac{\partial w_1}{\partial t_2}(t_1, t_2) - \frac{\partial^2 w_2}{\partial t_1 t_2}(t_1, t_2) - 2\frac{\partial^3 w_3}{\partial t_1^3}(t_1, t_2) + w_3(t_1, t_2) \;=\; 0,$$

with signal space $\mathcal{A} = \mathcal{C}^\infty(\mathbb{R}^2, \mathbb{R})$, can be written as

$$\mathcal{B} \;=\; \ker R, \qquad R \;=\; \begin{pmatrix} (z_1 + z_2) & (-z_1 z_2) & (1 - 2z_1^3) \end{pmatrix}.$$

For the case $\mathcal{A} = k^{\mathbb{Z}^n}$, the shift operators σ_i are invertible, and so we consider the polynomial matrices to have entries in the ring $k[z_1, \ldots, z_n, z_1^{-1}, \ldots, z_n^{-1}]$ of Laurent polynomials; that is, in this case only we take $\mathcal{D} = k[z_1, \ldots, z_n, z_1^{-1}, \ldots, z_n^{-1}]$ throughout. We will not refer to the Laurent polynomial ring again; however, it should be borne in mind that Laurent polynomials replace polynomials for this particular signal space.

The two continuous and two discrete signal spaces listed above all possess certain important abstract properties which give our behaviors a far richer structure than would otherwise be the case, as explained in the ground-breaking work of Oberst [14]. Throughout the remainder of the paper, we take $\mathcal{A}$ to be one of these four spaces, and all behaviors considered will be differential or difference behaviors as defined by (2.1) or (2.2) according to $\mathcal{A}$. Thus, when we write "$\mathcal{B} \subseteq \mathcal{A}^q$," we implicitly assume not only that $\mathcal{A}$ is one of the listed signal spaces but also that $\mathcal{B}$ can be described by differential or difference equations as above. In particular, all behaviors are given by linear equations with constant coefficients.

It will occasionally be necessary to refer to module-theoretic properties of behaviors. For any $r \in \mathcal{D}$ and any $w \in \mathcal{A}^q$, rw is an element of $\mathcal{A}^q$ defined as $(rI_q)w$, where rI_q is that polynomial matrix given by r times the identity.

Given a differential or difference behavior $\mathcal{B} \subseteq \mathcal{A}^q$, it will sometimes be necessary to refer to the *orthogonal module* $\mathcal{B}^\perp$, defined by

$$\mathcal{B}^\perp \;=\; \{v \in \mathcal{D}^{1,q} \mid vw \;=\; 0 \;\forall\, w \in \mathcal{B}\}, \tag{2.4}$$

where the meaning of vw is simply the $1 \times q$ polynomial matrix v applied to the trajectory w in the usual way. Thus $\mathcal{B}^\perp$ is the set (actually a submodule of $\mathcal{D}^{1,q}$) of all polynomial equations satisfied by the behavior. Given a kernel representation

$\mathcal{B} = \ker R$, we in fact have that $\mathcal{B}^\perp$ is the set of all $\mathcal{D}$-linear combinations of the rows of R.

The following important result effectively says that an inclusion of behaviors is equivalent to the reverse inclusion of the corresponding orthogonal modules.

THEOREM 2.1 (see [14, 2.61], [19, Prop. II.9]). *Let $\mathcal{B}_1 = \ker R_1$, $\mathcal{B}_2 = \ker R_2$ be two behaviors in $\mathcal{A}^q$ for some q. Then $\mathcal{B}_1$ is a* subbehavior *of $\mathcal{B}_2$ (i.e., $\mathcal{B}_1 \subseteq \mathcal{B}_2$) if and only if there exists a polynomial matrix L with $R_2 = LR_1$.*

For a behavior $\mathcal{B} \subseteq \mathcal{A}^q$, we also say that a given matrix $M \in \mathcal{D}^{q,l}$ is an *image representation* of $\mathcal{B}$ if

$$\mathcal{B} \;=\; \operatorname{im} M \;:=\; \{w \in \mathcal{A}^q \mid w = Mv \text{ for some } v \in \mathcal{A}^l\}. \tag{2.5}$$

Not all behaviors have image representations.

2.2. Free variables, autonomous behaviors, and regular behaviors.

DEFINITION 2.2 (see [19, Def. III.11]). *Let $\mathcal{B} \subseteq \mathcal{A}^q$. The set of variables $\{w_i \,|\, i \in \Phi\}$ for some subset Φ of $\{1, \ldots, q\}$ is said to be a* set of free variables *if the mapping $\rho : \mathcal{B} \to \mathcal{A}^\Phi$, which projects a trajectory onto the components of Φ, is surjective.*

The maximum size of a set of free variables is called the number of free variables *of $\mathcal{B}$ and is denoted by $m(\mathcal{B})$.*

We will often abuse this nomenclature somewhat by referring to certain variables w_i as free when we actually mean that the set of such variables is free.

If R is a kernel representation of $\mathcal{B}$, then $m(\mathcal{B}) = q - \operatorname{rank} R$ (where the rank is defined over the ring $\mathcal{D}$ or, equivalently, over the field $k(z_1, \ldots, z_n)$); see [14, Thm. 2.69]. It can also be shown [35, remarks following Def. 6] that for any behavior $\mathcal{B}$ with subbehavior $\mathcal{B}'$, we have

$$m(\mathcal{B}) \;=\; m(\mathcal{B}') + m(\mathcal{B}/\mathcal{B}'); \tag{2.6}$$

the meaning of $\mathcal{B}/\mathcal{B}'$ will be discussed in section 2.4. Now it is a well-known algebraic fact that $\mathcal{B} + \overline{\mathcal{B}}$ is isomorphic (as a module) to the factor of $\mathcal{B} \oplus \overline{\mathcal{B}}$ by $\mathcal{B} \cap \overline{\mathcal{B}}$. Consequently, we have the following equation for any $\mathcal{B}, \overline{\mathcal{B}} \subseteq \mathcal{A}^q$:

$$m(\mathcal{B}) + m(\overline{\mathcal{B}}) \;=\; m(\mathcal{B} + \overline{\mathcal{B}}) + m(\mathcal{B} \cap \overline{\mathcal{B}}). \tag{2.7}$$

We will use the following definition of autonomous systems; an equivalent condition in terms of trajectories is given for discrete behaviors in [7, 35].

DEFINITION 2.3. *A behavior $\mathcal{B}$ is called* autonomous *if it has no free variables, i.e., if $m(\mathcal{B}) = 0$.*

LEMMA 2.4 (see [7, 35]). *The following are equivalent for any behavior $\mathcal{B} = \ker R$.*

1. *$\mathcal{B}$ is autonomous.*
2. *There exists a nonzero polynomial r with $rw = 0$ for all $w \in \mathcal{B}$.*
3. *R has full column rank.*

A behavior is said to be *regular* if it has a full row rank kernel representation [19]. For $n = 1$, all behaviors are regular; this fails for $n \geq 2$, as can be seen from the simple 2D differential behavior

$$\mathcal{B} \;=\; \ker \begin{pmatrix} z_1 \\ z_2 \end{pmatrix},$$

consisting of all constant functions, which cannot be described as the kernel of a single polynomial operator.

2.3. Input/output structures and controllability. Inputs (assumed to be free) and outputs are defined in the behavioral framework as follows.

DEFINITION 2.5 (see [14, Thm. 2.69], [19, Def. IV.8], [30, Def. VIII.3], [35, Def. 12]). *A* (free) input/output structure *on the behavior $\mathcal{B}$ is a partitioning of the system variables $w = (u, y)$, such that the set of variables u is free and the* zero-input behavior $\mathcal{B}_{0,y}$*, defined by*

$$\mathcal{B}_{0,y} \;=\; \{(u, y) \in \mathcal{B} \;\mid\; u = 0\}, \tag{2.8}$$

is autonomous.

Equivalently, we can consider a partitioning $R = (-Q\ P)$ of any kernel representation R of $\mathcal{B}$ (to within a column permutation), where the columns of Q correspond to the input variables u and the columns of P correspond to the output variables y, and we have the condition

$$\text{rank } R \;=\; \text{rank } P \;=\; \text{number of columns of } P. \tag{2.9}$$

It is easy to show that the number of inputs is equal to m, the number of free variables. In particular, the number of inputs and number of outputs of a behavior are independent of the input/output structure. We will denote the number of outputs of a behavior $\mathcal{B} \subseteq \mathcal{A}^q$ by $p(\mathcal{B}) = q - m(\mathcal{B})$; then for $\mathcal{B} = \ker R$ we have

$$p(\mathcal{B}) \;=\; \text{rank } R. \tag{2.10}$$

For a given free input/output structure, any behavior $\mathcal{B}$ has a unique *transfer (function) matrix* $G \in k(z_1, \dots, z_n)^{p,m}$ characterized by the equation $PG = Q$; see [14, Thm. 2.69] and also [19, p. 75], [30] for the 2D/1D cases. Collecting together all behaviors with a given input/output structure (i.e., which have the same number of system variables and both admit the given partitioning of those variables into inputs and outputs) and the same transfer matrix with respect to that input/output structure, we obtain a *transfer class*. The transfer class turns out to be independent of the input/output structure and resulting transfer matrix, and the transfer classes partition the set of behaviors. Furthermore, each transfer class has a unique element which is minimal with respect to set inclusion [14, Thm. 7.21], [19, p. 76].

Next, recall the notion of a minimal left annihilator (MLA) [19, p. 24]. The following is not the usual definition but is equivalent to it [35, Lem. 7].

DEFINITION 2.6. *Suppose that $R \in \mathcal{D}^{g,q}$. Then R is called an MLA (of M) if there exists a matrix $M \in \mathcal{D}^{q,h}$ for some h with* $\ker R = \text{im } M$.

Every polynomial matrix has an MLA [14, Lem. 2.27], [35, Lem. 7].

It is also shown in [35, Lem. 10] that a given matrix E is an MLA (of some matrix) precisely when it satisfies the condition of generalized factor left primeness introduced in [14, Thm. 7.21], [38]. Furthermore, a full row rank matrix $E \in \mathcal{D}^{g,q}$ is an MLA precisely when E is minor left prime, i.e., when the gth order minors of E have no nontrivial common factor.

We can now define and characterize controllability. It will be convenient to first introduce the notion of concatenability of trajectories.

DEFINITION 2.7. *Let $\mathcal{B}$ be a behavior, $w^{(1)}, w^{(2)} \in \mathcal{B}$, and let T_1, T_2 be subsets of the signal domain T. We say that $w^{(1)}$ and $w^{(2)}$ are* concatenable in $\mathcal{B}$ with respect to (T_1, T_2) *if the following condition, dependent on the signal space $\mathcal{A}$, holds.*

$\mathcal{A} = \mathcal{C}^\infty(\mathbb{R}^n, k)$ or $\mathcal{A} = \mathcal{D}'(\mathbb{R}^n, k)$. There exists $w^{(0)} \in \mathcal{B}$ such that $w^{(0)}$ agrees with $w^{(1)}$ on T_1 and with $w^{(2)}$ on T_2.

$\mathcal{A} = k^T, T = \mathbb{N}^n$ or $\mathbb{Z}^n$. For any $b_1, b_2 \in T$, there exists $w^{(0)} \in \mathcal{B}$ with

$$w^{(0)}(t) = \begin{cases} w^{(1)}(t-b_1), & t \in T_1 \text{ and } t-b_1 \in T, \\ w^{(2)}(t-b_2), & t \in T_2 \text{ and } t-b_2 \in T. \end{cases} \tag{2.11}$$

Concatenability therefore expresses the possibility of being able to find a trajectory $w^{(0)}$ which looks like the trajectory $w^{(1)}$ on the set T_1 and like $w^{(2)}$ on T_2. In the discrete case, we further require that this should be possible with $w^{(1)}$ and $w^{(2)}$ "positioned anywhere" within the signal domain, i.e., with the origins of their coordinate systems located at any points b_1, b_2. This is necessary in the case $T = \mathbb{N}^n$ to allow for the possibility that the behavior may not be forward shift-invariant [36].

The behavioral definition of controllability is due to Willems in the 1D case [30] and to various authors in the nD discrete/continuous cases [15, 19, 21, 35, 36]. The metric $d(\cdot,\cdot)$ on the signal domain T in the discrete case can be arbitrary.

DEFINITION 2.8. *A differential behavior $\mathcal{B}$ is* controllable *if, for any two open sets $T_1, T_2 \subseteq \mathbb{R}^n$ with disjoint closures, any pair of trajectories of $\mathcal{B}$ are concatenable in $\mathcal{B}$ with respect to (T_1, T_2).*

A difference behavior $\mathcal{B}$ with signal domain $T = \mathbb{Z}^n$ or $T = \mathbb{N}^n$ is controllable (with separation distance ρ) *if $\rho > 0$ has the property that for any sets $T_1, T_2 \subseteq T$ with $d(T_1, T_2) > \rho$, any pair of trajectories of $\mathcal{B}$ are concatenable in $\mathcal{B}$ with respect to (T_1, T_2).*

For the purposes of the definition of controllability, in the discrete case $T = \mathbb{N}^n$ we need consider only $b_1 = 0$ in Definition 2.7, whereas for $T = \mathbb{Z}^n$ we can take $b_1 = b_2 = 0$ without loss of generality [36].

The characterization of controllability is due to many authors [4, 14, 15, 17, 19, 21, 30, 35, 36, 38].

THEOREM 2.9. *The following are equivalent for a behavior $\mathcal{B} \subseteq \mathcal{A}^q$ with kernel representation R.*

1. *$\mathcal{B}$ is controllable.*
2. *$\mathcal{B}$ is minimal in its transfer class.*
3. *$\mathcal{B}$ has an image representation.*
4. *R is an MLA.*
5. *$\mathcal{B}$ is a divisible module, i.e., for any nonzero $r \in \mathcal{D}$ and any $w \in \mathcal{B}$, there exists $w' \in \mathcal{B}$ with $rw' = w$.*
6. *$\mathcal{B}$ has no proper subbehaviors with the same number of free variables.*

Any behavior has a "controllable-autonomous decomposition" $\mathcal{B} = \mathcal{B}^c + \mathcal{B}^a$ [35, Thm. 7]; the argument given in that paper applies equally well to continuous systems. In this decomposition, the *controllable part* $\mathcal{B}^c \subseteq \mathcal{B}$ is uniquely determined as the minimal element of the transfer class of $\mathcal{B}$. An example of an *autonomous part* is $\mathcal{B}_{0,y}$ for any free input/output structure (u, y) on $\mathcal{B}$ [34, Lem. 3.9]. Note that in the nD context we do not require $\mathcal{B}^c \cap \mathcal{B}^a = 0$.

The following result has not previously appeared in the literature.

COROLLARY 2.10. *The following are equivalent for any subbehavior $\mathcal{B}'$ of $\mathcal{B}$.*

1. *$\mathcal{B}'$ contains the controllable part $\mathcal{B}^c$ of $\mathcal{B}$.*
2. *$\mathcal{B}'$ and $\mathcal{B}$ have the same number of free variables.*
3. *$\mathcal{B}'$ and $\mathcal{B}$ are in the same transfer class.*

Proof. We prove $1 \Rightarrow 2 \Rightarrow 3 \Rightarrow 1$. Suppose first that $\mathcal{B}'$ contains $\mathcal{B}^c$. Then we must have $m(\mathcal{B}^c) \leq m(\mathcal{B}') \leq m(\mathcal{B})$. However, since $\mathcal{B}^c$ is in the same transfer class as $\mathcal{B}$, it in particular has the same number of free variables, and therefore $m(\mathcal{B}) = m(\mathcal{B}')$

also. Now let $(\mathcal{B}')^c$ denote the controllable part of $\mathcal{B}'$; this is in the transfer class of $\mathcal{B}'$, so has the same number of free variables, and we find

$$m((\mathcal{B}')^c) \;=\; m(\mathcal{B}') \;=\; m(\mathcal{B}).$$

However, by [33, Thm. 7], $\mathcal{B}^c$ is the unique controllable subbehavior of $\mathcal{B}$ with the same number of free variables as $\mathcal{B}$; therefore, $\mathcal{B}^c = (\mathcal{B}')^c$. Since $\mathcal{B}'$ and $\mathcal{B}$ have the same controllable part, they are in the same transfer class. Finally, if $\mathcal{B}'$ and $\mathcal{B}$ are in the same transfer class, then they have the same controllable part, and so, in particular, $\mathcal{B}^c \subseteq \mathcal{B}'$. □

The following definition is not the original one given for strong controllability [19], which has not yet been usefully extended from 2D to nD or to the continuous case.

DEFINITION 2.11. *A matrix $R \in \mathcal{D}^{g,q}$ is called* zero left prime *if its gth order minors generate $\mathcal{D}$ as an ideal. A behavior $\mathcal{B}$ is called* strongly controllable *if it has a zero left prime kernel representation.*

LEMMA 2.12. *A behavior $\mathcal{B} \subseteq \mathcal{A}^q$ is strongly controllable if and only if it is a direct summand of $\mathcal{A}^q$ (the complementary summand being also a differential/difference behavior).*

Proof. Suppose that $\mathcal{B} = \ker R$ with R zero left prime. Then it is well known that R has a right inverse Y over $\mathcal{D}$, and it is now routine to show that $\mathcal{A}^q = \ker R \oplus \operatorname{im} Y$. Conversely, if $\mathcal{B}$ is a direct summand of $\mathcal{A}^q$, then the "orthogonal module" $\mathcal{B}^\perp$ of all system equations is a direct summand of $\mathcal{D}^{1,q}$. A standard algebraic argument now yields that there exists a polynomial matrix R which is zero left prime such that the rows of R generate $\mathcal{B}^\perp$. Hence $\ker R = \mathcal{B}$. □

Any zero left prime matrix is minor left prime so, in particular, an MLA; therefore, any strongly controllable behavior is also controllable. Since minor left primeness and zero left primeness are equivalent in the 1D case, the converse holds for 1D systems but not for $n \geq 2$. Strong controllability as defined above has previously been termed "rectifiability" and is equivalent to the concepts of free-controllability and flatness due to Fliess and coworkers (e.g., [6, 12]).

2.4. Factors and sums of behaviors. Due to the categorical duality of Oberst, if $\mathcal{B}' \subseteq \mathcal{B} \subseteq \mathcal{A}^q$ are behaviors, then the factor space (module) $\mathcal{B}/\mathcal{B}'$ also admits the structure of a behavior [14, Thm. 2.56(iii)]. This can be seen by choosing a kernel representation $R' \in \mathcal{D}^{g',q}$ for $\mathcal{B}'$. The restriction of this operator map from $\mathcal{A}^q$ to $\mathcal{B}$ has kernel $\mathcal{B}'$, and so its image is isomorphic to $\mathcal{B}/\mathcal{B}'$:

$$R'\mathcal{B} \;\cong\; \mathcal{B}/\mathcal{B}'. \tag{2.12}$$

This is an isomorphism of behaviors in the sense that it is the dual of an isomorphism of the corresponding finitely generated modules [33, Sect. 2.5]. Such isomorphisms preserve many important system-theoretic properties such as controllability, strong controllability, autonomy, number of free variables, etc., and for most purposes we may consider them to be "the same behavior." Note that $R'\mathcal{B}$ is dependent upon the choice of R'; different $R'\mathcal{B}$'s may even have different numbers of system variables.

The next result, adapted from [34, Thm. 5.7], demonstrates the construction of a kernel representation for a factor behavior.

LEMMA 2.13. *Let $\mathcal{B}' \subseteq \mathcal{B}$ be behaviors, where $\mathcal{B}'$ has the kernel representation R', and KR' is a kernel representation of $\mathcal{B}$ for some K. Let C be an MLA of R', and set*

$$L \;=\; \begin{pmatrix} K \\ C \end{pmatrix}.$$

Then L is a kernel representation of $R'\mathcal{B} \cong \mathcal{B}/\mathcal{B}'$. In the case where R' has full row rank, K itself is a kernel representation of $R'\mathcal{B} \cong \mathcal{B}/\mathcal{B}'$.

Proof. For the first claim, follow the first part of the proof of [34, Thm. 5.7], which does not depend upon the special choice $\mathcal{B}' = \mathcal{B}^c$ in that result. The second claim is immediate from the fact that a full row rank matrix has 0 as an MLA. ∎

Alternatively, if given two kernel representations $\mathcal{B} = \ker R, \mathcal{B}' = \ker R'$, construct an MLA $(C\ C')$ of $\binom{R'}{R}$. By the procedure for elimination of variables [8], [14, Cor. 2.38], $\ker C = R'\mathcal{B}$.

Of particular interest are the factors of controllable or autonomous behaviors. If $\mathcal{B}$ is controllable and $\mathcal{B}'$ is a subbehavior of $\mathcal{B}$, then it is easy to show from the divisibility characterization in Theorem 2.9 that $\mathcal{B}/\mathcal{B}'$ is also controllable. Similarly, from condition 2 in Lemma 2.4 we find that if $\mathcal{B}$ is autonomous, then so is $\mathcal{B}/\mathcal{B}'$.

We can also construct a representation for the sum of two given behaviors; this technique is implicit in some of the work of Valcher, e.g., [25, Lem. 4.3].

LEMMA 2.14. *Suppose $\mathcal{B} = \ker R, \overline{\mathcal{B}} = \ker \overline{R}$ are subbehaviors of $\mathcal{A}^q$. Let $(C\ \overline{C})$ be an MLA of $\binom{R}{\overline{R}}$. Then $CR = -\overline{C}\,\overline{R}$ is a kernel representation of $\mathcal{B} + \overline{\mathcal{B}}$.*

Proof. Let $R, \overline{R}, C$, and $\overline{C}$ be as stated. Suppose first that $CRw = 0$. Then

$$\begin{pmatrix} Rw \\ 0 \end{pmatrix} \in \ker (C\ \overline{C}) = \operatorname{im} \begin{pmatrix} R \\ \overline{R} \end{pmatrix},$$

so there exists w' with $Rw = Rw'$ and $\overline{R}w' = 0$. Hence $w = (w - w') + w'$ with $w - w' \in \mathcal{B}$ and $w' \in \overline{\mathcal{B}}$. Conversely if $w = w_1 + w_2$, $w_1 \in \mathcal{B}$, $w_2 \in \overline{\mathcal{B}}$, then $CRw = CRw_2 = -\overline{C}\,\overline{R}w_2 = 0$. Hence $\mathcal{B} + \overline{\mathcal{B}} = \ker CR$. ∎

3. Regular interconnection. In the behavioral approach, interconnections of systems are described by intersections of the corresponding behaviors (e.g., [3, 30, 32]). Thus if $\mathcal{B} = \ker R$ and $\overline{\mathcal{B}} = \ker \overline{R}$, then $\mathcal{B} \cap \overline{\mathcal{B}} = \ker \frac{R}{\overline{R}}$ is the behavior of the system obtained by interconnecting $\mathcal{B}$ and $\overline{\mathcal{B}}$. In some cases, the physical interconnection is not along all input/output channels, e.g., in a series interconnection. Such an interconnection can, however, still be represented in this framework by extending each of the original systems to include the variables of the other (without constraints). Such variations are discussed in detail in the work of Weiland and Stoorvogel [27, 28]. Various types of interconnections have also been discussed and related to the module-theoretic approach by Fliess and Bourlès [5]. In this dual domain, interconnection is described as a fibered sum of modules [5].

A prerequisite for an interconnection to be describable in terms of feedback is that it should be "regular" [32]. This has nothing to do with the concept of a "regular" behavior. The definition that we give here is different from the original definition of Willems, but the two are equivalent in both 1D and nD cases.

Throughout this section and the remainder of the paper, $\mathcal{B}', \mathcal{B}$, and $\overline{\mathcal{B}}$ denote three (linear, shift-invariant) differential/difference behaviors in the same space $\mathcal{A}^q$. $\mathcal{B}$ can be thought of as the plant, $\overline{\mathcal{B}}$ as the controller, and $\mathcal{B}'$ as the controlled system.

DEFINITION 3.1. *The interconnection $\mathcal{B} \cap \overline{\mathcal{B}}$ is said to be a* regular interconnection *if the sets $\mathcal{B}^\perp$ and $\overline{\mathcal{B}}^\perp$ of system equations intersect trivially.*

If $\mathcal{B}' \subseteq \mathcal{B}$ and there exists $\overline{\mathcal{B}}$ such that $\mathcal{B}' = \mathcal{B} \cap \overline{\mathcal{B}}$ and this is a regular interconnection, we say that $\mathcal{B}'$ is achievable from $\mathcal{B}$ by regular interconnection.

With this new definition, regular interconnection expresses the idea of "restricting what is not yet restricted." In a regular interconnection, the controller imposes new restrictions on the plant; it does not reimpose restrictions that are already present. In

this sense, the controller in a regular interconnection has no redundancy. A feedback controller is a simple example of a regular interconnection; the controller imposes restrictions only on the plant input, which is unrestricted in the plant. Feedback interconnections (with no assumptions of causality) can be formally defined in the current framework as follows.

DEFINITION 3.2 (see [20, 32]). *The interconnection $\mathcal{B} \cap \overline{\mathcal{B}}$ is said to be* implementable as a feedback interconnection *if there exist input/output structures (u, y) on $\mathcal{B}$ and $(\overline{u}, \overline{y})$ on $\overline{\mathcal{B}}$ such that y and $\overline{y}$ are disjoint sets of variables, and $(u \cap \overline{u}, y \cup \overline{y})$ is an input/output structure on $\mathcal{B} \cap \overline{\mathcal{B}}$.*

We now give a number of conditions characterizing regular interconnection; in particular, we show that Definition 3.1 is equivalent to the original definition given in [32]. The last condition, established in [20], shows that regular interconnection is both necessary and sufficient for the existence of a feedback control structure.

LEMMA 3.3. *Take $\mathcal{B} = \ker R$, $\overline{\mathcal{B}} = \ker \overline{R}$. Then the following are equivalent.*

1. *$\mathcal{B} \cap \overline{\mathcal{B}}$ is a regular interconnection.*
2. *$p(\mathcal{B}) + p(\overline{\mathcal{B}}) = p(\mathcal{B} \cap \overline{\mathcal{B}})$.*
3. *$\operatorname{rank} R + \operatorname{rank} \overline{R} = \operatorname{rank} \left(\begin{smallmatrix} R \\ \overline{R} \end{smallmatrix}\right)$.*
4. *$\mathcal{B} + \overline{\mathcal{B}} = \mathcal{A}^q$.*
5. *$\mathcal{B} \cap \overline{\mathcal{B}}$ can be implemented as a feedback interconnection.*

Proof. The equivalence of conditions 2 and 3 is obvious from the equality of the rank and the number of outputs. Now subtract $2q$ from each side of (2.7) to obtain

$$p(\mathcal{B} \cap \overline{\mathcal{B}}) + p(\mathcal{B} + \overline{\mathcal{B}}) = p(\mathcal{B}) + p(\overline{\mathcal{B}}). \tag{3.1}$$

This holds for any $\mathcal{B}, \overline{\mathcal{B}}$. Thus condition 2 is equivalent to the condition

$$p(\mathcal{B} + \overline{\mathcal{B}}) = 0,$$

which is trivially equivalent to the condition 4. Next, note that any equation satisfied by both $\mathcal{B}$ and $\overline{\mathcal{B}}$ is satisfied by $\mathcal{B}+\overline{\mathcal{B}}$, and vice versa. Since $\mathcal{A}^q$ is the only (linear, shift-invariant) differential/difference behavior satisfying no nontrivial system equation, conditions 1 and 4 are equivalent.

Finally, equivalence of conditions 1 and 5 is shown in [20] for the case $\mathcal{A} = k^{\mathbb{Z}^n}$; the proof techniques given there apply also to the other signal spaces. □

Note that an interconnection satisfying the conditions of Lemma 3.3 is called a feedback interconnection in [10, 31], where the term regular (feedback) interconnection is reserved for something more specific. Weiland and Stoorvogel [27, 28] have another nonequivalent definition of regular interconnection. The term "achievable" is used in a similar way by Polderman and Mareels [16] but not in the context of *regular* interconnections.

Condition 3 in Lemma 3.3 provides an algorithmic test for the regularity of a given interconnection. Condition 2 intuitively says that the plant and controller cannot share output variables in a regular interconnection (though this interpretation depends upon a suitable choice of input/output structures and is hard to formalize). Condition 4 says that any q-tuple of signals can be decomposed as a sum of a plant trajectory and a controller trajectory; further interpretation of this is still open. One immediate consequence of condition 4 is that regularity of interconnection is preserved by extending either the plant or the controller behavior. This condition provides us with an alternative test of regularity of an interconnection: if $\mathcal{B} = \ker R$, $\overline{\mathcal{B}} = \ker \overline{R}$, and $(C\ \overline{C})$ is an MLA of $\left(\begin{smallmatrix} R \\ \overline{R} \end{smallmatrix}\right)$, then by Lemma 2.14 $\mathcal{B} \cap \overline{\mathcal{B}}$ is a regular interconnection

if and only if $CR = 0$. This test says that $\mathcal{B} \cap \overline{\mathcal{B}}$ is a regular interconnection precisely when any relation on the rows of $\left(\begin{smallmatrix} R \\ \overline{R} \end{smallmatrix}\right)$ can be decomposed into a relation on the rows of R and a relation on the rows of $\overline{R}$.

Lemma 3.3 shows that regular and feedback interconnections are equal in a certain sense. Moreover, it is shown in [20] (following [32] for the 1D case) that if $\mathcal{B}' \subseteq \mathcal{B}$ is achievable from $\mathcal{B}$ by regular interconnection and if input/output structures on $\mathcal{B}'$ and $\mathcal{B}$ are given, where the inputs of $\mathcal{B}'$ are a subset of the inputs of $\mathcal{B}$, then $\mathcal{B}'$ can be achieved from $\mathcal{B}$ by control in a standard feedback loop. In other words, achievability by regular interconnection is sufficient for the existence of a feedback controller, even when inputs and outputs for plant and controlled system are assigned a priori. Since we have no new observations to make on the subject of feedback, we will discuss it no further in this paper; see [20] for further details.

Example 3.4. 1. We begin with a 1D example, taken over the signal space $\mathcal{A} = \mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$. Consider the behavior $\mathcal{B}$ and subbehavior $\mathcal{B}'$ given by

$$\mathcal{B} = \ker R, \quad R = \begin{pmatrix} z-1 & z-1 & z^2+z-2 \end{pmatrix},$$
$$\mathcal{B}' = \ker R', \quad R' = \begin{pmatrix} -z & 1 & 1 \\ z-1 & 0 & z-1 \end{pmatrix}.$$

We can achieve $\mathcal{B}'$ by interconnection with the following controller:

$$\overline{\mathcal{B}} = \ker \overline{R}, \quad \overline{R} = \begin{pmatrix} -z^2 & z & z \\ z-1 & 0 & z-1 \end{pmatrix}.$$

In fact, we have

$$\begin{pmatrix} R \\ \overline{R} \end{pmatrix} = \hat{L} R', \quad R' = \hat{K} \begin{pmatrix} R \\ \overline{R} \end{pmatrix}$$

with

$$\hat{L} = \begin{pmatrix} z-1 & z+1 \\ z & 0 \\ 0 & 1 \end{pmatrix}, \quad \hat{K} = \begin{pmatrix} -1 & 1 & z+1 \\ 0 & 0 & 1 \end{pmatrix}.$$

Unfortunately, $\mathcal{B}$ has rank $R = 1$ output, and $\overline{\mathcal{B}}$ has rank $\overline{R} = 2$ outputs, whereas $\mathcal{B} \cap \overline{\mathcal{B}} = \mathcal{B}'$ has rank $R' = 2$ outputs. The interconnection is therefore not regular, and a feedback structure connecting $\mathcal{B}$ and $\overline{\mathcal{B}}$ cannot exist.

2. For $\mathcal{A} = \mathcal{C}^\infty(\mathbb{R}^3, \mathbb{R})$, consider the interconnection

$$\mathcal{B} \cap \overline{\mathcal{B}} = \ker \left(\begin{array}{ccc} 0 & z_1 & -z_2 \\ -z_1 & 0 & z_3 \\ \hline 0 & z_2 & z_1 z_2 \\ 0 & z_3 & z_1 z_3 \end{array} \right).$$

It is easy from the rank condition to check that this interconnection is regular. The interconnection may be implemented by means of a feedback loop: take the first two variables to be outputs in $\mathcal{B}$ and inputs in $\overline{\mathcal{B}}$, and the third variable to be an input in $\mathcal{B}$ and an output in $\overline{\mathcal{B}}$; all three variables are then outputs in the intersection.

3. For the same signal space and behavior $\mathcal{B}$, look at the following interconnection with a different controller $\overline{\mathcal{B}}$:

$$\mathcal{B} \cap \overline{\mathcal{B}} = \ker \left(\begin{array}{ccc} 0 & z_1 & -z_2 \\ -z_1 & 0 & z_3 \\ \hline z_2 & -z_3 & 0 \end{array} \right).$$

The interconnected system is given by the *curl* or *rot* operator, which is singular. Since the ranks of the representations of $\mathcal{B}$ and $\overline{\mathcal{B}}$ add to up $3 > 2$, the interconnection is not regular. Indeed, $(z_1 z_2 \ -z_1 z_3 \ 0)w = 0$ is a system equation of both $\mathcal{B}$ and $\overline{\mathcal{B}}$. This interconnection cannot be implemented through feedback: the intersection has an input, which should therefore be a common input to both plant and controller, and by counting we find this is impossible without the nonfeedback phenomenon of a common output.

We now consider the condition of achievability by regular interconnection.

LEMMA 3.5. *Suppose $\mathcal{B}' \subseteq \mathcal{B} \subseteq \mathcal{A}^q$. Then the following are equivalent.*

1. *$\mathcal{B}'$ is achievable from $\mathcal{B}$ by regular interconnection.*
2. *$\mathcal{B}/\mathcal{B}'$ is a direct summand of $\mathcal{A}^q/\mathcal{B}'$.*

In the case where $\mathcal{B}'$ is a regular behavior, a further equivalent condition is that $\mathcal{B}/\mathcal{B}'$ is strongly controllable.

Also, for any $\mathcal{B}$ the following are equivalent.

(i) *$\mathcal{B}$ is strongly controllable.*

(ii) *Any $\mathcal{B}' \subseteq \mathcal{B}$ is achievable from $\mathcal{B}$ by regular interconnection.*

(iii) *The subbehavior $\{0\}$ is achievable from $\mathcal{B}$ by regular interconnection.*

Proof. If $\mathcal{B}'$ can be achieved from $\mathcal{B}$ by regular interconnection, then there exists a $\overline{\mathcal{B}} \subseteq \mathcal{A}^q$ with $\mathcal{B} \cap \overline{\mathcal{B}} = \mathcal{B}'$ and $\mathcal{B} + \overline{\mathcal{B}} = \mathcal{A}^q$. It follows immediately that $\mathcal{B}/\mathcal{B}' \cap \overline{\mathcal{B}}/\mathcal{B}' = \{0\}$ and $\mathcal{B}/\mathcal{B}' + \overline{\mathcal{B}}/\mathcal{B}' = \mathcal{A}^q/\mathcal{B}'$, i.e., $\mathcal{B}/\mathcal{B}'$ is a direct summand of $\mathcal{A}^q/\mathcal{B}'$ and conversely. Now suppose that $\mathcal{B}'$ is a regular behavior with a full row rank kernel representation R'. Then, as discussed in section 2.4, $\mathcal{B}/\mathcal{B}'$ can be identified with $R'\mathcal{B}$, and $\mathcal{A}^q/\mathcal{B}' = \operatorname{im} R'$, which by [14, pp. 24–25] equals $\mathcal{A}^{p(\mathcal{B}')}$, as R' has full row rank. Furthermore, there is a natural one-to-one correspondence between the subbehaviors of $\mathcal{A}^{p(\mathcal{B}')}$ and the subbehaviors of $\mathcal{A}^q/\mathcal{B}'$. Hence $\mathcal{B}/\mathcal{B}'$ is a direct summand of $\mathcal{A}^q/\mathcal{B}'$ if and only if $R'\mathcal{B}$ is a direct summand of $\mathcal{A}^{p(\mathcal{B}')}$, and by Lemma 2.12 this is precisely the condition of strong controllability of $R'\mathcal{B} \cong \mathcal{B}/\mathcal{B}'$.

Next, if $\mathcal{B}$ is strongly controllable, then $\mathcal{B}$ is a direct summand of $\mathcal{A}^q$, and so $\{0\}$ is achievable from $\mathcal{B}$ by regular interconnection, say, $\{0\} = \mathcal{B} \cap \overline{\mathcal{B}}$, $\mathcal{B} + \overline{\mathcal{B}} = \mathcal{A}^q$, and conversely. Hence (i) and (iii) are equivalent. Given such a controller $\overline{\mathcal{B}}$, for any given $\mathcal{B}' \subseteq \mathcal{B}$ we have $\mathcal{B} \cap (\overline{\mathcal{B}} + \mathcal{B}') = \mathcal{B}'$. Since $\mathcal{B} + (\overline{\mathcal{B}} + \mathcal{B}') = \mathcal{A}^q$, this interconnection is regular, and so $\mathcal{B}'$ is achievable from $\mathcal{B}'$ by regular interconnection. Hence (ii) and (iii) are equivalent. ∎

Note that in the 1D case every behavior is regular, and so $\mathcal{B}'$ is achievable from $\mathcal{B}$ by regular interconnection if and only if $\mathcal{B}/\mathcal{B}'$ is strongly controllable. In the nD case this is not so, and the condition does not generally apply. For example, take $\mathcal{B} = \mathcal{A}^q$. Then for any $\ker R' = \mathcal{B}' \subseteq \mathcal{B}$, $\mathcal{B}'$ is certainly achievable from $\mathcal{B}$ by regular interconnection (take the controller $\overline{\mathcal{B}} = \mathcal{B}'$), whereas $\mathcal{B}/\mathcal{B}' \cong \operatorname{im} R'$ is controllable but not in general strongly controllable.

Recall that for 1D systems controllability and strong controllability are equivalent. Equivalence of (i)–(iii) in Lemma 3.5 therefore gives us again the 1D result [32] that any subbehavior of a given 1D behavior $\mathcal{B}$ is achievable from $\mathcal{B}$ by regular interconnection if and only if $\mathcal{B}$ is controllable. In the nD case, even for $n = 2$, controllability and strong controllability are no longer equivalent, and we find that achievability by regular interconnection is generally a strong property. In particular, not every autonomous part, or even every minimal autonomous part, of a given behavior $\mathcal{B}$ can be achieved from $\mathcal{B}$ by regular interconnection, even when $\mathcal{B}$ is controllable. We see this by taking $\mathcal{B}$ to be a controllable but not strongly controllable behavior, e.g.,

$$\mathcal{B} \ = \ \ker \begin{pmatrix} z_1 - 1 & z_2 - 1 \end{pmatrix}.$$

Then $\mathcal{B}' = \{0\}$ cannot be achieved from $\mathcal{B}$ by regular interconnection.

We provide examples of behaviors which can and cannot be achieved by regular interconnection in the next section.

3.1. Constructing controllers. As shown in Lemma 3.5, achievability by regular interconnection can be tested through a direct summand condition. Conditions and an algorithm for testing whether one behavior is a direct summand of another have recently been given by Bisiacco and Valcher [2], extending some previous work of Valcher [26, Thm. 3.4]. Furthermore, this work is constructive in that it allows the construction of a complementary direct summand when one exists.

Using these techniques, we can test for achievability by regular interconnection and construct a suitable controller as follows. For brevity, we omit some of the details, which can be found by inspecting the results and proofs in [2].

1. Suppose we are given $\mathcal{B} = \ker R$ and $\mathcal{B}' = \ker R'$ with $\mathcal{B}' \subseteq \mathcal{B} \subseteq \mathcal{A}^q$. Begin by constructing a kernel representation L of $R'\mathcal{B} \cong \mathcal{B}/\mathcal{B}'$, as described in Lemma 2.13. Construct also an MLA C of R', so that $\ker C \cong \mathcal{A}^q/\mathcal{B}'$. Then we must find whether $\ker L$ is a direct summand of $\ker C$.

2. Apply the method of Bisiacco and Valcher. Construct a matrix K with $KL = C$, and extend K to $\hat{K}$ by adding rows which form an MLA of C. Now $\ker L$ is a direct summand of $\ker C$ if and only if $\hat{K}$ and L are *internally zero skew-prime*, i.e., there exists polynomial matrices X, Y with $X\hat{K} + LY = I$.

3. Internal zero skew-primeness can be tested using Gröbner basis techniques, since $X\hat{K} + LY = I$ is a system of linear equations in the entries of X and Y, and we simply want to know whether a solution exists and to find one if it does. Accordingly, if no such X, Y exist, then $\mathcal{B}'$ is not achievable from $\mathcal{B}$ by regular interconnection.

4. If, on the other hand, we can compute X and Y with $X\hat{K} + LY = I$, then we can extend this equation to a "doubly coprime factorization"

$$\begin{pmatrix} X & L \\ W_1 & W_2 \end{pmatrix} \begin{pmatrix} \hat{K} & Z_1 \\ Y & Z_2 \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix},$$

in which the two matrices on the left-hand side are mutual unimodular inverses (a unimodular matrix being a square matrix with an inverse over the polynomial ring). Since $\binom{L}{W_2}$ is zero right prime, $\ker L \cap \ker W_2 = 0$. Furthermore, $(\hat{K}\ Z_1)$ must be an MLA of $\binom{L}{W_2}$, so from Lemma 2.14 we find that $\hat{K}L$ is a kernel representation of $\ker L + \ker W_2$. However, by definition of $\hat{K}$ we have that $\ker \hat{K}L = \ker C$, so $\ker W_2$ is the complementary summand of $\ker L$ in $\ker C$.

5. Set $\overline{\mathcal{B}} = \ker W_2 R'$. It is now easy to check that $\mathcal{B} \cap \overline{\mathcal{B}} = \mathcal{B}'$ and $\mathcal{B} + \overline{\mathcal{B}} = \mathcal{A}^q$. Hence this interconnection is regular.

In the case where $\mathcal{B}'$ is regular, a simpler algorithm can be applied; we need only construct $R'\mathcal{B}$ and test to see whether it is strongly controllable. In the case where it is, we can find a full row rank kernel representation L of $R'\mathcal{B}$, which must be zero left prime. We then construct a matrix W_2 such that $\binom{L}{W_2}$ is unimodular (see, e.g., [1, 9]), and $W_2 R'$ gives us the required controller as in the above algorithm.

Example 3.6. 1. If (u, y) is a free input/output structure on $\mathcal{B}$, then $\mathcal{B}_{0,y}$ is achievable from $\mathcal{B}$ by regular interconnection. A suitable controller is

$$\overline{\mathcal{B}} := \left\{ \begin{pmatrix} u \\ y \end{pmatrix} \in \mathcal{A}^{m+p} \;\middle|\; u = 0 \right\}.$$

It is easy to see that $\mathcal{B} + \overline{\mathcal{B}} = \mathcal{A}^q$; from this we find by Lemma 3.3 that $\mathcal{B} \cap \overline{\mathcal{B}} = \mathcal{B}_{0,y}$ is a regular interconnection.

2. Let us return to the 1D example in 3.4.1, in which we previously found a nonregular interconnection:

$$\mathcal{B} = \ker R, \quad R = \begin{pmatrix} z-1 & z-1 & z^2+z-2 \end{pmatrix},$$
$$\mathcal{B}' = \ker R', \quad R' = \begin{pmatrix} -z & 1 & 1 \\ z-1 & 0 & z-1 \end{pmatrix}.$$

The given kernel representations for the behaviors $\mathcal{B}$ and $\mathcal{B}'$ are related by

$$R = LR', \quad L = \begin{pmatrix} z-1 & z+1 \end{pmatrix}.$$

Since R' has full row rank, L is a kernel representation for $R'\mathcal{B}$. It is zero left prime, proving that $R'\mathcal{B}$ is controllable or, equivalently, strongly controllable, and is therefore extendable by the Quillen–Suslin theorem to a unimodular matrix, e.g.,

$$\begin{pmatrix} L \\ \overline{L} \end{pmatrix} = \begin{pmatrix} z-1 & z+1 \\ 1 & 1 \end{pmatrix}.$$

A suitable controller is given by

$$\overline{\mathcal{B}} = \ker \overline{L} R' = \ker \begin{pmatrix} -1 & 1 & z \end{pmatrix},$$

which has a single output. We have $p(\mathcal{B}) + p(\overline{\mathcal{B}}) = p(\mathcal{B}')$, so the interconnection is regular. It can be implemented via a feedback loop: choose the first variable to be an input in $\overline{\mathcal{B}}$ and an output in $\mathcal{B}, \mathcal{B}'$, the second variable to be an input in $\mathcal{B}$ and an output in $\overline{\mathcal{B}}, \mathcal{B}'$, and the third variable to be an input in all three behaviors.

3. Consider the 2D behaviors given by the matrices

$$\mathcal{B} = \ker R, \quad \mathcal{B}' = \ker R', \quad R = \begin{pmatrix} z_1^2 - 1 & z_1 - z_2 \end{pmatrix}, \quad R' = \begin{pmatrix} z_1 + 1 & z_2 \\ 0 & -z_1 \end{pmatrix}.$$

In this case, R' has full row rank, and so from Lemma 3.5 we have that $\mathcal{B}'$ is achievable from $\mathcal{B}$ by regular interconnection if and only if $\mathcal{B}/\mathcal{B}'$ is strongly controllable. We have

$$R = LR', \qquad L = \begin{pmatrix} z_1 - 1 & z_2 - 1 \end{pmatrix},$$

from which L is a kernel representation of $R'\mathcal{B} \cong \mathcal{B}/\mathcal{B}'$. Since L is not zero left prime, $\mathcal{B}/\mathcal{B}'$ is not strongly controllable, and so $\mathcal{B}'$ is not achievable from $\mathcal{B}$ by regular interconnection, even though $\mathcal{B}$ is controllable.

4. Now look at the following 2D behaviors: $\mathcal{B} = \ker R, \mathcal{B}' = \ker R'$, where

$$R = \begin{pmatrix} z_1 z_2 & z_1 + 1 & z_2 \end{pmatrix}, \qquad R' = \begin{pmatrix} 0 & z_1 + z_2 + 1 & z_2 \\ z_1 z_2 & -z_1^2 - z_1 z_2 + 1 & -z_1 z_2 + z_2 \\ -z_1^2 - z_1 z_2 & z_1 + z_2 & 0 \end{pmatrix}.$$

Again we compute a kernel representation of $R'\mathcal{B}$,

$$R'\mathcal{B} = \ker L, \qquad L = \begin{pmatrix} z_1 & 1 & 0 \\ z_1^2 + z_1 z_2 - z_1 - z_2 & z_1 + z_2 & z_2 \end{pmatrix},$$

and an MLA C of R', which is given by the second row of L. Since R' is not full row rank (moreover $\mathcal{B}'$ is not regular), we proceed by applying the method of Bisiacco and Valcher. We have $\ker L \subseteq \ker C$, and we wish to know whether $\ker L$ is a direct

summand of $\ker C$. A kernel representation of $L(\ker C)$ is $\hat{K} = (0 \;\; 1)$, and we have that $\mathcal{B}'$ is achievable from $\mathcal{B}$ by regular interconnection if and only if $\hat{K}$ and L are internally zero skew-prime. This is indeed the case; we have

$$\left(\begin{array}{c|ccc} -1 & z_1 & 1 & 0 \\ 1 & z_1^2+z_1z_2-z_1-z_2 & z_1+z_2 & z_2 \end{array}\right) \left(\begin{array}{cc} 0 & 1 \\ \hline 1-z_2 & 1-z_2 \\ z_1z_2-z_1+1 & z_1z_2-z_1+1 \\ -z_1-z_2 & -z_1-z_2 \end{array}\right) \;=\; I,$$

and the matrix W_2 in the algorithm above is accordingly constructed as

$$\left(\begin{array}{ccc} z_1z_2 - z_1 + 1 & z_2 - 1 & 0 \\ z_1^2 + z_1 - 1 & z_1 + 1 & 1 \end{array}\right),$$

from which we obtain, as a suitable controller,

$$\overline{\mathcal{B}} \;=\; \ker \overline{R}, \qquad \overline{R} = W_2R' \;=\; \left(\begin{array}{ccc} z_1z_2^2 - z_1z_2 & z_1z_2 + 2z_2 & z_2^2 \\ z_1^2z_2 - z_1^2 & z_1^2 + 2z_1 & z_1z_2 \end{array}\right).$$

Since the ranks of R, $\overline{R}$, and $\left(\frac{R}{\overline{R}}\right)$ are 1, 1, and 2, respectively, $\mathcal{B} \cap \overline{\mathcal{B}}$ is certainly a regular interconnection. It remains to confirm that $\mathcal{B} \cap \overline{\mathcal{B}} = \mathcal{B}'$. However, we find

$$\left(\begin{array}{c} R \\ \overline{R} \end{array}\right) \;=\; U \left(\begin{array}{ccc} 0 & z_1 + z_2 + 1 & z_2 \\ z_1z_2 & -z_1^2 - z_1z_2 + 1 & -z_1z_2 + z_2 \\ -z_1^2 - z_1z_2 & z_1 + z_2 & 0 \end{array}\right) \;=\; UR',$$

where

$$U \;=\; \left(\begin{array}{ccc} z_1 & 1 & 0 \\ z_1z_2 - z_1 + 1 & z_2 - 1 & 0 \\ z_1^2 + z_1 - 1 & z_1 + 1 & 1 \end{array}\right)$$

is a unimodular matrix, proving that $\mathcal{B} \cap \overline{\mathcal{B}} = \mathcal{B}'$, as predicted by the theory.

Lomadze and Zerz [11] have recently provided a direct method for testing for achievability by regular interconnection and controller construction. Their algorithms address the problem more directly, but we will not discuss their work here, since the homological techniques involved are beyond the scope of this paper.

4. Set-controllability. In this section we introduce and characterize the new concept of set-controllability, which formalizes the idea of being able to drive any system trajectory into a prescribed trajectory set, in practice a subbehavior $\mathcal{B}'$. This concept has not previously appeared in the literature, even in the 1D case. The formalization is necessarily slightly different for the different signal spaces of interest.

DEFINITION 4.1. *Let $\mathcal{B}' \subseteq \mathcal{B} \subseteq \mathcal{A}^q$. Then we say that $\mathcal{B}$ is* set-controllable to $\mathcal{B}'$ *if the following condition, dependent on the signal space $\mathcal{A}$, holds.*

$\mathcal{A} = \mathcal{C}^\infty(\mathbb{R}^n, k)$ or $\mathcal{A} = \mathcal{D}'(\mathbb{R}^n, k)$: For any $w \in \mathcal{B}$, there exists $w' \in \mathcal{B}'$ such that, for any open sets $T_1, T_2 \subseteq \mathbb{R}^n$ with disjoint closures, w and w' are concatenable in $\mathcal{B}$ with respect to (T_1, T_2).

$\mathcal{A} = k^T, T = \mathbb{N}^n$ or $\mathbb{Z}^n$: There exists a $\rho > 0$ such that for any $w \in \mathcal{B}$, there exists $w' \in \mathcal{B}'$ such that, for any $T_1, T_2 \subseteq T$ with $d(T_1, T_2) \geq \rho$, w and w' are concatenable in $\mathcal{B}$ with respect to (T_1, T_2).

In the above definition, $w \in \mathcal{B}$ is interpreted as the given system trajectory, and $w' \in \mathcal{B}'$ is some trajectory in the desired subsystem into which w can be controlled,

by means of some driving trajectory $w^* \in \mathcal{B}$ (the concatenating trajectory $w^{(0)}$ in Definition 2.7). As in the definition of controllability, it is sufficient for the definition of set-controllability in the discrete case to consider only $b_1 = 0$ in Definition 2.7, and for $T = \mathbb{Z}^n$ we can take $b_2 = 0$ as well.

In each case, note that $w' \in \mathcal{B}'$ can be chosen independently of the regions T_1 and T_2. This sounds like a very stringent requirement, but as we will now show, set-controllability as defined above admits a variety of interesting characterizations.

4.1. Characterizing set-controllability. Here follows one of our main results.

THEOREM 4.2. *Let $\mathcal{B}' \subseteq \mathcal{B}$. The following are equivalent.*

1. *$\mathcal{B}$ is set-controllable to $\mathcal{B}'$.*
2. *$\mathcal{B}/\mathcal{B}'$ is controllable.*
3. *$\mathcal{B} = \mathcal{B}^c + \mathcal{B}'$.*
4. *Any autonomous part of $\mathcal{B}'$ is also an autonomous part of $\mathcal{B}$.*
5. *$\mathcal{B}$ has no proper subbehaviors with the same number of free variables which contain $\mathcal{B}'$.*
6. *$\mathcal{B}$ has no proper subbehaviors in the same transfer class which contain $\mathcal{B}'$.*

Proof. 1 $\Rightarrow$ 2. Suppose that $\mathcal{B}$ is set-controllable to $\mathcal{B}'$, and let R' be any kernel representation of $\mathcal{B}'$. We will prove that the behavior $R'\mathcal{B}$ is controllable. Since $R'\mathcal{B} \cong \mathcal{B}/\mathcal{B}'$, and controllability is (by the divisibility condition in Theorem 2.9) preserved by isomorphism, this is sufficient. We treat the discrete and continuous cases in turn.

The case $\mathcal{A} = k^T, T = \mathbb{N}^n$ or $\mathbb{Z}^n$: Let $\rho > 0$ be the distance given in the definition of set-controllability, and take $d(t_1, t_2) := \sum_{i=1}^n |(t_1)_i - (t_2)_i|$. Note that without loss of generality we can take R' to have entries in $\mathcal{D}$ regardless of T. Hence we can use the notation

$$R' \;=\; \sum_{a \in \mathbb{N}^n} R'_a z_1^{a_1} \cdots z_n^{a_n}$$

with coefficient matrices R'_a over k:

$$\operatorname{supp}(R') := \{a \in \mathbb{N}^n : R'_a \neq 0\}.$$

Set

$$\Delta := \max\{d(0, a) \mid a \in \operatorname{supp}(R')\}$$

and $\tau := \rho + 2\Delta$. We will prove that $R'\mathcal{B}$ is controllable with separation distance τ.

Let $v \in R'\mathcal{B}$ be arbitrary, say, $v = R'w$ for some $w \in \mathcal{B}$. Now there exists $w' \in \mathcal{B}'$ with the property given in the definition of set-controllability. Let $T_1, T_2 \in T$ be such that $d(T_1, T_2) \geq \tau$. It suffices by linearity to prove that v and 0 are concatenable in $R'\mathcal{B}$ with respect to (T_1, T_2). Therefore, choose $b_1, b_2 \in T$ arbitrarily, and extend T_1 and T_2 to T_3 and T_4, respectively, according to

$$T_3 \;:=\; \{t + a \mid t \in T_1, a \in \operatorname{supp}(R')\},$$

and similarly for T_4. Now for any $t_1 + a_1 \in T_3$, $t_2 + a_2 \in T_4$ with $t_1 \in T_1$, $t_2 \in T_2$, $a_1, a_2 \in \operatorname{supp}(R')$, we have

$$d(t_1, t_2) \;\leq\; d(t_1, t_1 + a_1) + d(t_1 + a_1, t_2 + a_2) + d(t_2 + a_2, t_2),$$

and, consequently, $d(T_3, T_4) \geq \rho$. Hence, by set-controllability, there exists $w^* \in \mathcal{B}$ satisfying

$$w^*(t) \;=\; \begin{cases} w(t-b_1) & \text{if } t \in T_3 \text{ and } t-b_1 \in T, \\ w'(t-b_2) & \text{if } t \in T_4 \text{ and } t-b_2 \in T. \end{cases} \tag{4.1}$$

Set $v^* = R'w^*$; we now find that for any $t \in T_1$ with $t - b_1 \in T$, we have

$$v^*(t) \;=\; \sum_{a \in \operatorname{supp}(R')} R'_a w^*(t+a).$$

Now $t + a - b_1$ must be in T for any $a \in \operatorname{supp}(R')$, and $t + a \in T_3$, which by (4.1) gives us

$$\begin{aligned} v^*(t) &= \sum_{a \in \operatorname{supp}(R')} R'_a w(t+a-b_1) \\ &= (R'w)(t-b_1) \\ &= v(t-b_1). \end{aligned}$$

Similarly, for any $t \in T_2$ with $t - b_2 \in T$, we have

$$v^*(t) \;=\; (R'w')(t-b_2) \;=\; 0.$$

Hence v^* drives the given trajectory v to the zero trajectory. This is sufficient to prove that $R'\mathcal{B}$ is controllable.

The case $\mathcal{A} = \mathcal{C}^\infty(\mathbb{R}^n, k)$ or $\mathcal{A} = \mathcal{D}'(\mathbb{R}^n, k)$: Let U be any open subset of $\mathbb{R}^n$, and let V be any closed subset whose interior contains the closure of U. Let $v \in R'\mathcal{B}$ be arbitrary, say, $v = R'w$ for some $w \in \mathcal{B}$. Since $\mathcal{B}$ is set-controllable to $\mathcal{B}'$, $\exists w' \in \mathcal{B}'$ and $w^* \in \mathcal{B}$ such that w^* agrees with w on U and with w' on V^c, the complement of V. Now define $v^* = R'w^* \in R'\mathcal{B}$. Since the support of $R'(w^* - w')$ is contained in the support of $w^* - w'$ (the differential operator R' is a local operator), we have that $v^* = R'(w^* - w')$ vanishes on V^c. On the other hand, $v - v^* = R'(w - w^*)$ vanishes on U, so v^* agrees with v on U. This proves that v has a "cutoff" $v^* \in R'\mathcal{B}$ with respect to U and V, which by [15, Lemma 3.3] is sufficient to prove that $R'\mathcal{B}$ is controllable.

$2 \Rightarrow 3$. Now suppose that $\mathcal{B}/\mathcal{B}'$ is controllable, and consider the behavior $\mathcal{B}/(\mathcal{B}^c + \mathcal{B}')$. This behavior is a factor of $\mathcal{B}/\mathcal{B}'$ but is also a factor of the behavior $\mathcal{B}/\mathcal{B}^c$. However, $\mathcal{B}/\mathcal{B}^c$ is autonomous by (2.6). By the remarks in section 2.4, $\mathcal{B}/(\mathcal{B}^c + \mathcal{B}')$ is therefore both controllable and autonomous, and therefore equal to $\{0\}$.

$3 \Rightarrow 1$. Suppose that $\mathcal{B} = \mathcal{B}^c + \mathcal{B}'$. Then any trajectory $w = w^c + w'$ of $\mathcal{B}$ can be driven to $\mathcal{B}'$ by driving w^c to 0. Hence $\mathcal{B}$ is set-controllable to $\mathcal{B}'$.

$3 \Leftrightarrow 4$. Let $\mathcal{B}'^a$ be an autonomous part of $\mathcal{B}'$, and let $\mathcal{B}'^c$ be the controllable part. We argue as follows:

$$\mathcal{B} \;=\; \mathcal{B}^c + \mathcal{B}' \;=\; \mathcal{B}^c + \mathcal{B}'^c + \mathcal{B}'^a. \tag{4.2}$$

Now $\mathcal{B}^c + \mathcal{B}'^c$ contains $\mathcal{B}^c$ and is contained in $\mathcal{B}$ so by Corollary 2.10 has the same number of free variables as $\mathcal{B}$ and $\mathcal{B}^c$. But it is easy to see that $\mathcal{B}^c + \mathcal{B}'^c$ is itself controllable, which by condition 6 of Theorem 2.9 implies that we must have $\mathcal{B}^c = \mathcal{B}^c + \mathcal{B}'^c$. Equation (4.2) now becomes $\mathcal{B} = \mathcal{B}^c + \mathcal{B}'^a$, i.e., $\mathcal{B}'^a$ is an autonomous part of $\mathcal{B}$. The converse is easy.

$2 \Leftrightarrow 5 \Leftrightarrow 6$. By condition 6 of Theorem 2.9, controllability of $\mathcal{B}/\mathcal{B}'$ means that no proper subbehavior of $\mathcal{B}/\mathcal{B}'$ can have the same number of free variables as $\mathcal{B}/\mathcal{B}'$.

By (2.6), an equivalent statement is 5, and the equivalence of 5 and 6 is immediate from Corollary 2.10. □

Condition 2 in Theorem 4.2 suggests that we can think of set-controllability to $\mathcal{B}'$ as "controllability to within $\mathcal{B}'$." We will use this characterization heavily in what follows.

Next, note that any $\mathcal{B}$ is set-controllable to $\mathcal{B}$. In fact, this is the only possibility for set-controllability when $\mathcal{B}$ is autonomous. Indeed, the condition "$\mathcal{B}$ is set-controllable to $\mathcal{B}' \Rightarrow \mathcal{B}' = \mathcal{B}$" characterizes autonomy. Controllability may also be characterized in terms of set-controllability (compare with Lemma 3.5 (i)–(iii)).

COROLLARY 4.3. *The following are equivalent.*

1. $\mathcal{B}$ *is controllable.*
2. $\mathcal{B}$ *is set-controllable to* $\mathcal{B}'$ *for any subbehavior* $\mathcal{B}'$.
3. $\mathcal{B}$ *is set-controllable to* $\{0\}$.

Thus controllability expresses the ability to control a given system in the above sense into any desired subsystem. Notice a more general property: if $\mathcal{B}$ is set-controllable to $\mathcal{B}'$, and $\mathcal{B}' \subseteq \mathcal{B}'' \subseteq \mathcal{B}$, then $\mathcal{B}$ is also set-controllable to $\mathcal{B}''$.

Example 4.4. 1. From condition 3 in Theorem 4.2, $\mathcal{B}$ is set-controllable to $\mathcal{B}'$ for any autonomous part $\mathcal{B}'$ of $\mathcal{B}$. In particular, $\mathcal{B}$ is set-controllable to $\mathcal{B}_{0,y}$ for any free input/output structure (u, y) (any trajectory is driven into $\mathcal{B}_{0,y}$ by setting the inputs to 0).

2. In Example 3.6.3, the subbehavior $\mathcal{B}'$ was shown not to be achievable from the controllable $\mathcal{B}$ by regular interconnection. However, we showed that $\mathcal{B}/\mathcal{B}' \cong \ker (z_1 - 1 \quad z_2 - 1)$, which by Theorem 4.2 demonstrates that $\mathcal{B}$ is set-controllable to $\mathcal{B}'$, as predicted by Corollary 4.3.

From condition 2 of Theorem 4.2 and (for example) the divisibility characterization of controllability, we can easily prove the intuitive result that set-controllability is transitive. In particular, if a behavior is set-controllable to a controllable subbehavior, then it must itself be controllable.

It is interesting to consider set-controllability of $\mathcal{B}$ to $\mathcal{B}'$ from the point of view of the subbehavior $\mathcal{B}'$. For any behavior $\mathcal{B}$, Theorem 4.2 tells us that $\mathcal{B}^c + \mathcal{B}'$ is set-controllable to $\mathcal{B}'$. This behavior $\mathcal{B}^c + \mathcal{B}'$ is also easily shown to be in the same transfer class as $\mathcal{B}$. Furthermore, there cannot be another behavior $\mathcal{B}^*$ in this transfer class which is also set-controllable to $\mathcal{B}'$, since $\mathcal{B}^*$ and $\mathcal{B}$ must have the same controllable part, and condition 3 of Theorem 4.2 therefore requires that $\mathcal{B}^* = \mathcal{B}^c + \mathcal{B}'$. This means that each transfer class contains at most one element which is set-controllable to $\mathcal{B}'$ for fixed $\mathcal{B}'$. When $\mathcal{B}'$ is autonomous, each transfer class contains exactly one such element. It follows that two distinct elements in the same transfer class cannot both be set-controllable to $\mathcal{B}'$ for any $\mathcal{B}'$, i.e., they have in a sense no controllability properties in common!

From these observations, we might go on to define a "$\mathcal{B}'$-transfer class" for fixed $\mathcal{B}'$, being a subset of an ordinary transfer class consisting of all those behaviors which contain $\mathcal{B}'$. Then each $\mathcal{B}'$-transfer class has a unique minimal element, and such elements are precisely the behaviors set-controllable to $\mathcal{B}'$.

It seems reasonable to assume set-controllability to $\mathcal{B}'$ as at least a necessary condition for control under any paradigm. Not only the plant, but also the controller, must be set-controllable to $\mathcal{B}'$, where $\mathcal{B}'$ is the controlled system. This means that, when designing a controller for a given problem, it suffices to identify the transfer class of that controller, for then the controller itself is uniquely identified as the single behavior in that transfer class which is set-controllable to $\mathcal{B}'$. This suggests that the

transfer function matrix approach may be an appropriate initial tool in nD controller design.

4.2. Testing for set-controllability. We now discuss algorithmic tests for set-controllability. These are based on conditions 2 and 3 in Theorem 4.2. We have the isomorphism

$$R'\mathcal{B} \cong \mathcal{B}/\mathcal{B}',$$

which enables us to test for set-controllability to $\mathcal{B}'$ by constructing a kernel representation E for $R'\mathcal{B}$ and testing to see whether E is an MLA. If E is an MLA, then $R'\mathcal{B}$ is controllable and so $\mathcal{B}$ is set-controllable to $\mathcal{B}'$, and conversely. Methods for finding a kernel representation of $R'\mathcal{B}$ have been discussed in section 2.4.

An alternative method for testing for set-controllability of $\mathcal{B}$ to $\mathcal{B}'$, which may be more efficient when a representation R^c of $\mathcal{B}^c$ is already given, is to test for the inclusion $\mathcal{B} \subseteq \mathcal{B}^c + \mathcal{B}'$. By Lemma 2.14, a kernel representation of $\mathcal{B}^c + \mathcal{B}'$ is given as $CR^c = -C'R'$, where $(C\ C')$ is an MLA of $\binom{R^c}{R'}$.

These algorithms require procedures for the following basic problems:

1. construction of a minimal left/right annihilator of a given matrix;
2. construction of a kernel representation for the controllable part of a given $\ker R$;
3. a test for inclusion $\ker R_1 \subseteq \ker R_2$, and the construction of a matrix L with $LR_1 = R_2$ when the inclusion holds;
4. a test to see whether a given R is an MLA.

The first problem is effectively the problem of constructing a syzygy module in commutative algebra, which can be solved through Gröbner basis techniques. The second problem is solved for a given R by constructing a minimal right annihilator M; the controllable part is then $\operatorname{im} M$, a kernel representation of which can be found as any MLA R^c of M. This algorithm is explained in [14, pp. 144–145], [35, 38]; see also the related algorithms in [18, section 4]. The third problem is another standard Gröbner basis problem, while the last reduces to the previous three, as it is equivalent to testing $\ker R \subseteq \ker R^c$.

4.3. Regular interconnection and set-controllability. We have now examined in outline two different paradigms for the control of 1D/nD systems. Regular interconnection is based on the notion of adding restrictions to systems and is related to classical feedback. Set-controllability, on the other hand, formalizes the concept of "on-line" control by steering system trajectories. For a 1D system, we know from the work of Willems [32] that controllability is equivalent to achievability of any subbehavior by means of regular interconnection. This leads to two immediate questions.

- What happens when the given behavior $\mathcal{B}$ is not controllable? In particular, if $\mathcal{B}$ is a 1D behavior which is set-controllable to some subbehavior $\mathcal{B}'$ (but not necessarily actually controllable), is this sufficient for $\mathcal{B}'$ to be achievable from $\mathcal{B}$ by regular interconnection? What about the converse?
- What is the situation for nD behaviors?

The next result answers these questions.

THEOREM 4.5. *If $\mathcal{B}'$ can be achieved from $\mathcal{B}$ by regular interconnection, then $\mathcal{B}$ is set-controllable to $\mathcal{B}'$. The converse holds for $n = 1$ but not for $n \geq 2$.*

Proof. Suppose that $\mathcal{B}'$ can be achieved from $\mathcal{B}$ by regular interconnection. Then by Lemma 3.5, $\mathcal{B}/\mathcal{B}'$ is a direct summand of the controllable behavior $\mathcal{A}^q/\mathcal{B}'$ and so is isomorphic to a factor of it. By the divisibility condition, any factor of a controllable

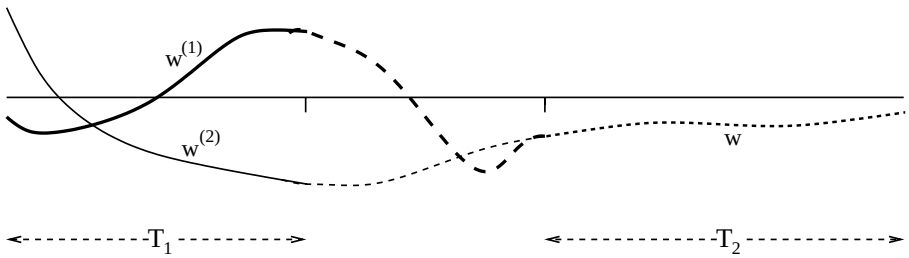


FIG. 4.1. *Mergeable behaviors.*

behavior is controllable, and so $\mathcal{B}$ is set-controllable to $\mathcal{B}'$. In the 1D case, suppose, conversely, that $\mathcal{B}$ is set-controllable to $\mathcal{B}'$. Then $\mathcal{B}/\mathcal{B}'$ is controllable, or, equivalently, strongly controllable, $\mathcal{B}'$ is regular, and so by Lemma 3.5 $\mathcal{B}'$ is achievable from $\mathcal{B}$ by regular interconnection. For any $n \geq 2$, the following is a controllable but not strongly controllable behavior:

$$\mathcal{B} \;=\; \ker\,(z_1 - 1 \;\; z_2 - 1).$$

Hence $\{0\}$ is not achievable from $\mathcal{B}$ by regular interconnection, although $\mathcal{B}$ is set-controllable to $\{0\}$. □

More generally, we have seen at the end of section 3 that not every autonomous part is achievable from a given (controllable) system by regular interconnection. On the other hand, a given behavior is set-controllable to any autonomous part.

The conclusion here is that "on-line" control to a desired subsystem is often possible in situations where there exists no suitable controller forming a regular interconnection with the plant. On the other hand, for a 1D system set-controllable to a given subsystem, it is always possible to construct a (not necessarily causal) feedback controller resulting in that subsystem.

4.4. Mergeable behaviors. We now discuss briefly a related controllability-type concept, which describes the possibility of interconnecting two systems on-line (e.g., a plant and controller) by steering their given trajectories to a common trajectory. This is a third possible paradigm for control of nD systems. While this concept has, to our knowledge, not previously appeared in the literature, similar ideas are the notions of a *compliance-free interconnection* [30] and *interconnectability with finite lag* [27, 28].

DEFINITION 4.6. *Let $\mathcal{B}$ and $\overline{\mathcal{B}}$ be two subbehaviors of $\mathcal{A}^q$. Then we say that $\mathcal{B}$ and $\overline{\mathcal{B}}$ are* mergeable *if the following condition, dependent on the signal space $\mathcal{A}$, holds.*

$\mathcal{A} = \mathcal{C}^\infty(\mathbb{R}^n, k)$ or $\mathcal{A} = \mathcal{D}'(\mathbb{R}^n, k)$. For any $w^{(1)} \in \mathcal{B}$, $w^{(2)} \in \mathcal{B}^{(2)}$, there exists $w \in \mathcal{B} \cap \overline{\mathcal{B}}$ such that for any open sets T_1, T_2 with disjoint closures, $w^{(1)}$ and w are concatenable in $\mathcal{B}$ with respect to (T_1, T_2), and $w^{(2)}$ and w are concatenable in $\overline{\mathcal{B}}$ with respect to (T_1, T_2).

$\mathcal{A} = k^T, T = \mathbb{N}^n$ or $\mathbb{Z}^n$. There exists a $\rho > 0$ such that, for any $w^{(1)} \in \mathcal{B}$, $w^{(2)} \in \mathcal{B}^{(2)}$, there exists $w \in \mathcal{B} \cap \overline{\mathcal{B}}$ such that for any $T_1, T_2 \subseteq T$ with $d(T_1, T_2) \geq \rho$, there exists $w \in \mathcal{B} \cap \overline{\mathcal{B}}$ such that $w^{(1)}$ and w are concatenable in $\mathcal{B}$ with respect to (T_1, T_2), and $w^{(2)}$ and w are concatenable in $\overline{\mathcal{B}}$ with respect to (T_1, T_2).

Thus mergeability formalizes the idea that any two trajectories in the given behaviors $\mathcal{B}$ and $\overline{\mathcal{B}}$ can be controlled to the same trajectory "in finite time." This concept is illustrated in the 1D case by Figure 4.1.

The next result characterizes mergeability and is new even in the 1D case; it also illustrates the usefulness of the set-controllability concept.

THEOREM 4.7. *Let $\mathcal{B}, \overline{\mathcal{B}} \subseteq \mathcal{A}^q$. Then $\mathcal{B}$ and $\overline{\mathcal{B}}$ are mergeable if and only if $\mathcal{B} + \overline{\mathcal{B}}$ is controllable.*

Proof. Consider the behavior $\mathcal{B} \oplus \overline{\mathcal{B}}$, which is naturally a subbehavior of $\mathcal{A}^{2q}$. Now $\mathcal{B} \cap \overline{\mathcal{B}}$ can be treated as a subbehavior of $\mathcal{B} \oplus \overline{\mathcal{B}}$ under the map $\iota : w \mapsto (w, w)$. With this embedding in mind, it is clear that $\mathcal{B}$ and $\overline{\mathcal{B}}$ are mergeable if and only if $\mathcal{B} \oplus \overline{\mathcal{B}}$ is set-controllable to $(\mathcal{B} \cap \overline{\mathcal{B}})$.

Now the image of the map ι is equal to the kernel of the projection $\pi : \mathcal{B} \oplus \overline{\mathcal{B}} \mapsto \mathcal{B} + \overline{\mathcal{B}}$, $(w^{(1)}, w^{(2)}) \mapsto w^{(1)} - w^{(2)}$, and therefore $\mathcal{B} + \overline{\mathcal{B}}$ is naturally isomorphic to the factor $(\mathcal{B} \oplus \overline{\mathcal{B}})/(\mathcal{B} \cap \overline{\mathcal{B}})$. The result now follows from condition 2 of Theorem 4.2. ∎

It can be seen that behaviors which are regularly interconnectable are mergeable, but not conversely. Also, a pair of mergeable behaviors are each set-controllable to their intersection, though this condition is not sufficient for mergeability.

5. Regular extended interconnection. In this section we will extend the framework of regular interconnection to include a concept equivalent in the nD case to set-controllability. The key is to introduce additional variables and to construct controllers in the extended variable space. We can think of these additional control variables as being variables internal to the system, which cannot be directly affected by means of a regular interconnection. By allowing restrictions to be placed directly on such internal variables, we can achieve a larger set of controlled behaviors than is possible through regular interconnection. In the 1D case, however, no additional power is gained through the ability to restrict latent or internal variables. This seems to be due to the fact that, in the 1D case, we can always choose our latent variable descriptions (see below) to be observable, so that the values of the latent variables can be determined from those of the system variables. Restrictions on such latent variables are therefore equivalent to restrictions on the manifest variables.

An idea similar to the use of latent variables is the partitioning of system variables in the plant $\mathcal{B}$ into variables w, which we wish to control, and variables v, which we are not interested in controlling but which we can nevertheless influence by interconnection. This framework has been used by Polderman and Mareels [16] and by Trentelman and Willems [23, 24] as a starting point for more advanced control theories in the behavioral context. In this paper we take the complementary approach of starting with the variables w, which we wish to control, and introducing the additional variables v, which we can also influence.

Recall that a behavior $\mathcal{B}_{w,v} \subseteq \mathcal{A}^{q+l}$ is said to be a *latent variable description* of a behavior $\mathcal{B}_w$ if

$$\mathcal{B}_w \;=\; \{w \in \mathcal{A}^q \mid \exists\, v \in \mathcal{A}^l \text{ such that } (w, v) \in \mathcal{B}_{w,v}\}. \tag{5.1}$$

The variables w are called *manifest variables*, and the variables v are called *latent variables* or auxiliary variables. Latent variable descriptions are often represented in "ARMA form" as

$$\mathcal{B}_{w,v} \;=\; \ker\,(-\hat{R}\;M) \;=\; \{(w, v) \in \mathcal{A}^{q+l} \mid \hat{R}w \;=\; Mv\} \tag{5.2}$$

for some polynomial matrices $\hat{R}$, M. These ideas are described in [8, 19, 21, 30]. The *manifest behavior* $\mathcal{B}_w$ can be constructed from the *full behavior* $\mathcal{B}_{w,v}$ as follows [8], [14, Cor. 2.38]: if E is an MLA of M, then $\mathcal{B}_w = \ker E\hat{R}$. Note our use of $\hat{R}$ rather than

the conventional R; this is to avoid the suggestion that the kernel of the left-hand operator is equal to our $\mathcal{B} = \mathcal{B}_w$. However, we do of course have that $\ker \hat{R} \subseteq \mathcal{B}_w$. Analogously to (5.1), we define the behavior $\mathcal{B}_v$.

5.1. Free latent variable descriptions. If we wish to perform a control action to drive a system trajectory into some desired subbehavior, our first problem is that we can normally affect only free variables. Control of the free input variables of the system unfortunately can only determine the controlled trajectory to within $\mathcal{B}_{0,y}$, and this may not be good enough. It therefore becomes necessary to look for free latent variables, which leads us to introduce the following concept.

DEFINITION 5.1. *A latent variable description $\mathcal{B}_{w,v} \subseteq \mathcal{A}^{q+l}$ is said to be a* free latent variable description *if the variables v are free variables, i.e., if $\mathcal{B}_v = \mathcal{A}^l$.*

We can characterize free latent variable descriptions as follows.

LEMMA 5.2. *Let $\mathcal{B}_{w,v} = \ker(-\hat{R}\ M)$ be a latent variable description of the behavior $\mathcal{B} = \mathcal{B}_w$. Then the following are equivalent.*

1. *$\mathcal{B}_{w,v}$ is a free latent variable description.*
2. *$\operatorname{rank} \hat{R} = \operatorname{rank}(-\hat{R}\ M)$.*
3. *$\operatorname{im} M \subseteq \operatorname{im} \hat{R}$.*
4. *M is an image representation of $\hat{R}\mathcal{B}$.*

Proof. $1 \Rightarrow 2$. Suppose $\mathcal{B}_{w,v}$ is a free latent variable representation. Then $m(\mathcal{B}_v) = l$, the number of variables v. Now $\mathcal{B}_v = \mathcal{B}_{w,v}/\mathcal{B}_{w,0}$, so by additivity of the number of free variables we have $m(\mathcal{B}_{w,v}) - m(\mathcal{B}_{w,0}) = l$. However, $m(\mathcal{B}_{w,v}) = (q+l) - \operatorname{rank}(-\hat{R}\ M)$, and $m(\mathcal{B}_{w,0}) = q - \operatorname{rank} \hat{R}$, from which $\operatorname{rank} \hat{R} = \operatorname{rank}(-\hat{R}\ M)$ as required.

$2 \Rightarrow 3$. Suppose that $\operatorname{rank} \hat{R} = \operatorname{rank}(-\hat{R}\ M)$. Then every column of M is linearly dependent on the columns of $\hat{R}$, i.e., there exist a nonsingular diagonal polynomial matrix D and a polynomial matrix X such that $\hat{R}X = MD$. Since D has full row rank, $\operatorname{im} D = \mathcal{A}^l$ by [14, pp. 24–25], and consequently $\operatorname{im} M = \operatorname{im} \hat{R}X \subseteq \operatorname{im} \hat{R}$ as required.

$3 \Rightarrow 4$. Suppose that $\operatorname{im} M \subseteq \operatorname{im} \hat{R}$. Then for any $x \in \operatorname{im} M$, $x = Mv = \hat{R}w$ for some w, v, which implies $w \in \mathcal{B}$, and so $x \in \hat{R}\mathcal{B}$. Conversely, if $x \in \hat{R}\mathcal{B}$, then $x = \hat{R}w$ for some $w \in \mathcal{B}$, so there must exist v with $x = \hat{R}w = Mv$, and now $x \in \operatorname{im} M$. This proves that $\operatorname{im} M = \hat{R}\mathcal{B}$.

$4 \Rightarrow 1$. Suppose that M is an image representation of $\hat{R}\mathcal{B}$. Then for any $v \in \mathcal{A}^l$, $Mv \in \hat{R}\mathcal{B}$, and so $v \in \mathcal{B}_v$. Hence $\mathcal{B}_{w,v}$ is a free latent variable representation as required. □

We now look at some interesting special cases of free latent variable descriptions.

Example 5.3. 1. Free input/output structures: A behavior $\mathcal{B} = \mathcal{B}_{y,u}$ with a given free input/output structure is clearly a free latent variable description of the manifest output behavior $\mathcal{B}_y$. Conversely, in the case where $\hat{R}w = Mv$ defines a free latent variable description and the behavior $\ker \hat{R}$ is furthermore autonomous, we have a free input/output structure on $\mathcal{B}_{w,v}$ with inputs v and outputs w. An interesting open question is whether the resulting transfer matrix and minimal element of the transfer class then have any interpretation in terms of the latent variable description.

2. A latent variable description arising from a controllable-autonomous decomposition $\mathcal{B} = \mathcal{B}^c + \mathcal{B}^a$, $\mathcal{B}^c = \operatorname{im} M$, $\mathcal{B}^a = \ker \hat{R}$ of $\mathcal{B}$: In this situation,

$$\hat{R}w \ = \ \hat{R}Mv$$

defines a free latent variable description of $\mathcal{B}$. It is a latent variable description of $\mathcal{B}$

since if $\hat{R}w = \hat{R}Mv$, then $w - Mv \in \ker \hat{R}$, so $w \in \operatorname{im} M + \ker \hat{R}$, and conversely. The latent variables v are free from the fact that $\operatorname{im} M = \mathcal{B}^c \subseteq \mathcal{B}$.

3. Consider a classical 1D state-space model, given in the form

$$\begin{pmatrix} B & 0 \\ D & -I \end{pmatrix} \begin{pmatrix} u \\ y \end{pmatrix} = \begin{pmatrix} zI - A \\ -C \end{pmatrix} x, \tag{5.3}$$

where the matrices A, B, C, and D are matrices over the field k. This is a latent variable description of the manifest behavior $\mathcal{B}_{u,y}$ (the behavior of the input and output variables only). By Lemma 5.2, it is free precisely when

$$\operatorname{rank} \begin{pmatrix} zI - A & -B & 0 \\ -C & -D & I \end{pmatrix} = \operatorname{rank} \begin{pmatrix} -B & 0 \\ -D & I \end{pmatrix}. \tag{5.4}$$

However, the rank of the left-hand side is equal to the number of output variables plus the number of state variables, whereas the rank of the right-hand side is equal to the number of output variables plus the rank of B. Therefore, (5.3) is a free latent variable description of $\mathcal{B}_{u,y}$ if and only if B has full row rank.

5.2. Extended interconnection. Having introduced an appropriate class of latent variable descriptions, we now discuss interconnections on the extended variable spaces given by latent variable descriptions.

DEFINITION 5.4. *Suppose that $\mathcal{B}' \subseteq \mathcal{B} \subseteq \mathcal{A}^q$. If there exists a latent variable description $\mathcal{B}_{w,v} \subseteq \mathcal{A}^{q+l}$ of $\mathcal{B}$ such that*

$$\mathcal{B}'_{w,v} = \mathcal{B}_{w,v} \cap \overline{\mathcal{B}_{w,v}} \tag{5.5}$$

for some $\overline{\mathcal{B}_{w,v}}, \mathcal{B}'_{w,v} \subseteq \mathcal{A}^{q+l}$ with $\mathcal{B}'$ equal to the manifest behavior $\mathcal{B}'_w$ of $\mathcal{B}'_{w,v}$, then we say that $\mathcal{B}'$ is achievable from $\mathcal{B}$ by extended interconnection*, and we call* (5.5) *an* extended interconnection*. If the variables w are free in $\overline{\mathcal{B}_{w,v}}$, then* (5.5) *is called a* latent interconnection*, and $\overline{\mathcal{B}_{w,v}}$ is called a* latent controller*. If* (5.5) *is a regular interconnection, i.e., if*

$$p(\mathcal{B}'_{w,v}) = p(\mathcal{B}_{w,v}) + p(\overline{\mathcal{B}_{w,v}}),$$

then it is called a regular extended interconnection *or a* regular latent interconnection*, appropriately. Hence we define* achievability by regular extended/latent interconnection*.*

Achievability by regular extended interconnection therefore describes the possibility of achieving the desired subsystem by extending the variable space using some latent variable description, applying a regular interconnection, and then projecting onto the manifest variables. Latent interconnections are extended interconnections in which the only restrictions applied are on the latent variables. To make a clear distinction, we will refer to an "ordinary" interconnection (i.e., one which restricts only the manifest variables w) as a *manifest interconnection*, and, similarly, we will refer to *manifest controllers*. Latent and manifest controllers are investigated by Kuijper in [10].

Given a latent variable description represented by the formula $\hat{R}w = Mv$, a general extended interconnection is described by the polynomial matrix

$$\left(\begin{array}{cc} -\hat{R} & M \\ \hline N_1 & N_2 \end{array}\right).$$

The case $N_2 = 0$ describes a manifest interconnection in extended interconnection form, and the case $N_1 = 0$ is an important special case of a latent interconnection.

Any manifest interconnection can therefore be expressed trivially as an extended interconnection. Also, any latent interconnection is equivalent to some manifest interconnection (in the sense that there exists a manifest interconnection resulting in the same controlled manifest behavior). To see this, note that any latent controller $N_2 v = 0$ is equivalent in this sense to a latent controller of the form $(LM)v = 0$, which is in turn equivalent to the manifest controller $(L\hat{R})w = 0$. However, we can obtain any subbehavior of the full behavior $\mathcal{B}_w$ by some manifest interconnection, whereas only those subbehaviors containing $\ker \hat{R}$ are achievable by latent interconnection (see [16, 23, 24] for discussions along these lines). For a *fixed* latent variable description, more behaviors can generally be achieved by manifest interconnection than by latent interconnection.

However, any manifest interconnection can be expressed as a latent interconnection on *some* suitably chosen latent variable description, e.g., the behavior $\mathcal{B} = \ker R$ can be extended to the latent variable description given by

$$\begin{pmatrix} I \\ 0 \end{pmatrix} w \;=\; \begin{pmatrix} I \\ R \end{pmatrix} v,$$

and the manifest controller $\overline{R}w = 0$ is equivalent to the latent controller $\overline{R}v = 0$. Furthermore, any regular interconnection becomes a regular latent interconnection in this way. In particular, any subbehavior achievable from $\mathcal{B}$ by regular interconnection is achievable from $\mathcal{B}$ by regular latent interconnection. We will shortly see that the converse of this is not true.

It is important to make a distinction between the behavior $\mathcal{B}'_{w,v}$ obtained by extended interconnection and the behavior

$$\mathcal{B}^*_{w,v} \;:=\; \{(w,v) \in \mathcal{B}_{w,v} \mid w \in \mathcal{B}'_w\}$$

obtained from the manifest $\mathcal{B}' = \mathcal{B}'_w$. It can be easily shown that the latter is equal to $\mathcal{B}'_{w,v} + \mathcal{B}_{0,v}$. Another trap is to suppose that the manifest controlled behavior $\mathcal{B}'_w$ is given by the intersection of the original manifest behavior $\mathcal{B}_w$ and the manifest behavior $\overline{\mathcal{B}_w}$ of the controller. This is generally not the case, e.g., for a latent interconnection we have $\overline{\mathcal{B}_w} = \mathcal{A}^q$, and certainly $\mathcal{B}_w \cap \mathcal{A}^q \neq \mathcal{B}'_w$, except when the interconnection is trivial.

Example 5.5. 1. Consider the behaviors $\mathcal{B}, \mathcal{B}'$ from Example 3.6.3. We saw there that $\mathcal{B}'$ is not achievable from $\mathcal{B}$ by regular interconnection. Now a free latent variable description of $\mathcal{B}$ is given by

$$\begin{pmatrix} z_1 + 1 & 0 \\ 0 & z_2 \\ 0 & -z_1 \end{pmatrix} w \;=\; \begin{pmatrix} z_2 - z_1 \\ z_2(z_1 - 1) \\ -z_1(z_1 - 1) \end{pmatrix} v.$$

It is easy to verify that the compound matrix $(-\hat{R}\ M)$ in this case has rank 2, so that condition 2 of Lemma 5.2 holds. Now consider the controller (which restricts both manifest and latent variables) given by

$$\begin{pmatrix} 0 & z_1(z_2 - 1) \\ 0 & z_1(z_1 - 1) \end{pmatrix} w \;=\; \begin{pmatrix} z_1^2(z_2 - 1) \\ z_1^2(z_1 - 1) \end{pmatrix} v.$$

We can see that this is a regular extended interconnection. The resulting manifest controlled behavior is computed by elimination of latent variables as the subbehavior $\mathcal{B}'$ specified earlier.

Free latent variable descriptions are particularly simple as regards latent interconnection:

LEMMA 5.6. *Any latent interconnection on a free latent variable description is regular.*

Proof. Suppose that $\mathcal{B}_{w,v} \subseteq \mathcal{A}^{q+l}$ is a free latent variable description of $\mathcal{B}_w$, and that $\mathcal{B}'_{w,v} = \mathcal{B}_{w,v} \cap \overline{\mathcal{B}_{w,v}}$ is a latent interconnection. Then the variables v are free in $\mathcal{B}_{w,v}$, and the variables w are free in $\overline{\mathcal{B}_{w,v}}$. Hence $\mathcal{B}_{w,v} + \overline{\mathcal{B}_{w,v}} = \mathcal{A}^{q+l}$, so by Lemma 3.3 the latent interconnection is regular. ∎

5.3. Set-controllability and extended interconnection. Our final main result shows that set-controllability is precisely equivalent to regular interconnection using latent variables.

THEOREM 5.7. *Let $\mathcal{B}' \subseteq \mathcal{B}$. The following are equivalent.*

1. *$\mathcal{B}$ is set-controllable to $\mathcal{B}'$.*
2. *$\mathcal{B}$ admits a free latent variable description $R'w = Mv$ with $\ker R' = \mathcal{B}'$.*
3. *$\mathcal{B}'$ is achievable from $\mathcal{B}$ by latent interconnection on a free latent variable description.*
4. *$\mathcal{B}'$ is achievable from $\mathcal{B}$ by regular latent interconnection on a free latent variable description.*
5. *$\mathcal{B}'$ is achievable from $\mathcal{B}$ by regular latent interconnection.*
6. *$\mathcal{B}'$ is achievable from $\mathcal{B}$ by regular extended interconnection.*

Proof. Suppose that $\mathcal{B}$ is set-controllable to $\mathcal{B}'$, and let R' be an image representation of $\mathcal{B}'$. Then by Theorem 4.2 $R'\mathcal{B} \cong \mathcal{B}/\mathcal{B}'$ is controllable and so admits an image representation M. Now the manifest behavior of $R'w = Mv$ is equal to $\mathcal{B}$, since $R'w = Mv \;\Rightarrow\; R'w \in R'\mathcal{B} \;\Rightarrow\; w \in \mathcal{B} + \mathcal{B}' = \mathcal{B}$, whereas $w \in \mathcal{B}$ implies $R'w \in \operatorname{im} M$. Hence $R'w = Mv$ is a latent variable description of $\mathcal{B}$ and is free by Lemma 5.2. This establishes condition 2. Given that condition, we achieve $\mathcal{B}'$ through the trivial latent interconnection $Iv = 0$. By Lemma 5.6, this is a regular latent (extended) interconnection. This establishes $1 \Rightarrow 2 \Rightarrow 3 \Rightarrow 4 \Rightarrow 5 \Rightarrow 6$. Finally, suppose that $\mathcal{B}'$ is achievable from $\mathcal{B}$ by regular extended interconnection, and let $\mathcal{B}'_{w,v}$ and $\mathcal{B}_{w,v}$ be the corresponding extended behaviors. By Theorem 4.5, $\mathcal{B}_{w,v}$ is set-controllable to $\mathcal{B}'_{w,v}$, and it follows from the definition of set-controllability that $\mathcal{B}_w$ is set-controllable to $\mathcal{B}'_w$, i.e., $\mathcal{B}$ is set-controllable to $\mathcal{B}'$. This completes the proof. ∎

The proof of Theorem 5.7 shows that $\mathcal{B}' = \ker R'$ can be achieved by regular extended or latent interconnection on the free latent variable description $R'w = Mv$ for some M. In fact, the choice of free latent variable descriptions allowing control to $\mathcal{B}'$ is much wider.

LEMMA 5.8. *If $\mathcal{B}$ is set-controllable to $\mathcal{B}'$, then $\mathcal{B}'$ is achievable from $\mathcal{B}$ by regular latent interconnection on any free latent variable description of $\mathcal{B}$ given by an equation of the form*

$$\hat{R}w \;=\; Mv,$$

where $\ker \hat{R} \subseteq \mathcal{B}'$.

Proof. Let $\hat{R}$ be such that $\hat{R}w = Mv$ is a free latent variable description of $\mathcal{B}$, and also $\ker \hat{R} \subseteq \mathcal{B}'$. There must exist an L with $\ker L\hat{R} = \mathcal{B}'$. Now imposition of the controller $LMv = 0$ on the extended behavior is a latent interconnection and regular

by Lemma 5.6. The manifest behavior of the controlled system is then

$$\begin{aligned}\{w \in \mathcal{A}^q \mid \exists\ v \text{ with } \hat{R}w = Mv,\ LMv = 0\} \\ = \{w \in \mathcal{A}^q \mid \exists\ v \text{ with } \hat{R}w = Mv,\ L\hat{R}w = 0\} \\ = \mathcal{B}'.\end{aligned}$$

Hence $\mathcal{B}'$ is achievable by regular latent interconnection on $\hat{R}w = Mv$. □

Example 5.9. 1. Returning to the 1D state-space model in item 3 in Example 5.3, we see that if the matrix B has full row rank, then the full behavior $\mathcal{B}_{x,u,y}$ is set-controllable to $\mathcal{B}_{0,u,y}$, i.e., to the subbehavior of all trajectories for which the state variables remain 0.

An interesting extension of the results in this section would be the consideration of which manifest behaviors can be achieved by restricting only the free latent variables in a (not necessarily free) latent variable description. We expect that this again gives only the subbehaviors to which $\mathcal{B}_w$ is set-controllable.

5.4. On-line control. We have seen that a behavior $\mathcal{B}$ is set-controllable to a given subbehavior $\mathcal{B}'$ if and only if $\mathcal{B}$ has a latent variable description of the form

$$R'w = Mv, \qquad \ker R' = \mathcal{B}', \tag{5.6}$$

in which the latent variables v are free. The question then arises: how can we actually perform the control, i.e., if we are given a trajectory $w \in \mathcal{B}$ in the region $T_1 \subseteq T$, how can we drive w into the desired subbehavior $\mathcal{B}'$ in the region $T_2 \subseteq T$? The obvious answer is by setting the variables v to 0 in T_2.

Unfortunately, the obvious answer is wrong. Setting v to 0 on T_2 enforces only $R'w = 0$ on T_2, which remarkably is not enough to establish that w agrees on T_2 with a trajectory of $\mathcal{B}'$. For example, with $\mathcal{A} = \mathbb{R}^{\mathbb{Z}^n}$, $n = 2$, we might have the behavior $\mathcal{B}' = \ker(z_1 - 1)$ consisting of functions which are constant in the t_1-direction. Take $T_2 = \mathbb{Z}^2 \backslash \{(0,0)\}$, and take w^* to be a trajectory given by

$$w^*(t_1, t_2) = \begin{cases} 1 & \text{if } t_1 \geq 1 \text{ and } t_2 = 0, \\ 0 & \text{elsewhere.} \end{cases}$$

Then $(z_1 - 1)w^*$ clearly vanishes on T_2, although w^* does not agree on T_2 with any trajectory of $\mathcal{B}'$. We can easily manufacture continuous and even 1D examples of this phenomenon.

We can get around this problem by making a change of latent variables. Suppose we are considering a free latent variable description of $\mathcal{B}$ of the form

$$\hat{R}w = Mv, \tag{5.7}$$

where, of course, $\ker \hat{R}$ is contained in the desired subbehavior $\mathcal{B}'$. By condition 2 of Lemma 5.2, the columns of M must be linearly dependent (over the ring $\mathcal{D}$) on the columns of $\hat{R}$. Hence there exists a diagonal nonsingular polynomial matrix D and a polynomial matrix L satisfying $\hat{R}L = MD$. Since D has full row rank, it follows that $\operatorname{im} MD = \operatorname{im} M$, and so

$$\hat{R}w = \hat{R}Lv \tag{5.8}$$

is another free latent variable description of $\mathcal{B}$. For such a latent variable description, a given trajectory w can be steered into the subbehavior $\mathcal{B}'$ by control of the variables

v. Since if we can drive a trajectory into $\ker \hat{R}$, we can certainly drive it into any larger behavior $\mathcal{B}'$, it suffices to prove this in the case $\hat{R} = R'$, $\ker R' = \mathcal{B}'$. In this case, the control action is the setting of the latent variables v to 0 in a suitable region. The next result demonstrates this; for brevity it is stated and proved in the continuous case only.

LEMMA 5.10. *Let $\mathcal{A} = \mathcal{C}^\infty(\mathbb{R}^n, k)$ or $\mathcal{A} = \mathcal{D}'(\mathbb{R}^n, k)$. Suppose that $\mathcal{B}$ has a free latent variable description of the form*

$$R'w \;=\; R'Lv,$$

where L has l columns and $\mathcal{B}' = \ker R'$. Then $\mathcal{B}$ is set-controllable to $\mathcal{B}'$, and, furthermore, any given trajectory of $\mathcal{B}$ can be driven into $\mathcal{B}'$ by control of the variables v.

More formally, let $w \in \mathcal{B}$, and suppose $v \in \mathcal{A}^l$ is such that $R'w = R'Lv$. Then for any open sets $T_1, T_2 \subseteq \mathbb{R}^n$ with disjoint closures, there exists a trajectory $v^* \in \mathcal{A}^l$ which agrees with v on T_1, such that for any $w^* \in \mathcal{B}$ satisfying the conditions

1. *$R'w^* = R'Lv^*$, and*
2. *w^* agrees with w on T_1,*

w^* agrees with a trajectory of $\mathcal{B}'$ on T_2.

Proof. The fact that $\mathcal{B}$ is set-controllable to $\mathcal{B}'$ is immediate from Theorem 5.7. Now let $w \in \mathcal{B}$ and $v \in \mathcal{A}^l$ satisfy $R'w = R'Lv$. Let $v^* \in \mathcal{A}^l$ be any trajectory which agrees with v on T_1 and vanishes on T_2; v^* exists because $\mathcal{A}^l$ is controllable. Now let w^* be a trajectory agreeing with w on T_1 and satisfying $R'w^* = R'Lv^*$; note that $w^* \in \mathcal{B}$. For any $t_2 \in T_2$ we now have

$$w^*(t_2) \;=\; (w^* - Lv^*)(t_2) + (Lv^*)(t_2) \;=\; (w^* - Lv^*)(t_2).$$

Since $w^* - Lv^* \in \ker R' = \mathcal{B}'$, we have that w^* agrees with a trajectory of $\mathcal{B}'$ on T_2. ∎

The interpretation of the various trajectories in the statement of Lemma 5.10 is as follows. The initial system trajectory is w, and v is some corresponding latent variable trajectory. Now w is controlled by setting v appropriately outside the region T_1. (This is formalized by having v^* agree with v on T_1, so that v^* is the latent trajectory "that actually occurs".) The trajectory w^* is required to fit the given "initial data" w on T_1 and is assumed to result from the given controlled latent variable trajectory v^*, so it must also satisfy $R'w^* = R'Lv^*$. The lemma now guarantees that any such w^* will agree with a trajectory of the subbehavior $\mathcal{B}'$ on the region T_2. Therefore, the control action of choosing v^* appropriately is guaranteed to steer w^* into $\mathcal{B}'$ on T_2.

It is worth noting that when adjusting Lemma 5.10 to the discrete case, it is necessary that v^* should agree with v on a suitably large extension of T_1. Note also that in the proof of Lemma 5.10, the resulting trajectory $w' = w^* - Lv^* \in \mathcal{B}'$ depends upon the choice of T_1, unlike in the definition of set-controllability.

6. Conclusions. We have looked at "control" from two points of view: as regular interconnection and as the ability to drive trajectories. We have seen that for 1D systems theory these paradigms are essentially different perspectives on the same phenomenon, whereas in the nD case they differ. Regular interconnection of nD systems appears to be a very strong condition. However, we have shown that the subsystems to which a given system is set-controllable are precisely those which are obtainable by regular interconnection in an extended trajectory space. In order to control nD trajectories by interconnection, it seems necessary to introduce these latent variables.

One important issue to be addressed in the future is that of causality. In the nD framework, this is most naturally defined with respect to a cone [15, 25]; a cohesive theory of 2D causality in the behavioral framework is emerging in the work of Napoli and Zampieri [13, 37]. Combining the causality theory with the theory of regular interconnections will surely be a difficult but important step in the future development of nD control theory.

Acknowledgment. We would like to thank two anonymous reviewers for their very helpful comments and suggestions.

REFERENCES

[1] M. Bisiacco, E. Fornasini, and G. Marchesini, *Dynamic regulation of 2D systems: A state-space approach*, Linear Algebra Appl., 122/123/124 (1989), pp. 195–218.
[2] M. Bisiacco and M. Valcher, *A note on the direct sum decomposition of two-dimensional behaviors*, Trans. Circuits Systems V Fund. Theory Appl., (2001), to appear.
[3] G. Conte and A. Perdon, *Composition of I/O behavioral systems*, in Systems and Networks: Mathematical Theory and Applications (MTNS 93), V. Helmke, R. Mennicken, and J. Saurer, eds., Akademie Verlag, Berlin, 1994, pp. 103–108.
[4] M. Fliess, *Controllability revisited*, in Mathematical System Theory (The Influence of R. E. Kalman), A. Antoulas, ed., Springer-Verlag, Berlin, 1991, pp. 463–474.
[5] M. Fliess and H. Bourlès, *Discussing some examples of linear system interconnections*, Systems Control Lett., 27 (1996), pp. 1–7.
[6] M. Fliess, J. Lévine, P. Martin, and P. Rouchon, *Flatness and defect of nonlinear systems: Introductory theory and examples*, Internat. J. Control, 61 (1995), pp. 1327–1361.
[7] E. Fornasini, P. Rocha, and S. Zampieri, *State space realization of 2D finite-dimensional behaviors*, SIAM J. Control Optim., 31 (1993), pp. 1502–1517.
[8] J. Komorník, P. Rocha, and J. Willems, *Closed subspaces, polynomial operators in the shift, and ARMA representations*, Appl. Math. Lett., 4 (1991), pp. 15–19.
[9] V. Kucera, *Discrete Linear Control—The Polynomial Matrix Equation Approach*, John Wiley, New York, 1979.
[10] M. Kuijper, *Why do stabilizing controllers stabilize?*, Automatica J. IFAC, 31 (1995), pp. 621–625.
[11] V. Lomadze and E. Zerz, *Control and interconnection revisited: The linear multidimensional case*, in Proceedings of the Second International Workshop on Multidimensional Systems, Technical University Press, Zielona Góra, Poland, 2000, pp. 77–82.
[12] H. Mounier, *Propriétés Structurelles des Systèmes Linéaires A Retards: Aspects Théoriques et Pratiques*, Ph.D. thesis, Université Paris-Sud, Paris, France, 1995.
[13] M. Napoli and S. Zampieri, *2D proper rational matrices and causal input/output representations of 2D behavioral systems*, SIAM J. Control Optim., 37 (1999), pp. 1538–1552.
[14] U. Oberst, *Multidimensional constant linear systems*, Acta Appl. Math., 20 (1990), pp. 1–175.
[15] H. Pillai and S. Shankar, *A behavioral approach to control of distributed systems*, SIAM J. Control Optim., 37 (1998), pp. 388–408.
[16] J. Polderman and I. Mareels, *A behavioral approach to adaptive control*, in The Mathematics of Systems and Control: From Intelligent Control to Behavioral Systems, J. Polderman and H. Trentelman, eds., Foundation Systems and Control, Groningen, The Netherlands, 1999, pp. 119–130.
[17] J.-F. Pommaret, *Partial Differential Equations and Group Theory: New Perspectives for Applications*, Math. Appl. 293, Kluwer, Dordrecht, The Netherlands, 1994.
[18] J.-F. Pommaret and A. Quadrat, *Generalized Bezout identity*, Appl. Algebra Engrg. Comm. Comput., 9 (1998), pp. 91–116.
[19] P. Rocha, *Structure and Representation of 2D Systems*, Ph.D. thesis, University of Groningen, Groningen, The Netherlands, 1990.
[20] P. Rocha, *Feedback control of multidimensional systems*, in Proceedings of the 14th International Symposium on the Mathematical Theory of Networks and Systems, Perpignan, France, 2000.
[21] P. Rocha and J. Willems, *Controllability of 2D systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 413–423.

[22] P. Rocha and J. Wood, *A foundation for the control theory of nD behaviors*, in Proceedings of the 13th International Symposium on the Mathematical Theory of Networks and Systems, A. Beghi, L. Finesso, and G. Picci, eds., Il Poligrafo, Padova, Italy, 1998, pp. 377–380.

[23] H. Trentelman, *A truly behavioral approach to the H_∞ control problem*, in The Mathematics of Systems and Control: From Intelligent Control to Behavioral Systems, J. Polderman and H. Trentelman, eds., Foundation Systems and Control, Groningen, The Netherlands, 1999, pp. 177–190.

[24] H. Trentelman and J. Willems, *H-infinity control in a behavioral context: The full information case*, IEEE Trans. Automat. Control, 44 (1999), pp. 521–536.

[25] M. Valcher, *Characteristic cones and stability properties of two-dimensional autonomous behaviors*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 47 (2000), pp. 290–302.

[26] M. Valcher, *On the decomposition of two-dimensional behaviors*, Multdimens. Systems Signal Process., 11 (2000), pp. 49–65.

[27] S. Weiland and A. Stoorvogel, *Rational representations of behaviors: Interconnectability and stabilizability*, Math. Control Signals Systems, 10 (1997), pp. 125–164.

[28] S. Weiland and A. Stoorvogel, *Stability and stabilizability of behaviors*, in Proceedings of the European Control Conference, Brussels, Belgium, 1997.

[29] J. Willems, *From time series to linear system—part* I: *Finite-dimensional linear time invariant systems*, Automatica J. IFAC, 22 (1986), pp. 561–580.

[30] J. Willems, *Paradigms and puzzles in the theory of dynamical systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 259–294.

[31] J. Willems, *Feedback in a behavioral setting*, in Systems, Models and Feedback: Theory and Applications, A. Isidori and T. Tam, eds., Birkhäuser, Boston, 1992, pp. 179–191.

[32] J. Willems, *On interconnections, control, and feedback*, IEEE Trans. Automat. Control, 42 (1997), pp. 458–472.

[33] J. Wood, *Modules and behaviors in nD systems theory*, Multidimens. Systems Signal Process., 11 (2000), pp. 11–48.

[34] J. Wood, U. Oberst, E. Rogers, and D. H. Owens, *A behavioral approach to the pole structure of one-dimensional and multidimensional linear systems*, SIAM J. Control Optim., 38 (2000), pp. 627–661.

[35] J. Wood, E. Rogers, and D. Owens, *Controllable and autonomous nD linear systems*, Multdimens. Systems Signal Process., 10 (1999), pp. 33–69.

[36] J. Wood and E. Zerz, *Notes on the definition of behavioral controllability*, Systems Control Lett., 37 (1999), pp. 31–37.

[37] S. Zampieri, *Causal input/output representation of 2D systems in the behavioral approach*, SIAM J. Control Optim., 36 (1998), pp. 1133–1146.

[38] E. Zerz, *Primeness of multivariate polynomial matrices*, Systems Control Lett., 29 (1996), pp. 139–146.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 1, pp. 135–148

OPTIMAL SEQUENTIAL VECTOR QUANTIZATION OF MARKOV SOURCES*

VIVEK S. BORKAR†, SANJOY K. MITTER‡, AND SEKHAR TATIKONDA§

Abstract. The problem of sequential vector quantization of a stationary Markov source is cast as an equivalent stochastic control problem with partial observations. This problem is analyzed using the techniques of dynamic programming, leading to a characterization of optimal encoding schemes.

Key words. optimal vector quantization, sequential source coding, Markov sources, control under partial observations, dynamic programming

AMS subject classifications. 94A29, 90E20, 90C39

PII. S0363012999365261

1. Introduction.

In this paper, we consider the problem of optimal sequential vector quantization of stationary Markov sources. In the traditional rate distortion framework, the well-known result of Shannon shows that one can achieve entropy rates arbitrarily close to the rate distortion function for suitably long lossy block codes [9]. Unfortunately, long block codes imply long delays in communication systems. In particular, control applications require causal coding and decoding schemes.

These concerns are not new, and there is a sizable body of literature addressing these issues. We shall briefly mention a few key contributions. Witsenhausen [24] looked at the optimal finite horizon sequential quantization problem for finite state encoders and decoders. His encoder had a fixed number of levels. He showed that the optimal encoder for a kth order Markov source depends on at most the last k symbols and the present state of the decoder's memory. Walrand and Varaiya [23] looked at the infinite horizon sequential quantization problem for sources with finite alphabets. Using Markov decision theory, they were able to show that the optimal encoder for a Markov source depends only on the current input and the current state of the decoder. Gaarder and Slepian [12] look at sequential quantization over classes of finite state encoders and decoders. Though they lay down several useful definitions, their results, by their own admission, are incomplete. Other related works include a neural network based scheme [17] and a study of optimality properties of codes in specific cases [3], [10]. Some abstract theoretical results are given in [19].

*Received by the editors December 20, 1999; accepted for publication (in revised form) December 6, 2000; published electronically May 31, 2001. A preliminary version of this paper appeared in the *Proceedings of the* 1998 *IEEE International Symposium on Information Theory*, IEEE Information Theory Society, Piscataway, NJ, 1998, p. 71.

http://www.siam.org/journals/sicon/40-1/36526.html

†School of Technology and Computer Science, Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400005, India (borkar@tifr.res.in). This work was done while this author was visiting the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology. The research of this author was supported by a Homi Bhabha fellowship, NSF KDI: Learning, Adaptation, and Layered Intelligent Systems grant 6756500, and grant III 5(12)/96-ET of Department of Science and Technology, Government of India.

‡Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Room 35-403, Cambridge, MA 02139 (mitter@lids.mit.edu). The research of this author was supported by NSF KDI: Learning, Adaptation, and Layered Intelligent Systems grant ECS-9873451.

§University of California at Berkeley, Soda Hall, Room 485, Berkeley, CA 94720 (tatikond@eecs.berkeley.edu). The research of this author was supported by U.S. Army grant PAAL03-92-G-0115.

A formulation similar in spirit to ours (insofar as it aims to minimize a "Lagrangian distortion measure" described below) is studied in [7], [8]. They show empirically that one can make gains in performance by entropy coding the codewords. In [7] the entropy constrained vector quantization problem for a block is formulated and a Max–Lloyd-type algorithm is introduced. In [8] they introduce the conditional entropy constrained vector quantization problem and show that one should use conditional entropy coders when the codewords are not independent from block to block. In these papers there is more emphasis on synthesizing algorithms and less emphasis on proving rigorously the optimality of the schemes proposed. Along with this work there is a large literature on differential predictive coding, where one encodes the innovation. Other than the Gauss–Markov case, though, it is not apparent how one may prove the optimality of such innovation coding schemes. Herein we emphasize, through the dynamic programming formulation, the optimality properties of the sequential quantization scheme. This leads the way for the application of many powerful approximate dynamic programming tools.

In this paper we do not impose a fixed number of levels on the quantizer. The aim is to somehow jointly optimize the entropy rate of the quantized process (in order to obtain a better compression rate) as well as a suitable distortion measure. The traditional rate distortion framework [9] calls for the minimization of the former with a hard constraint on the latter. We shall, however, consider the analytically more tractable Lagrangian distortion measure of [7], [8], which is a weighted combination of the two. We approach the problem from a stochastic control viewpoint, treating the choice of the sequential quantizer as a control choice. The correct "state space" then turns out to be the space of conditional laws of the underlying process given the quantizer outputs, these conditional laws serving as the "state" or "sufficient statistics." The "state dynamics" is then given by the appropriate nonlinear filter. While this is very reminiscent of the finite state quantizers studied, e.g., in [16], the state space here is not finite, and the state process has the familiar stochastic control interpretation as the output of a nonlinear filter. We then consider the "separated" or "certainty equivalent" control problem of controlling this nonlinear filter so as to minimize an appropriately transformed Lagrangian distortion measure. This problem can be analyzed in the traditional dynamic programming framework. This in turn can be made a basis for computational schemes for near-optimal code design.

To summarize, the main contributions of this paper are as follows.

(i) We formulate a stochastic control problem equivalent to the optimal vector quantization problem. In the process, we make precise the passage from the source output to its encoded version in a manner that ensures the well-posedness of the control problem.

(ii) We underscore the crucial role of the process of conditional laws of the source given the quantized process as the correct "sufficient statistics" for the problem.

(iii) We analyze the equivalent control problem by using the methodology of Markov decision theory. This opens up the possibility of using the computational machinery of Markov decision theory for code design.

Specifically, we consider a pair of a "state process" $\{X_n\}$ and an associated "observation process" $\{Y_n\}$, given by the dynamics

$$X_{n+1} = g(X_n, \xi_n),\ Y_{n+1} = h(X_n, \xi'_n),$$

where $\{\xi_n\}, \{\xi'_n\}$ are independently and identically distributed (i.i.d.) driving noise processes. We quantize Y_{n+1} into its quantized version q_{n+1} that has a finite range and

is selected based on the "history" $q^n \triangleq [q_0, q_1, \ldots, q_n]$. The aim then is to minimize the long run average of the Lagrangian distortion measure $R_n = E[H(q_{n+1}/q^n) + \lambda||Y_n - \bar{q}_n||^2]$, where $\lambda > 0$ is a prescribed constant, $H(\cdot/\cdot)$ is the conditional entropy, and $\bar{q}_n$ is the best estimate of Y_n given q_n.

Let π_n be the regular conditional law of X_n given q^n for $n \geq 0$. From π_n, one can easily derive the regular conditional law of Y_{n+1} given q^n. Using Bayes's rule, $\{\pi_n\}$ can be evaluated recursively by a nonlinear filter. Furthermore, one can express R_n as the expected value of a function of π_n and a "control" process Q_n alone. ($\{Q_n\}$ is, in fact, the finite set depicting the range of the vector quantization of Y_{n+1} prior to its encoding into a fixed finite alphabet.) This allows us to consider the equivalent problem of controlling $\{\pi_n\}$ with the aim of minimizing the long run average of the R_n recast as above. This then fits the framework of traditional Markov decision theory and can be approached by dynamic programming. As usual, one has to derive the dynamic programming equations for the average cost control problem by a "vanishing discount" argument applied to the associated infinite horizon discounted control problem for which the dynamic programming equation is easier to justify.

The structure of the paper is as follows. In section 2, we describe the sequential quantization problem and introduce the formalism. Section 3 derives the equivalent control problem. This is analyzed in section 4 using the formalism of Markov decision theory.

2. Sequential quantization. This section formulates the sequential vector quantization problem. In particular, it describes the passage from the observation process to its quantized version, which in turn gets mapped into its encoding with respect to a fixed alphabet. We also lay down our key assumptions which, apart from making the coding scheme robust, also make its subsequent control formulation well-posed. The section concludes with a precise statement of this "long run average cost" control problem with partial observations that is equivalent to our original vector quantization problem.

Throughout, for a Polish (i.e., complete separable metric) space $X, P(X)$ will denote the Polish space of probability measures on X with Prohorov topology [6, Chapter 2]. For a random process $\{Z_m\}$, set $Z^n = \{Z_m, 0 \leq m \leq n\}$, its past up to time n. Finally, K will denote a finite positive constant, depending on the context.

Let $\{X_n\}$ be an ergodic Markov process taking values in $R^s, s \geq 1$, with an associated "observation process" $\{Y_n\}$ taking values in $R^d, d \geq 1$. ($\{Y_n\}$ thus is the actual process being observed.) Their joint evolution is governed by a transition kernel $x \in R^s \to p(x, dz, dy) \in P(R^s \times R^d)$, as described below. We assume this map to be continuous and further, that $p(x, dz, dy) = \varphi(y, z|x)dzdy$ for a density $\varphi(\cdot, \cdot|\cdot) : R^d \times R^s \times R^s \to R^+$ that is continuous and strictly positive, and furthermore, $\varphi(y, z|\cdot)$ is Lipschitz uniformly in y, z.

The evolution law is as follows. For $A \subset R^s, B \subset R^d$ Borel,

$$P(X_{n+1} \in A, Y_{n+1} \in B/X^n, Y^n) = \int_{A \times B} p(X_n, dx, dy)$$
$$= \int_A \int_B \varphi(y, z|X_n)dydz.$$

Following [13], we call the pair $(\{X_n\}, \{Y_n\})$ a Markov source, though the terminology "hidden Markov model" is more common nowadays. We impose on $(\{X_n\}, \{Y_n\})$ the condition of "asymptotic flatness" described next. We assume that these processes

are given recursively by the dynamics

$$X_{n+1} = g(X_n, \xi_n), \tag{2.1}$$

$$Y_{n+1} = h(X_n, \xi'_n), \tag{2.2}$$

where $\{\xi_n\}, \{\xi'_n\}$ are i.i.d. R^m-valued (say) random variables independent of each other and of X_0, and $g : R^s \times R^m \to R^s$, $h : R^s \times R^m \to R^d$ are prescribed measurable maps satisfying

$$||g(x,y)||, ||h(x,y)|| \leq K(1 + ||x||) \ \ \forall\ y.$$

Equations (2.1) and (2.2) and the laws of $\{\xi_n\}, \{\xi'_n\}$ completely specify $p(x, dz, dy)$, and therefore the conditions we impose on the latter will implicitly restrict the choice of the former.

Let $(\{X_n(x)\}, \{Y_n(x)\}), (\{X_n(y)\}, \{Y_n(y)\})$ denote the solutions to (2.1), (2.2) for $X_0 = x$, respectively, y with the *same* driving noises $\{\xi_n\}, \{\xi'_n\}$. The assumption of asymptotic flatness then is that there exist $K > 0, 0 < \beta < 1$, such that

$$E[||X_n(x) - X_n(y)||] \leq K\beta^n ||x - y||, n \geq 0.$$

A simple example would be the case when $g(x,u) = \bar{g}(x) + u, h(x,u) = \bar{h}(x) + u$ for all x, u, where $\bar{g} : R^s \to R^s$ is a contraction with respect to some equivalent norm on R^s. This covers, e.g., the usual linear quadratic Gaussian (LQG) case when the state process is stable. Another example would be a discretization of continuous time asymptotically flat processes considered in [1], where a Lyapunov-type sufficient condition for asymptotic flatness is given. This assumption, one must add, is not required for our formulation of the optimization problem per se but will play a key role in our derivation of the dynamic programming equations in section 4.

Let $\sum = \{\alpha_1, \alpha_2, \ldots, \alpha_N\}$ be an ordered set that will serve as the alphabet for our vector quantizer. Let $\{q_n\}$ denote the $\sum$-valued process that stands for the "vector quantized" version of $\{Y_n\}$. The passage from $\{Y_n\}$ to $\{q_n\}$ is described below.

Let D denote the set of finite nonempty subsets of R^d with cardinality at most $N \geq 1$, satisfying the following.

(†) There exist $M > 0$ ("large") and $\triangle > 0$ ("small") such that

(i) $x \in A \in D$ implies $||x|| \leq M$,

(ii) $x = [x_1, \ldots, x_d], y = [y_1, \ldots, y_d]$ for $x, y \in A \in D$, $x \neq y$, implies $|x_i - y_i| > \triangle$ for all i.

We endow D with the Hausdorff metric which renders it a compact Polish space. For $A \in D$, let $l_A : R^d \to A$ denote the map that maps $x \in R^d$ to the element of A nearest to it with reference to the Euclidean norm $||\cdot||$, any tie being resolved according to some fixed priority rule. Let $i_A : A \to \sum$ denote the map that first orders the elements $\{a_1, \ldots, a_m\}$ of A lexicographically and then maps them to $\{\alpha_1, \ldots, \alpha_m\}$ preserving the order.

Let $\sum^\infty = \sum \times \sum \times \cdots$ (i.e., a one-sided countably infinite product. Analogous notation will be used elsewhere.) At each time n, a measurable map $\eta_n : \sum^{n+1} \to D$ is chosen. With $Q_n \stackrel{\Delta}{=} \eta_n(q^n)$, one sets

$$q_{n+1} = i_{Q_n}(l_{Q_n}(Y_{n+1})).$$

This defines $\{q_n\}$ recursively as the quantized process that is to be encoded and transmitted across a communication channel.

The explanation of this scheme is as follows. In case of a fixed quantizer, the finite subset of R^d to which the signal gets mapped can itself be identified with the alphabet $\sum$. In our case, however, this set will vary from one instant to another and therefore must be mapped to a fixed alphabet $\sum$ in a uniquely invertible manner. This is achieved through the map i_A. Assuming that the receiver knows ahead of time the deterministic maps $\{n_n(\cdot)\}$ (later on we argue that a single fixed $\eta(\cdot)$ will suffice), she can reconstruct Q_n as $\eta_n(q^n)$ on having received q^n by time n. In turn, she can reconstruct $i_{Q_n}^{-1}(q_{n+1}) = l_{Q_n}(Y_{n+1})$ as the vector quantized version of Y_{n+1}. The main contribution of the condition (†) is to render the map $A = \{a_1, \ldots, a_m\} \in D \to \{i_A(a_1), \ldots, i_A(a_m)\} \in \sum^*$ continuous. Not only does this make sense from the point of view of robust decoding, but it also makes the control problem we formulate later well-posed.

As mentioned in the introduction, our aim will be to jointly optimize over the choice of $\{\eta_n(\cdot)\}$ the average entropy rate of $\{q_n\}$ ($\approx$ the average code length if the encoding is done optimally) and the average distortion. The conventional rate distortion theoretic formulation would be to minimize the average entropy rate

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n-1} E[H(q_{m+1}/q^m)],$$

$H(\cdot)$ being the (conditional) Shannon entropy, subject to a hard constraint on the distortion

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n-1} E[||Y_m - \bar{q}_m||^2] \le K,$$

where $\bar{q}_m = i_{Q_{m-1}}^{-1}(q_m) = l_{Q_{m-1}}(Y_m)$. We shall, however, consider the simpler problem of minimizing the Lagrangian distortion measure

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n-1} E[H(q_{m+1}/q^m) + \lambda||Y_m - \bar{q}_m||^2], \tag{2.3}$$

where $\lambda > 0$ is a prescribed constant. One may think of λ as a Lagrange multiplier, though, strictly speaking, such an interpretation is lacking given our arbitrary choice thereof.

3. Reduction to the control problem. This section derives the "completely observed" optimal stochastic control problem equivalent to the optimal vector quantization problem described above. In this, we follow the usual "separation" idea of stochastic control by identifying the regular conditional law of state given past observations (in our case, past encodings of the actual observations) as the new state process for the completely observed control problem. The original cost function is rewritten in an equivalent form that displays it as a function of the new state and control processes alone. Under the assumptions of the previous section on the permissible vector quantization schemes (as reflected in our definition of D), the above controlled Markov process is shown to have a transition kernel continuous in the initial state and control. Finally, a relaxation of this control problem is outlined, which allows for a larger class of controls. This is purely a technical convenience required for the proofs of the next section and does not affect our control problem in any essential manner.

Let $\pi_n(dx) \in P(R^s)$ denote the conditional law of X_n given $q^n, n \geq 0$. A standard application of the Bayes rule shows that $\{\pi_n\}$ is given recursively by the nonlinear filter

$$\pi_{n+1}(dx') = \frac{\int\int I\{i_{Q_n}(l_{Q_n}(y)) = q_{n+1}\}\varphi(y, x'|x)dydx'\pi_n(dx)}{\int\int\int I\{i_{Q_n}(l_{Q_n}(y)) = q_{n+1}\}\varphi(y, z|x)dydz\pi_n(dx)}. \tag{3.1}$$

By (†), $l_A^{-1}(i_A^{-1}(a))$ contains an open subset of R^d for any a, A. Given this fact and the condition that $\varphi(\cdot, \cdot|\cdot) > 0$, it follows that the denominator above is strictly positive, and hence the ratio is well defined. The initial condition for the recursion (3.1) is $\pi_0 =$ the conditional law of X_0 given q_0. We assume q_0 to be the trivial quantizer, i.e., $q_0 \equiv 0$, say, so that $\pi_0 =$ the law of X_0. Thus defined, $\{\pi_n\}$ can be viewed as a $P(R^s)$-valued controlled Markov process with a D-valued "control" process $\{Q_n\}$. To complete the description of the control problem, we need to define our cost (2.3) in terms of $\{\pi_n\}, \{Q_n\}$. For this purpose, let $\bar{\varphi}(y|x) \stackrel{\Delta}{=} \int \varphi(y, z|x)dz$ for all $(x, y) \in R^s \times R^d$. Note that for $a \in \sum$,

$$\begin{aligned} P(q_{n+1} = a/q^n) &= E[E[I\{q_{n+1} = a\}/q^n, X^n]/q^n] \\ &= E\left[\int p(X_n, R^s, dy)I\{q_{n+1} = a\}/q^n\right] \\ &= \int \pi_n(dx) \int \bar{\varphi}(y|x)I\{i_{\eta_n(q^n)}(l_{\eta_n(q^n)}(y)) = a\}dy \\ &\stackrel{\Delta}{=} h_a(\pi_n, Q_n), \end{aligned}$$

where $h_a : P(R^s) \times D \to R$ is defined by

$$h_a(\pi, A) = \int \pi(dx)f_a(x, A)$$

with

$$f_a(x, A) = \int \bar{\varphi}(y|x)I\{i_A(l_A(y)) = a\}dy.$$

Also define

$$\begin{aligned} \hat{f}(x, A) &= \int \bar{\varphi}(y|x)||y - l_A(y)||^2 dy, \\ k(\pi, A) &= -\sum_a h_a(\pi, A) \log h_a(\pi, A), \\ r(\pi, A) &= \int \pi(dx)\hat{f}(x, A), \end{aligned}$$

where the logarithm is to the base 2. We assume $f_a(\cdot, A), \hat{f}(\cdot, A)$ to be Lipschitz uniformly in a, A. This would be implied in particular by the condition that $\bar{\varphi}(y/\cdot)$ be Lipschitz uniformly in y. Now (2.3) can be rewritten as

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n-1} E[k(\pi_m, Q_m) + \lambda r(\pi_m, Q_m)]. \tag{3.2}$$

Strictly speaking, we should consider the problem of controlling $\{\pi_n\}$ given by (3.1) so as to minimize the cost (3.2). We shall, however, introduce some further

simplifications, thereby replacing (3.2) by an approximation of the same. Let $\frac{1}{N} > \epsilon^\star > 0$ be a small positive constant. For $n \geq 1$, let $P_n^\star$ denote the simplex of probability vectors in R^n which have each component bounded from below by $\epsilon^\star$. That is,

$$P_n^\star = \left\{ x = [x_1, \ldots, x_n] \in R^n : x_i \in [\epsilon^\star, 1] \ \ \forall\ i, \quad \sum_i x_i = 1 \right\}.$$

Similarly, let

$$P_n = \left\{ x = [x_1, \ldots, x_n] \in R^n : x_i \in [0, 1] \ \ \forall\ i, \quad \sum_i x_i = 1 \right\}$$

denote the entire simplex of probability vectors in R^n. Let $\Pi_n : P_n \to P_n^\star$ denote the projection map. Let $h(\pi, A) = [h_{a_1}(\pi, A), \ldots, h_{a_m}(\pi, A)]$ for $A = \{a_1, \ldots, a_m\}$ and

$$\begin{aligned} \tilde{h}(\pi, A) &= \Pi_{|A|}(h(\pi, A)) \\ &\triangleq [\tilde{h}_{a_1}(\pi, A), \ldots, \tilde{h}_{a_m}(\pi, A)]. \end{aligned}$$

Note that

$$|\log \tilde{h}_a(\pi, A)| \leq -\log \epsilon^\star < \infty \quad \forall\ a, \pi, A. \tag{3.3}$$

Finally, let

$$\tilde{k}(\pi, A) = -\sum_a \tilde{h}_a(\pi, A) \log \tilde{h}_a(\pi, A).$$

The control problem we consider is that of controlling $\{\pi_n\}$ so as to minimize the cost

$$\limsup_{n \to \infty} \frac{1}{n} \sum_{m=0}^{n-1} E[\tilde{k}(\pi_n, Q_n) + \lambda r(\pi_n, Q_n)]. \tag{3.4}$$

Replacing $k(\cdot, \cdot)$ by $\tilde{k}(\cdot, \cdot)$ is a purely technical convenience to suit the needs of the developments to come in section 4. We believe that it should be possible to obtain the same results directly for (3.2), though possibly at the expense of a considerable additional technical overhead.

We shall analyze this problem using techniques of Markov decision processes. With this in mind, call $\{Q_n\}$ a stationary control policy if $Q_n = v(\pi_n)$ for all n for a measurable $v : P(R^s) \to D$. The map $v(\cdot)$ itself may be referred to as the stationary control policy by a standard abuse of notation. Let $(\pi, A) \in P(R^s) \times D \to \phi(\pi, A, d\pi') = P(P(R^s))$ denote the transition kernel of the controlled Markov process $\{\pi_n\}$.

LEMMA 3.1. *The map $\phi(\cdot, \cdot, d\pi')$ is continuous.*

Proof. It suffices to check that for $f \in C_b(P(R^s))$, the map $\int f(y)\phi(\cdot, \cdot, dy)$ is continuous. Let $(\mu_n, A_n) \to (\mu_\infty, A_\infty)$ in $P(R^s) \times D$. Then $\{\mu_n\}$ are tight, and therefore, for any $\epsilon > 0$, we can find a compact $S_\epsilon \subset R^s$ such that $\mu_n(S_\epsilon) > 1 - \epsilon$ for $n = 1, 2, \ldots, \infty$. Fix $\epsilon > 0$ and $S_\epsilon \subset R^s$. By the Stone–Weierstrass theorem, any $f \in C_b(P(R^s))$ can be approximated uniformly on S_ϵ by $\bar{f} \in C_b(P(R^s))$ of the form

$$\bar{f}(\mu) = F\left(\int f_1 d\mu, \ldots, \int f_l d\mu \right)$$

for some $l \geq 1, f_1, \ldots, f_l \in C_b(R^s)$ and $F \in C_b(R^l)$. Then

$$\left| \int f(y)\phi(\mu_n, A_n, dy) - \int f(y)\phi(\mu_\infty, A_\infty, dy) \right| \leq 4\epsilon K + \sup_{\mu \in S_\epsilon} |f(\mu) - \bar{f}(\mu)| + \left| \int \bar{f}(y)\phi(\mu_n, A_n, dy) - \int \bar{f}(y)\phi(\mu_\infty, A_\infty, dy) \right|. \tag{3.5}$$

Let

$$\nu_{ai}(\pi, A) = \int\int f_i(y) I\{i_A(l_A(y)) = a\} \bar{\varphi}(y|x) dy \pi(dx)$$

for $a \in \sum, 1 \leq i \leq l$. Direct verification leads to

$$\int \bar{f}(y)\phi(\pi, A, dy) = \sum_a h_a(\pi, A) F\left(\frac{\nu_{a1}(\pi, A)}{h_a(\pi, A)}, \ldots, \frac{\nu_{al}(\pi, A)}{h_a(\pi, A)} \right). \tag{3.6}$$

Note that for all a,

$$I\{i_{A_n}(l_{A_n}(y)) = a\} \to I\{i_{A_\infty}(l_{A_\infty}(y)) = 0\} \text{ almost everywhere (a.e.),}$$

because this convergence fails only on the boundaries of the regions $l^{-1}_{A_\infty}(b)$, $b \in A_\infty$, which have zero Lebesgue measure. (These are the so called *Voronoi* regions in vector quantization literature, viz., sets in the partition generated by the quantizer $l_{A_\infty}(\cdot)$.) Therefore, for all a, j,

$$f_j(y) I\{i_{A_n}(l_{A_n}(y)) = a\} \to f_j(y) I\{i_{A_\infty}(l_{A_\infty}(y)) = a\} \text{ a.e.}$$

If $x_n \to x_\infty$ in R^s, $\bar{\varphi}(y|x_n) \to \bar{\varphi}(y|x_\infty)$ for all y. Then by Scheffe's theorem [6, p. 26],

$$\bar{\varphi}(y|x_n)dy \to \bar{\varphi}(y|x_\infty)dy$$

in total variation. Hence for any a, j,

$$\int f_j(y) I\{i_{A_n}(l_{A_n}(y)) = a\} \bar{\varphi}(y|x_n) dy \to \int f_j(y) I\{i_{A_\infty}(l_{A_\infty}(y)) = a\} \bar{\varphi}(y|x_\infty) dy.$$

That is, the map

$$(x, A) \to \int f_j(y) I\{i_A(l_A(y)) = a\} \bar{\varphi}(y|x) dy$$

is continuous. It is clearly bounded. The continuity of $\nu_{ia}(\cdot, \cdot)$ follows. That of $h_a(\cdot, \cdot)$ follows similarly. The continuity of the sum in (3.6) then follows by one more application of Scheffe's theorem. Thus the last term on the right-hand side (RHS) of (3.5) tends to zero as $n \to \infty$. Since $\epsilon > 0$ was arbitrary and the second term on the RHS of (3.5) can be made arbitrarily small by a suitable choice of $\bar{f}$, the claim follows. □

We conclude this section with a description of a certain relaxation of this control problem wherein we permit a larger class of control policies, the so-called wide sense admissible controls used in [11]. Let $(\Omega, \mathcal{F}, P)$ denote the underlying probability space, where, without loss of generality, we may suppose that $\mathcal{F} = V_n \mathcal{F}_n$ for $\mathcal{F}_n = \sigma(X_i, Y_i, \xi_i, \xi_i', Q_i, i \leq n)$, $n \geq 0$. Define a new probability measure P_0 on $(\Omega, \mathcal{F})$ as

follows. Let $\psi_n : \sum^{n+1} \times R^m \to P(\sum)$ denote the regular conditional law of q_{n+1} given (q^n, Y_{n+1}) for $n \geq 0$. (Thus we are now allowing for a randomized choice of Q_n, i.e., Q_n is not necessarily a deterministic function of (q^n, Y_{n+1}).) Let $\Gamma \in P(\sum)$ be any fixed probability measure with full support. If, for $n \geq 0, P_n, P_{0n}$, we denote the restrictions of P, P_0 to $(\Omega, \mathcal{F}_n)$, respectively, then $P_n << P_{0n}$ with

$$\frac{dP_n}{dP_{0n}} = \prod_{m=0}^{n-1} \frac{\psi_n(q^m, Y_{m+1})(\{q_{m+1}\})}{\Gamma(\{q_{m+1}\})}, \quad n \geq 1.$$

Then, under $P_0, \{q_n\}$ are independent of $\{X_n, Y_n, \xi_n, \xi_n'\}$ and are i.i.d. with law Γ. We say that $\{Q_n\}$ is a wide sense admissible control if under $P_0, (q_{n+1}, q_{n+2}, \ldots)$ is independent of (q^n, Q^n) for $n \geq 0$. Note that this includes $\{Q_n\}$ of the type $Q_n = \eta_n(q^n)$ for suitable maps $\{\eta_n(\cdot)\}$.

It should be kept in mind that this allows explicit randomization in the choice of $\{Q_n\}$, whence the entropy rate expression in (3.2) or (3.4) is no longer valid. Nevertheless, we continue with wide sense admissible controls in the context of (3.1)–(3.4) because, for us, this is strictly a temporary technical device to facilitate proofs. The dynamic programming formulation that we shall finally arrive at in section 4 will permit us to return without any loss of generality to the apparently more restrictive class of $\{Q_n\}$ we started out with.

4. The vanishing discount limit. This section derives the dynamic programming equations for the equivalent "separated control problem" by extending the traditional "vanishing discount" argument to the present setup. Deriving the dynamic programming equations for the long run average cost control of the separated control problem has been an outstanding open problem in the general case. We solve it here by using in a crucial manner the asymptotic flatness assumption introduced earlier. It should be noted that this assumption was not required at all in the development thus far and is included purely for facilitating the vanishing discount limit argument that follows. In particular, it could be dispensed with altogether were we to consider the finite horizon or infinite horizon discounted cost. For an alternative set of conditions (also strong) under which the dynamic programming equations for the average cost control under partial observations have been derived, see [21].

Our first step will be to modify the construction at the end of section 3 so as to construct on a common probability space two controlled nonlinear filters with a common control process but differing in their initial condition. This allows us to compare discounted cost value functions for two different initial laws. In turn, this allows us to show that their difference, with one of the two initial laws fixed arbitrarily, remains bounded and equicontinuous with respect to a certain complete metric on the space of probability measures, as the discount factor approaches unity. (This is where one uses the condition of asymptotic flatness.) The rest of the derivation mimics the classical arguments in this field.

For $\alpha \in (0, 1)$, consider the discounted control problem of minimizing

$$J_\alpha(\pi_0, \{Q_n\}) = E\left[\sum_{n=0}^{\infty} \alpha^n (\tilde{k}(\pi_n, Q_n) + \lambda r(\pi_n, Q_n))\right] \tag{4.1}$$

over $\Phi \triangleq$ the set of all wide sense admissible controls, with the prescribed π_0. Define the associated value function $V_\alpha : P(R^s) \to R$ by

$$V_\alpha(\pi_0) = \inf_{\Phi} J(\pi_0, \{Q_n\}).$$

Standard dynamic programming arguments show that $V_\alpha(\cdot)$ satisfies

$$V_\alpha(\pi) = \min_A \left[k(\pi, A) + \lambda r(\pi, A) + \beta \int \phi(\pi, A, d\pi') V_\alpha(\pi') \right] \tag{4.2}$$

for $\pi \in P(R^s)$. We shall arrive at the dynamic programming equation for our original problem by taking a "vanishing discount" limit of a variant of (4.2). For this purpose, we need to compare $V_\alpha(\cdot)$ for two distinct values of its argument. In order to do so, we first set up a framework for comparing (4.1) for two choices of π_0 but with a "common" wide sense admissible control $\{Q_n\}$. This will be done by modifying the construction at the end of the preceding section. Let $(\Omega, \mathcal{F}, P_0)$ be a probability space on which we have (i) R^s-valued, possibly dependent random variables $\hat{X}_0, \tilde{X}_0$, with laws π_0, π_0', respectively; (ii) R^m-valued i.i.d. random processes $\{\xi_m\}, \{\xi_m'\}$, independent of each other and of $[\hat{X}_0, \tilde{X}_0]$ with laws as in (2.1), (2.2); and (iii) $\sum$-valued i.i.d. random sequences $\{\hat{q}_m\}, \{\tilde{q}_m\}$ with law Γ. Also defined on $(\Omega, \mathcal{F}, P_0)$ is a D-valued process $\{Q_n\}$ independent of $([\hat{X}_0, \tilde{X}_0], \{\xi_n\}, \{\xi_n'\}, \{\tilde{q}_n\})$ and satisfying the following. For $n \geq 0$, $(\hat{q}_{n+1}, \hat{q}_{n+2}, \ldots)$ is independent of $Q^n, \hat{q}^n$. Let $(\hat{X}_n, \hat{Y}_n), (\tilde{X}_n, \tilde{Y}_n)$ be solutions to (2.1), (2.2) with $\hat{X}_0, \tilde{X}_0$ as above. Without loss of generality, we may suppose that $\mathcal{F} = V_n \mathcal{F}_n$ with $\mathcal{F}_n = \sigma(\hat{X}^n, \tilde{X}^n, \hat{Y}^n, \tilde{Y}^n, \hat{q}^n, \tilde{q}^n, Q^n), n \geq 0$. Define a new probability measure P on $(\Omega, \mathcal{F})$ as follows. If P_n, P_{0n} denote the restrictions of P, P_0, respectively, to $(\Omega, \mathcal{F}_n), n \geq 0$, then $P_n << P_{0n}$ with

$$\frac{dP_n}{dP_{0n}} = \prod_{m=0}^{n-1} \frac{\psi_n(\hat{q}^n, \hat{Y}_{n+1})(\{\hat{q}_{n+1}, \}) \psi_n'(\tilde{q}^n, \tilde{Y}_{n+1})(\{\tilde{q}_{n+1}\})}{\Gamma(\{\hat{q}_{n+1}\}) \Gamma(\{\tilde{q}_{n+1}\})},$$

where the ψ_n (respectively, ψ_n') are the regular conditional laws of $Q_n(\hat{Y}_{n+1})$ given $(\hat{q}^n, \hat{Y}_{n+1})$ (respectively, of $Q_n(\tilde{Y}_{n+1})$ given $(\tilde{q}^n, \tilde{Y}_{n+1})$) for $n \geq 0$.

What this construction achieves is the identification of each wide sense admissible control $\{Q_n\}$ for initial law $\hat{\pi}_0$ with one wide sense admissible control for $\tilde{\pi}_0$. (This identification can be many-one.) By a symmetric argument that interchanges the roles of $\hat{\pi}_0$ and $\tilde{\pi}_0$, we can identify each wide sense admissible control for $\tilde{\pi}_0$ with one for $\hat{\pi}_0$. Now suppose that $V_\alpha(\hat{\pi}_0) \leq V_\alpha(\tilde{\pi}_0)$. Then for a wide sense admissible control $\{Q_n\}$ that is optimal for $\hat{\pi}_0$ (existence of this follows by standard dynamic programming arguments), we have

$$\begin{aligned} |V_\alpha(\hat{\pi}_0) - V_\alpha(\tilde{\pi}_0)| &= V_\alpha(\tilde{\pi}_0) - V_\alpha(\hat{\pi}_0) \\ &\leq J_\alpha(\tilde{\pi}_0, \{Q_n\}) - J_\alpha(\hat{\pi}_0, \{Q_n\}) \\ &\leq \sup_\Phi |J_\alpha(\tilde{\pi}_0, \{Q_n\}) - J_\alpha(\hat{\pi}_0, \{Q_n\})|, \end{aligned}$$

where we use the above identification. If $V_\alpha(\hat{\pi}_0) \geq V_\alpha(\tilde{\pi}_0)$, a symmetric argument applies. Thus we have proved the following lemma.

LEMMA 4.1.

$$|V_\alpha(\hat{\pi}_0) - V_\alpha(\tilde{\pi}_0)| \leq \sup_\Phi |J_\alpha(\hat{\pi}_0, \{Q_n\}) - J_\alpha(\tilde{\pi}_0, \{Q_n\})|.$$

Next, let $P_1(R^s) = \{\mu \in P(R^s) : \int ||x|| \mu(dx) < \infty\}$, topologized by the (complete) Vasserstein metric [20]

$$\rho(\mu_1, \mu_2) = \inf E[||X - Y||],$$

where the infimum is over all joint laws of (X, Y) such that the law of X (respectively, Y) is μ_1 (respectively, μ_2). We shall assume from now on that $\pi_0 \in P_1(R^s)$. Given the linear growth condition on $g(\cdot, y), h(\cdot, y)$ of (2.1), (2.2), uniformly in y, it is then easily deduced that $E[||X_n||] < \infty$ for all n and therefore $\pi_n \in P_1(R^s)$ almost surely (a.s.) for all n. Thus we may and do view $\{\pi_n\}$ as a $P_1(R^s)$-valued process. We then have the following lemma.

LEMMA 4.2. *For $\hat{\pi}_0, \tilde{\pi}_0 \in P_1(R^s)$ and $\alpha > 0, |V_\alpha(\hat{\pi}_0) - V_\alpha(\tilde{\pi}_0)| \le K\rho(\hat{\pi}_0, \tilde{\pi}_0)$.*

Proof. Let $\{\hat{\pi}_n\}, \{\tilde{\pi}_n\}$ be solutions to (3.1) with initial conditions $\hat{\pi}_0, \tilde{\pi}_0$, respectively, and a "common" wide sense admissible control $\{Q_n\} \in \Phi$. Then for $\{\hat{X}_n\}, \{\tilde{X}_n\}$ as above (with K denoting a generic positive constant that may change from step to step)

$$\begin{aligned}
&|E[r(\hat{\pi}_n, Q_n)] - E[r(\tilde{\pi}_n, Q_n]| \\
&\quad = |E[\hat{f}(\hat{X}_n, Q_n)] - E[\hat{f}(\tilde{X}_n, Q_n)]| \\
&\quad \le E[|\hat{f}(\hat{X}_n, Q_n) - \hat{f}(\tilde{X}_n, Q_n)|] \\
&\quad \le KE[||\hat{X}_n - \tilde{X}_n||]
\end{aligned}$$

(by the Lipschitz condition on $\hat{f}$)

$$\le K\beta^n E[||\hat{X}_0 - \tilde{X}_0||]$$

(by asymptotic flatness).

Now consider

$$|E[\tilde{k}(\hat{\pi}_n, Q_n)] - E[\tilde{k}(\tilde{\pi}_n, Q_n]|.$$

Suppose that $E[\tilde{k}(\hat{\pi}_n, Q_n)] \ge E[\tilde{k}(\tilde{\pi}_n, Q_n)]$. Then

$$\begin{aligned}
&|E[\tilde{k}(\hat{\pi}_n, Q_n)] - E[\tilde{k}(\tilde{\pi}_n, Q_n)]| \\
&\quad = E[\tilde{k}(\hat{\pi}_n, Q_n)] - E[\tilde{k}(\tilde{\pi}_n, Q_n)] \\
&\quad = E\left[\sum_a \tilde{h}_a(\tilde{\pi}_n, Q_n) \log \tilde{h}_a(\tilde{\pi}_n, Q_n)\right] - E\left[\sum_a \tilde{h}_a(\hat{\pi}_n, Q_n) \log \tilde{h}_a(\hat{\pi}_n, Q_n)\right] \\
&\quad = E\left[\sum_a \left(\tilde{h}_a(\tilde{\pi}_n, Q_n) \log \tilde{h}_a(\tilde{\pi}_n, Q_n) - \tilde{h}_a(\hat{\pi}_n, Q_n) \log \tilde{h}_a(\tilde{\pi}_n, Q_n)\right.\right. \\
&\qquad \left.\left. + \tilde{h}_a(\hat{\pi}_n, Q_n) \log \frac{\tilde{h}_a(\tilde{\pi}_n, Q_n)}{\tilde{h}_a(\hat{\pi}_n, Q_n)}\right)\right] \\
&\quad \le E\left[\sum_a (\tilde{h}_a(\tilde{\pi}_n, Q_n) - \tilde{h}_a(\hat{\pi}_n, Q_n)) \log \tilde{h}_a(\tilde{\pi}_n, Q_n))\right]
\end{aligned}$$

(by Jensen's inequality)

$$\begin{aligned}
&\le E\left[\sum_a (f_a(\tilde{X}_n, Q_n) - f_a(\hat{X}_n, Q_n)) \log \tilde{h}_a(\tilde{\pi}_n, Q_n))\right] \\
&\le KE[||\tilde{X}_n - \hat{X}_n||] \\
&\le K\beta^n E[||\tilde{X}_0 - \hat{X}_0||],
\end{aligned}$$

where we use (3.3) to arrive at the second to last inequality. A symmetric argument works if $E[\tilde{k}(\hat{\pi}_n, Q_n)] \le E[\tilde{k}(\tilde{\pi}_n, Q_n)]$, leading to the same conclusion. Combining

everything, we have

$$|E[\tilde{k}(\tilde{\pi}_n, Q_n) + \lambda r(\tilde{\pi}_n, Q_n)] - E[\tilde{k}(\hat{\pi}_n, Q_n) + \lambda r(\hat{\pi}_n, Q_n)]| \leq K\beta^n E[||\hat{X}_0 - \tilde{X}_0||].$$

Therefore, by Lemma 4.1,

$$|V_\alpha(\hat{\pi}_0) - V_\alpha(\tilde{\pi}_0)| \leq K \sum_n \beta^n \alpha^n E[||\hat{X}_0 - \tilde{X}_0||] \leq \frac{K}{1-\beta} E[||\hat{X}_0 - \tilde{X}_0||].$$

For any $\epsilon > 0$, we can render

$$E[||\hat{X}_0 - \tilde{X}_0||] \leq \rho(\hat{\pi}_0, \tilde{\pi}_0) + \epsilon$$

by suitably choosing the joint law of $(\hat{X}_0, \tilde{X}_0)$. Since $\epsilon > 0$ is arbitrary, the claim follows. □

Fix $\pi^\star \in P(R^s)$ and define $\bar{V}_\alpha(\pi) = V_\alpha(\pi) - V_\alpha(\pi^\star)$ for $\pi \in P(R^s), \alpha \in (0,1)$. By the above lemma, $\bar{V}_\alpha(\cdot)$ is bounded equicontinuous. Letting $\alpha \to 1$, we use the Arzela–Ascoli theorem to conclude that $\bar{V}_\alpha(\cdot)$ converges in $C(P_1(R^s))$ to some $V(\cdot)$ along a subsequence $\{\alpha(n)\}, \alpha(n) \to 1$. By dropping to a further subsequence if necessary, we may also suppose that $\{(1-\alpha(n))V_{\alpha(n)}(\pi^*)\}$, which is clearly bounded, converges to some $\gamma \in R$ as $n \to \infty$. These $V(\cdot), \gamma$ will turn out to be, respectively, the value function and optimal cost for our original control problem.

Our main result is the following theorem.

THEOREM 4.3.

(i) *$(V(\cdot), \gamma)$ solve the dynamic programming equation*

$$V(\pi) = \min_u \left(\tilde{k}(\pi, u) + \lambda r(\pi, u) + \int \phi(\pi, u, d\pi') V(\pi') - \gamma \right). \tag{4.3}$$

(ii) *γ is the optimal cost, independent of the initial condition. Furthermore, a stationary policy $v(\cdot)$ is optimal for any initial condition if*

$$v(\pi) \in \text{Argmin} \left(\tilde{k}(\pi, \cdot) + \lambda r(\pi, \cdot) + \int \phi(\pi, \cdot, d\pi') V(\pi') \right) \quad \forall\ \pi.$$

In particular, an optimal stationary policy exists.

(iii) *If $v(\cdot)$ is an optimal stationary policy and μ is a corresponding ergodic probability measure for $\{\pi_n\}$, then*

$$V(\pi) = \tilde{k}(\pi, v(\pi)) + \lambda r(\pi, v(\pi)) + \int \phi(\pi, v(\pi), d\pi') V(\pi') - \gamma, \ \mu\text{-a.s.}$$

Proof. For (i) rewrite (4.2) as

$$\bar{V}_\alpha(\pi) = \min_u \left(\tilde{k}(\pi, u) + \lambda r(\pi, u) + \alpha \int \phi(\pi, u, d\pi') \bar{V}_\alpha(\pi') - (1-\alpha) V_\alpha(\pi^\star) \right).$$

Let $\alpha \to 1$ along $\{\alpha(n)\}$ to obtain (4.3).

For (ii) note that the first two statements follow by a standard argument which may be found, e.g., in [15, Theorem 5.2.4, pp. 80–81]. The last claim follows from a standard measurable selection theorem—see, e.g., [22].

For (iii) note that the claim holds if "=" is replaced by "≤". If the claim is false, we can integrate both sides with respect to μ to obtain

$$\gamma < \int (\tilde{k}(\pi, v(\pi)) + \lambda r(\pi, v(\pi)))\mu(d\pi).$$

The RHS is the cost under $v(\cdot)$, whereby this inequality contradicts the optimality of $v(\cdot)$. The claim follows. □

This result opens up the possibility of exploiting the computational machinery of Markov decision theory (see, e.g., [2], [18], [21]) for code design.

Finally, we briefly consider the decoder's problem. If transmission is error free, the decoder can construct $\{\pi_n\}$ recursively given $\{q_n\}$ and the stationary policy $v(\cdot)$. Then $\{X_n\}, \{Y_n\}$ may be estimated by the maximum a posteriori (MAP) estimates:

$$\hat{X}_n = \text{argmax } \pi_n(\cdot),$$
$$\hat{Y}_n = \text{argmax} \left(\iint I\{i_{Q_{n-1}}(l_{Q_{n-1}}(\cdot)) = q_{n+1}\}\varphi(\cdot, z|x)dz \ \pi_{n-1}(dx) \right).$$

Suppose the decoder receives $\{q_n\}$ through a noisy but memoryless channel with input alphabet $\sum$ and output alphabet another finite set O, with transition probabilities $\tilde{p}(i,j), i \in D, j \in O$. Thus $\tilde{p}(i,j) \geq 0, \sum_l \tilde{p}(i,l) = 1$ for all i, j. Let d_n be the channel output at time n.

The decoder can estimate (X_n, Y_n) given $d^n, n \geq 0$, but this is no longer easy because we cannot reconstruct $\{Q_n\}$ exactly in absence of his knowledge of $\{\pi_n\}, \{q_n\}$. Thus he should estimate $\{q_n\}$ by $\{\hat{q}_n\}$, say (e.g., by maximum likelihood), given $\{d_n\}$ and use these estimates in place of $\{q_n\}$ in the nonlinear filter for $\{\pi_n\}$, giving an approximation $\{\hat{\pi}_n\}$ to $\{\pi_n\}$. The guess for Q_n then is $v(\hat{\pi}_n), n \geq 0$.

5. Conclusions and extensions. In this paper we have considered the problem of optimal sequential vector quantization of a stationary Markov source. We have formulated the problem as a stochastic control problem. We have used the methodology of Markov decision theory. Further, we have shown that the conditional law of the source given the quantized past is a sufficient statistic for the problem. Thus the optimal encoding scheme has a separated structure. The conditional laws are given recursively by the nonlinear filter described in (3.1). The optimal policy is characterized by Theorem 4.3.

The next step is to apply traditional Markov decision problem approximation techniques to compute approximate schemes. If we have access to training data, then we can use the tools of reinforcement learning. Here the idea is to parametrize the value function space or the control law itself and apply stochastic approximation techniques to optimize those parameters.

In general, the nonlinear filter recursion is very complicated. In the literature people have approximated this by a linear prediction of the mean. These linear predictive methods can be considered an approximation to the general nonlinear filter.

REFERENCES

[1] G. Basak and R. N. Bhattacharya, *Stability in distribution for a class of singular diffusions*, Ann. Probab., 20 (1992), pp. 312–321.

[2] D. Bertsekas and J. Tsitsiklis, *Neurodynamic Programming*, Athena Scientific, Belmont, MA, 1996.

[3] A. Bist, *Differential state quantization of high order Gauss-Markov processes*, in Proceedings of the IEEE Data Compression Conference, Snowbird, UT, 1994, pp. 62–71.
[4] V. S. Borkar, *Optimal Control of Diffusion Processes*, Pitman Lecture Notes in Math. 203, Longman Scientific and Technical, Harlow, UK, 1989.
[5] V. S. Borkar, *Topics in Controlled Markov Chains*, Pitman Lecture Notes in Math. 240, Longman Scientific and Technical, Harlow, UK, 1991.
[6] V. S. Borkar, *Probability Theory: An Advanced Course*, Springer-Verlag, New York, 1995.
[7] P. Chou, T. Lookabaugh, and R. Gray, *Entropy-constrained vector quantization*, IEEE Trans. Acoust. Speech Signal Process., 37 (1989), pp. 31–42.
[8] P. Chou and T. Lookabaugh, *Conditional entropy-constrained vector quantization of linear predictive coefficients*, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 1, Albuquerque, NM, 1990, pp. 197–200.
[9] T. Cover and J. Thomas, *Elements of Information Theory*, John Wiley, New York, 1991.
[10] J. G. Dunham, *An iterative theory for code design*, in Proceedings of the IEEE International Symposium on Information Theory, St. Jovite, QC, Canada, 1983, pp. 88–90.
[11] W. Fleming and E. Pardoux, *Optimal control for partially observed diffusions*, SIAM J. Control Optim., 20 (1982), pp. 261–285.
[12] N. T. Gaarder and D. Slepian, *On optimal finite-state digital transmission systems*, IEEE Trans. Inform. Theory, 28 (1982), pp. 167–186.
[13] R. E. Gallager, *Information Theory and Reliable Communication*, John Wiley, New York, 1968.
[14] P. Hall and C. C. Heyde, *Martingale Limit Theory and Its Applications*, Academic Press, New York, London, 1980.
[15] O. Hernandez-Lerma and J. B. Lasserre, *Discrete-Time Markov Control Processes*, Springer-Verlag, New York, 1996.
[16] J. C. Kieffer, *Stochastic stability of feedback quantization schemes*, IEEE Trans. Inform. Theory, 28 (1982), pp. 248–254.
[17] E. Levine, *Stochastic vector quantization and stochastic VQ with state feedback using neural networks*, in Proceedings of the IEEE Data Compression Conference, Snowbird, UT, 1996, pp. 330–339.
[18] S. P. Meyn, *Algorithms for optimization and stabilization of controlled Markov chains*, in SADHANA: Indian Academy of Sciences Proceedings in Engineering Sciences 24, Bangalore, 1999, pp. 339–368.
[19] D. Neuhoff and R. K. Gilbert, *Causal source codes*, IEEE Trans. Inform. Theory, 28 (1982), pp. 701–713.
[20] S. T. Rachev, *Probability Metrics and the Stability of Stochastic Models*, John Wiley, Chichester, UK, 1991.
[21] W. J. Runggaldier and L. Stettner, *Approximations of Discrete Time Partially Observed Control Problems*, Applied Maths. Monographs 6, Giardini Editori e Stampatori, Pisa, Italy, 1994.
[22] D. H. Wagner, *Survey of measurable selection theorems*, SIAM J. Control Optim., 15 (1977), pp. 859–903.
[23] J. Walrand and P. P. Varaiya, *Optimal causal coding-decoding problems*, IEEE Trans. Inform. Theory, 29 (1983), pp. 814–820.
[24] H. Witsenhausen, *On the structure of real-time source coders*, The Bell System Technical Journal, 58 (1979), pp. 1437–1451.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 1, pp. 149–165

EXPONENTIAL STABILITY OF AN ABSTRACT NONDISSIPATIVE LINEAR SYSTEM*

KANGSHENG LIU†, ZHUANGYI LIU‡, AND BOPENG RAO§

Abstract. In this paper we consider an abstract linear system with perturbation of the form

$$\frac{dy}{dt} = Ay + \varepsilon By$$

on a Hilbert space $\mathcal{H}$, where A is skew-adjoint, B is bounded, and ε is a positive parameter. Motivated by a work of Freitas and Zuazua on the one-dimensional wave equation with indefinite viscous damping [P. Freitas and E. Zuazua, *J. Differential Equations*, 132 (1996), pp. 338–352], we obtain a sufficient condition for exponential stability of the above system when B is not a dissipative operator. We also obtain a Hautus-type criterion for exact controllability of system (A, G), where G is a bounded linear operator from another Hilbert space to $\mathcal{H}$. Our result about the stability is then applied to establish the exponential stability of several elastic systems with indefinite viscous damping, as well as the exponential stabilization of the elastic systems with noncolocated observation and control.

Key words. linear elastic system, exponential stability, exact controllability, Hautus-type criterion, indefinite damping, stabilization

AMS subject classifications. 93D20, 93B05, 93D15, 35B35, 35B37

PII. S0363012999364930

1. Introduction.

We consider a linear evolution equation

$$\begin{cases} \dfrac{d}{dt}y(t) = \mathcal{A}_\varepsilon y(t) \equiv (A + \varepsilon B)y(t), \\ y(0) = y_0 \end{cases} \tag{1.1}$$

in a Hilbert space $\mathcal{H}$, where A is a densely defined, closed linear operator with domain $\mathcal{D}(\mathcal{A})$. We assume that

(H1) A is skew-adjoint ($A^* = -A$) and has a compact resolvent, and

(H2) B is a bounded linear operator on $\mathcal{H}$ with $\|B\| = N_b$.

Under assumptions (H1) and (H2), we know that the operator $\mathcal{A}_\varepsilon$ generates a C_0 semigroup $S_\varepsilon(t)$ on $\mathcal{H}$ (see [9]). In this paper we study mainly the exponential stability of the above system, i.e., that there exist $\mu > 0$ and $M \geq 1$ such that

$$\|S_\varepsilon(t)\| \leq Me^{-\mu t} \quad \forall\, t \geq 0. \tag{1.2}$$

When $\varepsilon = 1$, this problem has been investigated extensively for both bounded and unbounded operator B. These works are based on the assumption of the *dissipativeness* of B,

$$\mathrm{Re}\langle By, y\rangle \leq 0 \quad \forall\, y \in \mathcal{D}(A), \tag{1.3}$$

*Received by the editors December 8, 1999; accepted for publication (in revised form) November 10, 2000; published electronically May 31, 2001.
http://www.siam.org/journals/sicon/40-1/36493.html

†Department of Applied Mathematics, Zhejiang University, Hangzhou, 310027, China (ksliu@mail.hz.zj.cn). This author's work was partially supported by the National Key Project of China and by NSFC grant 69874034.

‡Department of Mathematics and Statistics, University of Minnesota, Duluth, MN 55812 (zliu@d.umn.edu).

§Institut de Recherche Mathématique Avancée, Université de Loius Pasteur de Strasbourg, 7 Rue René-Descartes, 67084 Strasbourg Cedex, France (rao@math.u-strasbg.fr).

which implies that the energy of the system, $E(t) = \|S_\varepsilon(t)\|^2$, is a decreasing function of time. Clearly, this is not a necessary condition for $E(t)$ being upper bounded by a function which tends to zero exponentially. A natural question to ask is the following. Without the dissipativeness of B, can we still obtain (1.2) under some extra conditions? This problem is quite significant in the control theory for distributed parameter systems because

(a) the optimality systems resulting from the regulators are nondissipative;
(b) the closed-loop systems by feedback with noncolocated sensors and actuators are nondissipative;
(c) the perturbations arising from undetermined parts of models are nondissipative in general.

Such a question was first raised in [2] for the one-dimensional wave equation

$$\begin{cases} w_{tt}(x,t) = w_{xx}(x,t) - d(x)w_t(x,t), \quad 0 < x < 1, \quad t > 0, \\ w(0,t) = w(1,t) = 0, \quad t > 0, \\ w(x,0) = w_0(x), \ \ w_t(x,0) = w_1(x), \quad 0 < x < 1, \end{cases} \tag{1.4}$$

where d is a smooth function and changes sign on $(0,1)$. It was conjectured that (1.2) holds if

$$I_n \equiv \int_0^1 d(x) \sin^2 n\pi x \mathrm{d}x \geq C_0 > 0, \quad n = 1, 2, \ldots. \tag{1.5}$$

It turns out that (1.5) is not enough to ensure exponential stability. When $\|d\|_{L^\infty}$ becomes large enough, there will be eigenvalues of the system (1.2) with positive real part (see [3]). Thus, in order to have exponential stability, the damping coefficient must not only satisfy (1.5), but also has a small L^∞ norm. Later on, Freitas and Zuazua [4] considered the modified system of (1.4):

$$\begin{cases} w_{tt}(x,t) = w_{xx}(x,t) - \varepsilon d(x)w_t(x,t), \quad 0 < x < 1, \quad t > 0, \\ w(0,t) = w(1,t) = 0, \quad t > 0, \\ w(0,t) = w_0(x), \ w_t(x,0) = w_1(x), \quad 0 < x < 1. \end{cases} \tag{1.6}$$

They proved that if $d \in BV(0,1)$ and the condition (1.5) holds, then there exist positive constants ε_0, M, ω, depending only on the function d, such that for all $0 < \varepsilon < \varepsilon_0$,

$$E(t) = \int_0^1 (|w_x|^2 + |w_t|^2)\mathrm{d}x \leq Me^{-\varepsilon\omega t}E(0) \ \ \forall t > 0, \tag{1.7}$$

for every finite energy solution of (1.6). Their result was further extended in [1] to the equation

$$w_{tt} = w_{xx} - 2\varepsilon d(x)w_t - b(x)w, \tag{1.8}$$

where $b \in L^1(0,1)$.

These works lead us to the current study in this paper. Instead of working on a particular PDE system, we would like to obtain a general result along the line developed in [2]. Although the shooting method used in [4] and [1] is no longer applicable to our abstract problem, the analysis in these papers does provide us with valuable information on how to impose additional conditions in order to guarantee (1.2). In the next section, we estimate the growth rate of the semigroup $S_\varepsilon(t)$, by

a formula for the type of a C_0 semigroup on a Hilbert space due to Huang [7] and Prüss [10]. The exponential stability of the semigroup follows from a negative growth rate. It is well known that the exponential stability of a linear system reversible in time is always connected with exact controllability of the corresponding system. Thanks to the fact that our sufficient condition for exponential stability of (1.1) is also necessary when $-B$ is symmetrical and nonnegative, in section 3 we obtain a Hautus-type criterion for exact controllability of system (A, G), where G is a bounded linear operator from another Hilbert space to $\mathcal{H}$. In section 4, we apply the result on exponential stability to several elastic systems (such as string, Euler–Bernoulli beam, Timoshenko beam, and two-dimensional Schrödinger equations) with indefinite viscous damping or nondissipative perturbation arising from feedback by noncolocated observation and control.

2. Sufficient condition for exponential stability. Under the condition (H1), there is an orthonormal base of $\mathcal{H}$ consisting of eigenvectors of A,

$$\{\phi_n \mid n = 1, 2, \ldots\}, \tag{2.1}$$

such that

$$\begin{cases} A\phi_n = \boldsymbol{i}\beta_n\phi_n, \quad n = 1, 2, \ldots, \ \beta_n \in \mathbb{R}, \\ 0 \le |\beta_1| \le |\beta_2| \le \cdots \le |\beta_n| \le |\beta_{n+1}| \to \infty. \end{cases} \tag{2.2}$$

We have taken multiple eigenvalues into account. Every eigenvalue has a finite multiplicity.

For each $\gamma > 0$, set

$$\Sigma_\gamma = \left\{ \psi = \sum_{n \in I_{\gamma,m}} a_n\phi_n \;\Big|\; \sum_{n \in I_{\gamma,m}} |a_n|^2 = 1, \ m \in \mathbb{N}, \ a_n \in \mathbb{C} \right\}, \tag{2.3}$$

where

$$I_{\gamma,m} = \{n \in \mathbb{N} | \quad |\beta_n - \beta_m| < \gamma\}. \tag{2.4}$$

Let

$$C_\gamma = \inf_{\psi \in \Sigma_\gamma} \mathrm{Re}\langle -B\psi, \psi\rangle. \tag{2.5}$$

Note that $I_{\gamma,m} = I_{\gamma,l}$ if $\beta_m = \beta_l$ and $C_{\gamma_1} \ge C_{\gamma_2}$ for $0 < \gamma_1 < \gamma_2$. We further assume that

$$C_\gamma > 0 \qquad \text{for some} \quad \gamma > 0. \tag{H3}$$

Denote the type of semigroup $S_\varepsilon(t)$ by

$$\omega_0(\mathcal{A}_\varepsilon) = \lim_{t \to \infty} \frac{\ln \|S_\varepsilon(t)\|}{t} \tag{2.6}$$

and the spectral bound of $\mathcal{A}_\varepsilon$ by

$$\sigma_0(\mathcal{A}_\varepsilon) = \sup\{\mathrm{Re}\lambda | \ \lambda \in \sigma(\mathcal{A}_\varepsilon)\}. \tag{2.7}$$

We shall use a result in [7, 10] which states

$$\omega_0(\mathcal{A}_\varepsilon) = \inf \left\{ s > \sigma_0(\mathcal{A}_\varepsilon) \mid \sup_{\mathrm{Re}\lambda = s} \|(\lambda I - \mathcal{A}_\varepsilon)^{-1}\| < +\infty \right\}. \tag{2.8}$$

THEOREM 2.1. *Under the assumptions* (H1)–(H3), *for every* $C \in [0, C_\gamma)$ *it holds that*

$$\omega_0(\mathcal{A}_\varepsilon) < -\varepsilon C \tag{2.9}$$

whenever

$$0 < \varepsilon < \frac{\gamma(\sqrt{N_b^2 + C_\gamma(C_\gamma - C)} - N_b)}{2C_\gamma\sqrt{N_b^2 - C^2}}. \tag{2.10}$$

In particular, $S_\varepsilon(t)$ *is exponentially stable if*

$$0 < \varepsilon < \frac{\gamma}{2N_bC_\gamma}\left(\sqrt{N_b^2 + C_\gamma^2} - N_b\right). \tag{2.11}$$

Proof. We will prove that for every $\sigma \geq -C$ there exists $\delta_\varepsilon > 0$ such that

$$\|(\varepsilon\sigma + \boldsymbol{i}\tau)y - \mathcal{A}_\varepsilon y\| \geq \delta_\varepsilon\|y\| \qquad \forall\ \tau \in \mathbb{R},\ y \in \mathcal{D}(\mathcal{A}_\varepsilon). \tag{2.12}$$

Since for $y \in \mathcal{D}(A)$, $\sigma, \tau \in \mathbb{R}$,

$$\|(\varepsilon\sigma + \boldsymbol{i}\tau)y - \mathcal{A}_\varepsilon y\|\|y\| \geq \mathrm{Re}\langle(\varepsilon\sigma + \boldsymbol{i}\tau)y - \mathcal{A}_\varepsilon y\ ,\ y\rangle \geq \varepsilon(\sigma - N_b)\|y\|^2,$$

(2.12) holds for all $\sigma > N_b$. If (2.12) is false for some $\sigma \in [-C, N_b]$, then there exist a sequence of real numbers τ_p and a sequence of normalized vectors $y_p \in \mathcal{D}(\mathcal{A}_\varepsilon)$ such that

$$((\varepsilon\sigma + \boldsymbol{i}\tau_p)I - \mathcal{A}_\varepsilon)y_p \equiv f_p \to 0 \quad \text{in} \quad \mathcal{H} \quad \text{as} \quad p \to +\infty. \tag{2.13}$$

From (2.13) we have

$$\sigma = \frac{1}{\varepsilon}\langle f_p - (\boldsymbol{i}\tau_p I - A)y_p + \varepsilon By_p\ ,\ y_p\rangle = \lim_{p\to+\infty}\mathrm{Re}\langle By_p, y_p\rangle. \tag{2.14}$$

Moreover, (2.13)–(2.14) imply that

$$\begin{aligned}\|(\boldsymbol{i}\tau_p I - A)y_p\|^2 &= \|f_p - \varepsilon(\sigma I - B)y_p\|^2\\ &\leq \|f_p\|^2 + 2\varepsilon\|\sigma I - B\|\|f_p\| + \varepsilon^2\|(\sigma I - B)y_p\|^2\\ &= o(1) + \varepsilon^2(\sigma^2 - 2\sigma\mathrm{Re}\langle By_p, y_p\rangle + \|By_p\|^2)\\ &\leq \varepsilon^2(N_b^2 - \sigma^2) + o(1).\end{aligned}$$

Thus for any $\delta > 0$ there exists $N \in \mathbb{N}$ such that

$$\|(\boldsymbol{i}\tau_p I - A)y_p\|^2 \leq \varepsilon^2(N_b^2 - \sigma^2 + \delta) \quad \forall\ p > N. \tag{2.15}$$

We expand y_p for $p > N$ in the eigenvectors of A:

$$y_p = \sum_{n=1}^{\infty}\langle y_p, \phi_n\rangle\phi_n. \tag{2.16}$$

Substituting (2.16) into (2.15) yields

$$\sum_{n=1}^{\infty}|\tau_p - \beta_n|^2|\langle y_p, \phi_n\rangle|^2 \leq \varepsilon^2(N_b^2 - \sigma^2 + \delta). \tag{2.17}$$

Choose $m = m(p)$ such that

$$|\tau_p - \beta_m| = \min\{|\tau_p - \beta_n| \mid n \in \mathbb{N}\}. \tag{2.18}$$

Then we have

$$\frac{\gamma}{2} < |\tau_p - \beta_n| \qquad \forall n \notin I_{\gamma,m}. \tag{2.19}$$

In fact, if $|\tau_p - \beta_m| \geq \frac{\gamma}{2}$, (2.19) holds obviously. If $|\tau_p - \beta_m| < \frac{\gamma}{2}$, then (2.19) follows from $|\tau_p - \beta_n| \geq |\beta_n - \beta_m| - |\tau_p - \beta_m|$. Combination of (2.17) and (2.19) gives that

$$\frac{\gamma^2}{4} \sum_{n \notin I_{\gamma,m}} |\langle y_p, \phi_n \rangle|^2 \leq \varepsilon^2 (N_b^2 - \sigma^2 + \delta). \tag{2.20}$$

Define

$$z_p = \sum_{n \in I_{\gamma,m}} \langle y_p, \phi_n \rangle \phi_n, \qquad p > N. \tag{2.21}$$

Then (2.20) implies that

$$\|y_p - z_p\| \leq \frac{2\varepsilon}{\gamma} \sqrt{N_b^2 - \sigma^2 + \delta}, \qquad 1 \geq \|z_p\|^2 \geq 1 - \frac{4\varepsilon^2}{\gamma^2}(N_b^2 - \sigma^2 + \delta). \tag{2.22}$$

Note that $-\sigma \leq C < C_\gamma \leq N_b$. Since the function

$$g(x) \equiv \frac{\sqrt{N_b^2 + C_\gamma(C_\gamma + x)} - N_b}{\sqrt{N_b^2 - x^2}} \tag{2.23}$$

is monotonically increasing on $(-C_\gamma, N_b)$ (see Supplement 1), the inequality (2.10) implies (2.11) and, therefore,

$$2N_b\varepsilon/\gamma < \frac{\sqrt{N_b^2 + C_\gamma^2} - N_b}{C_\gamma} = \frac{C_\gamma}{\sqrt{N_b^2 + C_\gamma^2} + N_b} < 1.$$

From (2.22) we know $z_p \neq 0$ if δ is small enough. Hence we have $z_p/\|z_p\| \in \Sigma_\gamma$ and

$$-\mathrm{Re}\langle Bz_p, z_p \rangle \geq C_\gamma \|z_p\|^2 \geq C_\gamma \left(1 - \frac{4\varepsilon^2}{\gamma^2}(N_b^2 - \sigma^2 + \delta)\right). \tag{2.24}$$

It follows from (2.14), (2.22), and (2.24) that

$$\begin{aligned}
\sigma &= \lim_{p \to +\infty} \mathrm{Re}\langle By_p, y_p \rangle \\
&\leq \sup_{p>N} [\mathrm{Re}\langle Bz_p, z_p \rangle + \mathrm{Re}\langle B(y_p - z_p), y_p \rangle + \mathrm{Re}\langle Bz_p, y_p - z_p \rangle] \\
&\leq \sup_{p>N} [-C_\gamma \|z_p\|^2 + (1 + \|z_p\|)\|B\|\|y_p - z_p\|] \\
&\leq -C_\gamma \left(1 - \frac{4\varepsilon^2}{\gamma^2}(N_b^2 - \sigma^2 + \delta)\right) + \frac{4N_b\varepsilon}{\gamma} \sqrt{N_b^2 - \sigma^2 + \delta}. \qquad (2.25)
\end{aligned}$$

We take $\delta \to 0$ in (2.25) to get

$$C_\gamma + \sigma \le 4C_\gamma \left(\frac{\varepsilon}{\gamma}\sqrt{N_b^2 - \sigma^2}\right)^2 + 4N_b \left(\frac{\varepsilon}{\gamma}\sqrt{N_b^2 - \sigma^2}\right). \tag{2.26}$$

If $\sigma = N_b$, then (2.26) is an obvious contradiction. For $-C \le \sigma < N_b$, (2.26) implies

$$\varepsilon \ge \frac{\gamma}{2C_\gamma} g(\sigma) \ge \frac{\gamma}{2C_\gamma} g(-C), \tag{2.27}$$

which contradicts (2.10).

Since $\mathcal{A}_\varepsilon$ also has a compact resolvent, from (2.12) we deduce that the resolvent of $\mathcal{A}_\varepsilon$ is bounded on $\varepsilon\sigma + \boldsymbol{i}\mathbb{R}$ for all $\sigma \ge -C$. By the resolvent equation, the resolvent is bounded on $\{\varepsilon\sigma + \boldsymbol{i}\tau \,|\, \sigma \ge -C - \delta, \tau \in \mathbb{R}\}$ for some $\delta > 0$ small enough. The proof is complete from (2.8). □

It is easy to see that (H3) is satisfied with $\gamma = \gamma_0$ when the following conditions hold.

(H4) The spectrum of A satisfies the gap condition

$$\inf\{|\beta_j - \beta_k| : j, k = 1, 2, \ldots, \beta_j \ne \beta_k\} \equiv \gamma_0 > 0. \tag{2.28}$$

(H5) For any normalized eigenvector ϕ of A,

$$-\mathrm{Re}\langle B\phi, \phi\rangle \ge C_0 > 0. \tag{2.29}$$

Corollary 2.2. *Assume that the conditions* (H1), (H2), (H4), *and* (H5) *hold. Then the semigroup $S_\varepsilon(t)$ is exponentially stable if*

$$0 < \varepsilon < \frac{\gamma_0}{2N_bC_0}\left(\sqrt{N_b^2 + C_0^2} - N_b\right). \tag{2.30}$$

Moreover, for every $C \in (0, C_0)$, it holds that

$$\omega_0(\mathcal{A}_\varepsilon) < -\varepsilon C \quad \forall\, 0 < \varepsilon < \frac{\gamma_0(\sqrt{N_b^2 + C_0(C_0 - C)} - N_b)}{2C_0\sqrt{N_b^2 - C^2}}. \tag{2.31}$$

Proof. This follows from Theorem 2.1 with $\gamma = \gamma_0$, $C_\gamma \ge C_0$, and the fact that the function

$$g_1(x) = \frac{\sqrt{N_b^2 + x(x - C)} - N_b}{x} = \left(\frac{N_b}{x - C} + \sqrt{\left(\frac{N_b}{x - C}\right)^2 + \frac{C}{x - C} + 1}\right)^{-1}$$

is monotonically increasing for $x > C \ge 0$. □

Remark 2.1. In the analysis above, we provided not only the sufficient conditions for the exponential stability of semigroup $S_\varepsilon(t)$ but also an explicit negative bound of the type $\omega_0(\mathcal{A}_\varepsilon)$ of semigroup $S_\varepsilon(t)$ with the perturbation parameter ε in an explicit range.

Remark 2.2. Chen et al. [2] discussed the exponential stability of (1.1) with $\varepsilon = 1$ and the dissipative operator B. In addition to assumptions similar to (H1)–(H3), they needed the condition that $\mathrm{Re}\langle By_p, y_p\rangle \to 0$ implies $By_p \to 0$ for any sequence of normalized vectors y_p. This condition does not hold generally if B is nondissipative. On the other hand, if $-B$ is symmetrical and nonnegative, i.e., $-B = -B^* \ge 0$, then

the assumption (H3) is necessary for the exponential stability of semigroup $S_\varepsilon(t)$ (see the proof of Theorem 3.1).

Remark 2.3. The spectral gap condition (H4) is very restrictive. Corollary 2.2 applies primarily to the one-dimensional problems. Actually, as we will see in section 4, the condition (H4) can be absent even for some PDE system on the region of one spatial dimension. Roughly speaking, (H5) means that the damping operator B is uniformly effective for all the normalized eigenvectors. When the spectral gap condition fails, (H3) means that the damping operator B is uniformly effective for all the normalized linear combinations of eigenvectors corresponding to the eigenvalues located in the γ-neighborhood of any eigenvalue.

3. Hautus-type criterion for exact controllability. Let $\mathcal{H}$ and U be Hilbert spaces. Consider the control system (A, G)

$$y(u,t) = e^{tA}y_0 + \int_0^t e^{(t-s)A}Gu(s)ds, \tag{3.1}$$

where A generates a C_0 semigroup e^{tA} on $\mathcal{H}$, $G \in \mathcal{L}(U;\mathcal{H})$, $y_0 \in \mathcal{H}$. When $\mathcal{H} = \mathbb{C}^n$ and $U = \mathbb{C}^m$ are finite dimensional, the famous Hautus lemma [6] says that the system (A, G) is controllable if and only if

$$\mathrm{Rank}[\lambda I - A,\ G] = n \qquad \forall\ \lambda \in \sigma(A), \tag{3.2}$$

or, equivalently,

$$\|(\lambda I - A)^* y\| + \|G^* y\|_U > 0 \qquad \forall\ \lambda \in \sigma(A),\ \|y\| = 1. \tag{3.3}$$

When $A^* = -A$, (3.3) is equivalent to

$$\|G^*\phi\|_U > 0 \qquad \forall\ \phi \text{ being normalized eigenvectors of } A. \tag{3.4}$$

In this section, we will give a counterpart of the special Hautus criterion (3.4) for infinite dimensional systems. We need the following Lemma given in Liu [8, Thm. 2.3]. Concerning the definitions of exact controllability and exponential stabilizability of (A, G), we refer the reader to [8].

LEMMA 3.1. *Let $A^* = -A$, $G \in \mathcal{L}(U;\mathcal{H})$. Then the following propositions are equivalent.*

(a) *The system (A, G) is exactly controllable.*
(b) *The system (A, G) is exponentially stabilizable.*
(c) *For every positive-definite self-adjoint $K \in \mathcal{L}(U)$ the operator $A - GKG^*$ generates an exponentially stable C_0 semigroup on $\mathcal{H}$.*

By Lemma 3.1 and the frequency domain condition for exponential stability [7, 10], Liu [8] gave a Hautus-type criterion for exact controllability of the second order conservative systems in Hilbert spaces. Also by Lemma 3.1, Zhou and Yamamoto [11] gave a counterpart of (3.3) for the conservative system (A, G), $A^* = -A$. Our result is the following.

THEOREM 3.2. *Suppose that the assumption* (H1) *holds and $G \in \mathcal{L}(U;\mathcal{H})$. Then the following propositions are equivalent.*

(a) *The system (A, G) is exactly controllable.*
(b) *The assumption* (H3) *holds for $B = -GG^*$; that is,*

$$\lim_{\gamma \to 0^+} \inf_{\psi \in \Sigma_\gamma} \|G^*\psi\|_U > 0. \tag{3.5}$$

(c) *There exists $F \in \mathcal{L}(\mathcal{H};U)$ such that the assumption* (H3) *holds for $B = GF$.*

Proof. The implication (b)⇒(c) is trivial; (c)⇒(a) follows readily from Theorem 2.1 and Lemma 3.1.

(a)⇒(b). By Lemma 3.1, the exact controllability of (A, G) is equivalent to the exponential stability of the semigroup $S_\varepsilon(t)$ with $B = -GG^*$, $\varepsilon > 0$. Thus it suffices to prove that (H3) is necessary for the exponential stability of $S_\varepsilon(t)$ when $-B = -B^* \geq 0$. In this case, $C_\gamma \geq 0$ for all $\gamma > 0$. If for any $\gamma > 0$, $C_\gamma = 0$, then there exist β_m and a normalized vector of the form

$$\psi_\gamma = \sum_{n \in I_{\gamma,m}} a_n \phi_n \tag{3.6}$$

such that

$$\|B\psi_\gamma\| \leq \|(-B)^{\frac{1}{2}}\| \langle -B\psi_\gamma, \psi_\gamma \rangle^{\frac{1}{2}} < \gamma.$$

Thus

$$\|(\boldsymbol{i}\beta_m I - \mathcal{A}_\varepsilon)\psi_\gamma\| \leq \varepsilon\gamma + \left\| \sum_{n \in I_{\gamma,m}} (\beta_m - \beta_n)\boldsymbol{i} a_n \phi_n \right\| \leq (1+\varepsilon)\gamma. \tag{3.7}$$

This means that the resolvent of $\mathcal{A}_\varepsilon$ is unbounded on $\boldsymbol{i}\mathbb{R}$ if it exists. Thus $S_\varepsilon(t)$ is not exponentially stable. □

Remark 3.1. If the spectral gap condition (H4) holds, then the condition (3.5) takes the form

$$\|G^*\phi\|_U \geq \delta > 0 \qquad \forall\ \phi \text{ being normalized eigenvectors of } A. \tag{3.8}$$

This is just a counterpart of the finite dimensional case (3.4).

4. Applications. In this section, we apply our result about exponential stability to the wave, beam, and two-dimensional Shrödinger equations with indefinite viscous damping or nondissipative perturbation arising from feedback by noncolocated observation and control.

Example 1. We consider the following one-dimensional wave equation with indefinite viscous damping:

$$\begin{cases} w_{tt}(x,t) = w_{xx}(x,t) - \varepsilon d(x) w_t(x,t), & 0 < x < 1, \quad t > 0, \\ w(0,t) = w_x(1,t) = 0, & t > 0, \\ w(0,t) = w_0(x), \quad w_t(x,0) = w_1(x), & 0 < x < 1, \end{cases} \tag{4.1}$$

where $d \in L^\infty(0,1)$ is real-valued. The underlying Hilbert space is

$$\mathcal{H} = \left\{ \begin{bmatrix} w \\ v \end{bmatrix} \in H^1(0,1) \times L^2(0,L) \Big| \ w(0) = 0 \right\}$$

with the inner product

$$\left\langle \begin{bmatrix} w_1 \\ v_1 \end{bmatrix}, \begin{bmatrix} w_2 \\ v_2 \end{bmatrix} \right\rangle = \int_0^1 [w_1' \bar{w}_2' + v_1 \bar{v}_2] \mathrm{d}x.$$

Define

$$\mathcal{D}(A) = \left\{ \begin{bmatrix} w \\ v \end{bmatrix} \Big| \ w \in H^2(0,1), v \in H^1(0,1), w(0) = v(0) = w'(1) = 0 \right\},$$

$$A = \begin{bmatrix} 0 & I \\ \partial_x^2 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 0 & -d(x) \end{bmatrix}, \quad y = \begin{bmatrix} w \\ w_t \end{bmatrix}.$$

Then the system (4.1) can be rewritten as (1.1). A has a complete orthonormal set of eigenfunctions

$$\phi_n^{\pm} = \frac{1}{(n+\frac{1}{2})\pi} \begin{bmatrix} \sin(n+\frac{1}{2})\pi x \\ \pm \boldsymbol{i}(n+\frac{1}{2})\pi \sin(n+\frac{1}{2})\pi x \end{bmatrix}$$

with eigenvalues

$$\boldsymbol{i}\beta_n^{\pm} = \pm \boldsymbol{i}\left(n+\frac{1}{2}\right)\pi, \qquad n = 0, 1, 2, \ldots.$$

It is easy to see that (H1), (H2), and (H4) are satisfied with $\gamma_0 = \pi$, and

$$-\langle B\phi_n^{\pm}, \phi_n^{\pm}\rangle = \int_0^1 d(x)\sin^2\left(n+\frac{1}{2}\right)\pi x \mathrm{d}x. \tag{4.2}$$

Thus (H5) holds if and only if

$$\inf_{n\geq 0}\int_0^1 d(x)[1-\cos(2n+1)\pi x]\mathrm{d}x > 0. \tag{4.3}$$

For example, we take

$$d(x) = 1 + \alpha\cos 2k\pi x, \quad \alpha \in \mathbb{R}, \tag{4.4}$$

with k being any positive integer. Then

$$\gamma_0 = \pi, \quad N_b = 1 + |\alpha|, \quad -\langle B\phi_n^{\pm}, \phi_n^{\pm}\rangle = \frac{1}{2} \quad \forall\, n \geq 0.$$

Therefore, $S_\varepsilon(t)$ is exponentially stable if

$$0 < \varepsilon < \frac{\pi}{2(1+|\alpha|)}\Big(\sqrt{4(1+|\alpha|)^2+1} - 2(1+|\alpha|)\Big). \tag{4.5}$$

Moreover, for every $C \in (0, \frac{1}{2})$, it holds that

$$\omega_0(\mathcal{A}_\varepsilon) < -\varepsilon C \qquad \forall\ 0 < \varepsilon < \frac{\pi(\sqrt{4(1+|\alpha|)^2+1-2C} - 2(1+|\alpha|))}{2\sqrt{(1+|\alpha|)^2 - C^2}}. \tag{4.6}$$

We can also choose $d(x)$ with local support, such as

$$d(x) = \begin{cases} \sin 4\pi x, & \frac{1}{4} \leq x \leq \frac{1}{2}, \\ 2\sin 4\pi x, & \frac{1}{2} \leq x \leq \frac{3}{4}, \\ 0, & \text{otherwise.} \end{cases} \tag{4.7}$$

For this case,

$$-\langle B\phi_n^{\pm}, \phi_n^{\pm}\rangle \geq \frac{3}{20\pi} \quad \forall\, n \geq 0.$$

Example 2. Let us consider the following system with distributed control and locally distributed observation:

$$\begin{cases} w_{tt}(x,t) + w_{xxxx}(x,t) = f(x,t), & 0 < x < 1, \quad t > 0, \\ w(0,t) = w(1,t) = w_{xx}(0,t) = w_{xx}(1,t) = 0, & t > 0, \\ w(0,t) = w_0(x), \ \ w_t(x,0) = w_1(x), & 0 < x < 1, \quad t > 0, \\ h(x,t) = w_t(x,t), & x \in (\frac{1}{2}, 1), \quad t > 0. \end{cases} \tag{4.8}$$

Our purpose is to find a bounded linear operator K from $L^2(\frac{1}{2},1)$ to $L^2(0,1)$ so that the feedback control law

$$f(\cdot,t) = -\varepsilon K h(\cdot,t), \qquad t>0, \tag{4.9}$$

stablizes the system (4.8) for some $\varepsilon>0$. The simplest form of operator K would be

$$[Kh](x) = \begin{cases} 2c(x)h(1-x), & 0<x<\frac{1}{2}, \\ 2d(x)h(x), & \frac{1}{2}<x<1, \end{cases} \tag{4.10}$$

where $c\in L^\infty(0,\frac{1}{2})$, $d\in L^\infty(\frac{1}{2},1)$ are real-valued. The underlying Hilbert space is

$$\mathcal{H} = [H^2(0,1)\cap H_0^1(0,1)]\times L^2(0,1)$$

with the inner product

$$\left\langle \begin{bmatrix} w_1 \\ v_1 \end{bmatrix}, \begin{bmatrix} w_2 \\ v_2 \end{bmatrix} \right\rangle = \int_0^1 [w_1''\bar{w}_2'' + v_1\bar{v}_2]\mathrm{d}x.$$

Define

$$\mathcal{D}(A) = \left\{ \begin{bmatrix} w \\ v \end{bmatrix} \,\middle|\, w\in H^4(0,1), v\in H^2(0,1), w,v\in H_0^1(0,1), w''(0)=w''(1)=0 \right\},$$

$$A = \begin{bmatrix} 0 & I \\ -\partial_x^4 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 0 & -KI_0 \end{bmatrix}, \quad y = \begin{bmatrix} w \\ w_t \end{bmatrix},$$

where I_0 is the embedding from $L^2(0,1)$ to $L^2(\frac{1}{2},1)$. Then the closed-loop system (4.8)–(4.9) can be rewritten as (1.1). A has a complete orthonormal set of eigenfunctions

$$\phi_n^\pm = \frac{1}{n^2\pi^2}\begin{bmatrix} \sin n\pi x \\ \pm \boldsymbol{i} n^2\pi^2 \sin n\pi x \end{bmatrix}$$

with eigenvalues

$$\boldsymbol{i}\beta_n^\pm = \pm \boldsymbol{i} n^2\pi^2, \qquad n=1,2,\ldots.$$

It is easy to see that (H1), (H2), and (H4) are satisfied with $\gamma_0 = 3\pi^2$. For K defined in (4.10), we have

$$\begin{aligned} -\langle B\phi_n^\pm, \phi_n^\pm\rangle &= 2\int_0^{\frac{1}{2}} c(x)\sin n\pi(1-x)\sin n\pi x\mathrm{d}x + 2\int_{\frac{1}{2}}^1 d(x)\sin^2 n\pi x\mathrm{d}x \\ &= \int_0^{\frac{1}{2}} [d(1-x)+(-1)^{n+1}c(x)](1-\cos 2n\pi x)\mathrm{d}x. \end{aligned} \tag{4.11}$$

Thus (H5) holds if and only if

$$\inf_{n\ge 1}\int_0^{\frac{1}{2}} [d(1-x)+(-1)^{n+1}c(x)](1-\cos 2n\pi x)\mathrm{d}x > 0. \tag{4.12}$$

For example, we take

$$c(x) = \cos 4\pi x, \quad d(x) = \frac{1}{2} + \cos 4\pi x. \tag{4.13}$$

Then

$$N_b = 3, \quad -\langle B\phi_n^{\pm}, \phi_n^{\pm}\rangle = \frac{1}{4} \quad \forall\, n \geq 1.$$

Therefore, $S_\varepsilon(t)$ is exponentially stable if

$$0 < \varepsilon < \frac{\pi^2}{2}\Big(\sqrt{145} - 12\Big). \tag{4.14}$$

Moreover, for every $C \in (0, \frac{1}{4})$, it holds that

$$\omega_0(\mathcal{A}_\varepsilon) < -\varepsilon C \qquad \forall\ 0 < \varepsilon < \frac{3\pi^2\Big(\sqrt{145 - 4C} - 12\Big)}{2\sqrt{9 - C^2}}. \tag{4.15}$$

Remark 4.1. Let $d \equiv 0$ in (4.10); then the observation and control are completely noncolocated. In this case, by (4.11) we know that K can be uniformly effective only for finite many eigenmodes. The problem of whether there exists $K \in \mathcal{L}(L^2(\frac{1}{2}, 1), L^2(0, 1))$ with $\mathrm{supp}Kh \subset (0, \frac{1}{2})$ for any $h \in L^2(\frac{1}{2}, 1)$ such that (H5) holds remains open.

Example 3. We consider the following two-dimensional Schrödinger equation with distributed control and locally distributed observation:

$$\begin{cases} \dfrac{\partial y}{\partial t}(x,t) = \boldsymbol{i}\Delta y(x,t) + f(x,t), \quad x \in \Omega = (0,a)\times(0,b), \quad t > 0, \\ y|_{\partial\Omega} = 0, \quad t > 0, \quad y(x,0) = y_0(x), \quad x \in \Omega, \\ h(x,t) = y(x,t), \qquad x \in \Omega_1 = (\frac{a}{2}, a)\times(0,b), \quad t > 0. \end{cases} \tag{4.16}$$

Let

$$\mathcal{H} = L^2(\Omega)$$

with the standard L^2 (complex) inner product. Define

$$A = \boldsymbol{i}\Delta, \quad \mathcal{D}(A) = H^2(\Omega) \cap H_0^1(\Omega). \tag{4.17}$$

Then the operator A is skew-adjoint and has eigenvalues

$$\lambda_{k,l} = \boldsymbol{i}\beta_{k,l} = \boldsymbol{i}\left(\frac{k^2}{a^2} + \frac{l^2}{b^2}\right)\pi^2, \quad k, l \in \mathbb{N},$$

and the corresponding normalized eigenfunctions

$$\phi_{k,l}(x) = \frac{2}{\sqrt{ab}} \sin\frac{k\pi x_1}{a} \sin\frac{l\pi x_2}{b}, \quad k, l \in \mathbb{N}.$$

Set the feedback control law

$$f(\cdot, t) = -\varepsilon K h(\cdot, t), \qquad K \in \mathcal{L}(L^2(\Omega_1), L^2(\Omega), \tag{4.18}$$

$$[Kh](x) = \begin{cases} c(x_1)h(a - x_1, x_2), & 0 < x_1 < \frac{a}{2},\ 0 < x_2 < b, \\ d(x_1)h(x), & \frac{a}{2} < x_1 < a,\ 0 < x_2 < b, \end{cases} \tag{4.19}$$

where $c \in L^\infty(0, \frac{a}{2})$, $d \in L^\infty(\frac{a}{2}, a)$ are real-valued. Then the closed-loop system (4.16)–(4.18) can be rewritten as (1.1) with $B = -KI_1$, where I_1 is the embedding from $L^2(\Omega)$ to $L^2(\Omega_1)$. When a/b is a rational number, the gap condition (H4) is true, but there are multiple eigenvalues. While a/b is an irrational number, the gap condition (H4) is false (see [2]). So, we have to check the condition (H3). We recount the eigenvalues and the corresponding normalized eigenfunctions:

$$\boldsymbol{i}\beta_n = \boldsymbol{i}\beta_{k_n,l_n}, \quad \beta_n \le \beta_{n+1}, \qquad \phi_n = \phi_{k_n,l_n}, \quad n \in \mathbb{N}.$$

Choose $\gamma = (\frac{\pi}{a})^2$; then

$$I_{\gamma,m} = \left\{ n \in \mathbb{N} \;\middle|\; \left| \frac{k_n^2 - k_m^2}{a^2} + \frac{l_n^2 - l_m^2}{b^2} \right| < \frac{1}{a^2} \right\}.$$

We note that for $p, q \in I_{\gamma,m}$, $p = q$ if and only if $l_p = l_q$ for any $\psi \in \Sigma_\gamma$,

$$\psi = \sum_{n \in I_{\gamma,m}} a_n \phi_n = \sum_{n \in I_{\gamma,m}} \frac{2a_n}{\sqrt{ab}} \sin\frac{k_n \pi x_1}{a} \sin\frac{l_n \pi x_2}{b}, \quad \sum_{n \in I_{\gamma,m}} |a_n|^2 = 1. \tag{4.20}$$

Using the orthogonality of $\{\sin(l\pi x_2/b)\}_{l=1}^\infty$ in $L^2(0,b)$, we have

$$\begin{aligned} -\langle B\psi, \psi\rangle &= \int_0^{\frac{a}{2}} c(x_1) \int_0^b \sum_{p,q \in I_{\gamma,m}} a_p \bar{a}_q \phi_p(a - x_1, x_2) \phi_q(x_1, x_2) \mathrm{d}x_2 \mathrm{d}x_1 \\ &\quad + \int_{\frac{a}{2}}^a d(x_1) \int_0^b \left| \sum_{n \in I_{\gamma,m}} \frac{2a_n}{\sqrt{ab}} \sin\frac{k_n \pi x_1}{a} \sin\frac{l_n \pi x_2}{b} \right|^2 \mathrm{d}x_2 \mathrm{d}x_1 \\ &= \frac{2}{a} \sum_{n \in I_{\gamma,m}} |a_n|^2 \int_0^{\frac{a}{2}} [d(a - x_1) + (-1)^{1+k_n} c(x_1)] \sin^2\frac{k_n \pi x_1}{a} \mathrm{d}x_1. \end{aligned} \tag{4.21}$$

Thus (H3) holds if and only if

$$\inf_{k \in \mathbb{N}} \int_0^{\frac{a}{2}} [d(a - x_1) + (-1)^{1+k} c(x_1)] \sin^2\frac{k\pi x_1}{a} \mathrm{d}x_1 > 0. \tag{4.22}$$

The sufficiency follows from (4.21). If (4.22) is false, then there exists a sequence ξ_n of positive integers such that

$$\lim_{n \to +\infty} \int_0^{\frac{a}{2}} [d(a - x_1) + (-1)^{1+\xi_n} c(x_1)] \sin^2\frac{\xi_n \pi x_1}{a} \mathrm{d}x_1 = \alpha \le 0.$$

Thus for the sequence $\phi_{\xi_n,1}$ of normalized eigenfunctions of A we have

$$-\langle B\phi_{\xi_n,1}\ ,\ \phi_{\xi_n,1}\rangle = \frac{2}{a} \int_0^{\frac{a}{2}} [d(a - x_1) + (-1)^{1+\xi_n} c(x_1)] \sin^2\frac{\xi_n \pi x_1}{a} \mathrm{d}x_1 \to \frac{2\alpha}{a} \le 0,$$

which implies that (H3) is false. For example, we take

$$c(x_1) = \cos\frac{4\pi x_1}{a}, \quad d(x_1) = \frac{1}{2} + \cos\frac{4\pi x_1}{a}. \tag{4.23}$$

Then

$$N_b = \frac{3}{2}, \qquad \gamma = \left(\frac{\pi}{a}\right)^2, \qquad C_\gamma = \frac{1}{4}.$$

Therefore, $S_\varepsilon(t)$ is exponentially stable if

$$0 < \varepsilon < \frac{\pi^2}{3a^2}\left(\sqrt{37} - 6\right). \tag{4.24}$$

Moreover, for every $C \in (0, \frac{1}{4})$, it holds that

$$\omega_0(\mathcal{A}_\varepsilon) < -\varepsilon C \qquad \forall\ 0 < \varepsilon < \frac{\pi^2\left(\sqrt{37 - 4C} - 6\right)}{a^2\sqrt{9 - 4C^2}}. \tag{4.25}$$

Example 4. We consider the following Timoshenko beam equation with indefinite viscous damping:

$$\begin{cases} u_{tt} = pu_{xx} - p\phi_x - \varepsilon d_1(x)u_t, & 0 < x < \pi,\ t > 0, \\ \phi_{tt} = q\phi_{xx} + pu_x - p\phi - \varepsilon d_2(x)\phi_t, & 0 < x < \pi,\ t > 0, \\ u(0,t) = u(\pi,t) = \phi_x(0,t) = \phi_x(\pi,t) = 0, & t > 0, \\ u(x,0) = u_0,\ u_t(x,0) = u_1,\ \phi(x,0) = \phi_0,\ \phi_t(x,0) = \phi_1, & 0 < x < \pi, \end{cases} \tag{4.26}$$

where $p,\ q > 0$ are constants and $d_1,\ d_2 \in L^\infty(0,\pi)$ are real-valued.

The underlying Hilbert space is

$$\mathcal{H} = H_0^1(0,\pi) \times L^2(0,\pi) \times H^1(0,\pi) \times L^2(0,\pi)$$

with the inner-product induced by the quadratic form of energy,

$$\|[u,v,\phi,\psi]^T\|^2 = \int_0^\pi \left(p|u_x - \phi|^2 + q|\phi_x|^2 + |v|^2 + |\psi|^2\right) \mathrm{d}x.$$

Define in $\mathcal{H}$

$$\mathcal{D}(A) = \{[u,v,\phi,\psi]^T \mid u, v, \in H_0^1(0,\pi),\ u, \phi \in H^2(0,\pi),\ \psi \in H^1(0,\pi)\},$$

$$A = \begin{bmatrix} 0 & I & 0 & 0 \\ p\partial_x^2 & 0 & -p\partial_x & 0 \\ 0 & 0 & I & 0 \\ p\partial_x & 0 & q\partial_x^2 - pI & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & -d_1(x) & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -d_2(x) \end{bmatrix}.$$

Then the system (4.26) can be rewritten as (1.1) by setting $y = [u, u_t, \phi, \phi_t]^T$. It is easy to verify that (H1) and (H2) are satisfied. To compute the eigenvalues of A, we solve the eigenequation

$$A[u,v,\phi,\psi]^T = \lambda[u,v,\phi,\psi]^T, \quad [u,v,\phi,\psi]^T \in \mathcal{D}(A).$$

Eliminating the unknowns v, ϕ, ψ, we obtain

$$\begin{cases} pqu_{xxxx} - (p+q)\lambda^2 u_{xx} + \lambda^2(\lambda^2 + p)u = 0, \\ u(0) = u(\pi) = u_{xx}(0) = u_{xx}(\pi) = 0. \end{cases}$$

A straightforward calculation leads to

$$\text{(4.27)}\qquad \begin{cases} \lambda_{n,1}^2 = -\frac{1}{2}[(p+q)n^2+p] + \frac{1}{2}\sqrt{[(p-q)n^2+p]^2+4pqn^2}, \\ \lambda_{n,2}^2 = -\frac{1}{2}[(p+q)n^2+p] - \frac{1}{2}\sqrt{[(p-q)n^2+p]^2+4pqn^2}, \end{cases}$$

$$\lambda_{n,i}^{\pm} = \pm \boldsymbol{i}\sqrt{-\lambda_{n,i}^2} \quad \text{(counting multiple eigenvalues)},$$

and the corresponding normalized eigenvectors

$$Z_{n,i}^{\pm} = \frac{1}{R_{n,i}}[\sin nx,\ \lambda_{n,i}^{\pm}\sin nx,\ S_{n,i}\cos nx,\ \lambda_{n,i}^{\pm}S_{n,i}\cos nx]^T, \qquad i=1,\ 2; \quad n\in\mathbb{N},$$

where

$$\begin{aligned} S_{n,i} &= \frac{1}{n}\left(n^2 + \frac{1}{p}\lambda_{n,i}^2\right) \quad (\neq 0) \\ &= \frac{1}{2pn}\left([(p-q)n^2-p] + (-1)^{1+i}\sqrt{[(p-q)n^2+p]^2+4pqn^2}\right), \end{aligned} \tag{4.28}$$

$$R_{n,i} = \sqrt{\frac{\pi}{2}}\left[p(n-S_{n,i})^2 + qn^2S_{n,i}^2 - \lambda_{n,i}^2 - \lambda_{n,i}^2S_{n,i}^2\right]^{\frac{1}{2}}. \tag{4.29}$$

These sequences have asymptotic expansions:

$$\text{(4.30)}\qquad \text{for } p=q, \qquad \begin{cases} \lambda_{n,j}^{\pm} = \pm\boldsymbol{i}\sqrt{p}(n + \frac{(-1)^j}{2} + O(\frac{1}{n})), & S_{n,j}^2 = 1 + O(\frac{1}{n}), \\ R_{n,j}^2 = 2\pi pn^2 + O(n), & j=1,\ 2, \end{cases}$$

$$\text{(4.31)}\qquad \text{for } p\neq q, \qquad \begin{cases} (i,j)=(1,2) \text{ for } p>q, & (i,j)=(2,1) \text{ for } p<q, \\ \lambda_{n,i}^2 = -qn^2 + O(1), & \lambda_{n,j}^2 = -pn^2 + O(1), \\ S_{n,i}^2 = (1-\frac{q}{p})^2n^2 + O(1), & S_{n,j}^2 = O(\frac{1}{n^2}), \\ R_{n,i}^2 = \pi q(1-\frac{q}{p})^2n^4 + O(n^2), & R_{n,j}^2 = \pi pn^2 + O(1). \end{cases}$$

In order to verify condition (H3), we need the following observations of the eigenvalues.

1. Each of the two branches of eigenvalues is distinct within itself. This can be verified by showing that $-\lambda_{n,i}^2, i=1,2$, are strictly monotonically increasing functions of n^2; see Supplement 2. Moreover, $\inf_{n\geq 1}|\lambda_{n+1,i}^{\pm} - \lambda_{n,i}^{\pm}| = \gamma_i > 0,\ i=1,2$, by the asymptotic expansions (4.30) and (4.31).
2. $\inf_{n\geq 1}|\lambda_{n,1}^{\pm} - \lambda_{n,2}^{\pm}| = \gamma_3 > 0$. This follows from the fact that $\lambda_{n,1}^2 - \lambda_{n,2}^2 > 0$ for all $n\geq 1$ and the asymptotic expansions (4.30) and (4.31).
3. Multiplicity of each eigenvalue is less than two. For any $1\leq m<n$, there exists $r=p/q$ such that $\lambda_{n,1}=\lambda_{m,2}$, i.e., double eigenvalues occur. This can be verified by showing that the function $f(r)=(\lambda_{n,1}^2-\lambda_{m,2}^2)/q$ changes sign on $(0,\infty)$ (see Supplement 3).
4. When $p\neq q$, the gap condition (H4) never holds. In fact, if $\sqrt{p/q}=k/l,\ k,l\in\mathbb{N}$, is a rational number, we have $\lambda_{kj,1}^2 - \lambda_{lj,2}^2 = O(1)$ for $p>q$ and $\lambda_{lj,1}^2 - \lambda_{kj,2}^2 = O(1)$ for $p<q$ as $j\to\infty$. See Supplement 4 for the case when $\sqrt{p/q}$ is an irrational number.

Choose $\gamma = \min\{\gamma_1, \gamma_2, \gamma_3, 2|\lambda_{1,1}^+|\}$. From the above observations, we know that

(4.32)
$$\Sigma_\gamma = \{c_1 Z_{n,1}^\pm + c_2 Z_{m,2}^\pm \mid |c_1|^2 + |c_2|^2 = 1,\ |\lambda_{n,1}^\pm - \lambda_{m,2}^\pm| < \gamma,\ m, n \in \mathbb{N},\ c_1, c_2 \in \mathbb{C}\}.$$

For any $c_1 Z_{n,1}^\pm + c_2 Z_{m,2}^\pm \in \Sigma_\gamma$, by $\gamma \le \gamma_3$ we have $n \neq m$, and

$$\begin{aligned}\Gamma(m, n, c_1, c_2) &\equiv -\langle B(c_1 Z_{n,1}^\pm + c_2 Z_{m,2}^\pm)\ ,\ c_1 Z_{n,1}^\pm + c_2 Z_{m,2}^\pm\rangle \\ &= \int_0^\pi d_1(x) \left| \frac{c_1 \lambda_{n,1}^+}{R_{n,1}} \sin nx + \frac{c_2 \lambda_{m,2}^+}{R_{m,2}} \sin mx \right|^2 \mathrm{d}x \\ &\quad + \int_0^\pi d_2(x) \left| \frac{c_1 \lambda_{n,1}^+ S_{n,1}}{R_{n,1}} \cos nx + \frac{c_2 \lambda_{m,2}^+ S_{m,2}}{R_{m,2}} \cos mx \right|^2 \mathrm{d}x. \qquad (4.33)\end{aligned}$$

Thus (H3) holds if and only if

$$\inf\{\Gamma(m, n, c_1, c_2) \mid m \neq n, |c_1|^2 + |c_2|^2 = 1, m, n \in \mathbb{N}, c_1, c_2 \in \mathbb{C}\} > 0. \qquad (4.34)$$

For example, we take

(4.35)
$$d_i(x) = \alpha_i(1 + \beta_i \cos k_i x), \quad \alpha_i \ge 0,\ \alpha_1^2 + \alpha_2^2 \neq 0,\ |\beta_i| < 2,\ k_i \in \mathbb{N},\ \ i = 1,\ 2.$$

Using

$$\left| \int_0^\pi \cos k_2 x \cos nx \cos mx \mathrm{d}x \right| \le \frac{\pi}{4},$$

$$\left| \int_0^\pi \cos k_1 x \sin nx \sin mx \mathrm{d}x \right| \begin{cases} = 0, & k_1 = 2n \text{ or } 2m, \\ \le \frac{\pi}{4}, & \text{otherwise}, \end{cases}$$

we deduce that

$$\begin{aligned}\Gamma(m, n, c_1, c_2) &\ge \frac{\pi}{2}\alpha_1 \left(1 - \frac{|\beta_1|}{2}\right) \left(\frac{-\lambda_{n,1}^2}{R_{n,1}^2}|c_1|^2 + \frac{-\lambda_{m,2}^2}{R_{m,2}^2}|c_2|^2 \right) \\ &\quad + \frac{\pi}{2}\alpha_2 \left(1 - \frac{|\beta_2|}{2}\right) \left(\frac{-\lambda_{n,1}^2 S_{n,1}^2}{R_{n,1}^2}|c_1|^2 + \frac{-\lambda_{m,2}^2 S_{m,2}^2}{R_{m,2}^2}|c_2|^2 \right) \\ &\ge \frac{\pi}{2} \min(\eta_1, \eta_2),\end{aligned}$$

where we have put

$$\eta_1 = \inf_{n \ge 1} \left(\alpha_1 \left(1 - \frac{|\beta_1|}{2}\right) \frac{-\lambda_{n,1}^2}{R_{n,1}^2} + \alpha_2 \left(1 - \frac{|\beta_2|}{2}\right) \frac{-\lambda_{n,1}^2 S_{n,1}^2}{R_{n,1}^2} \right),$$

$$\eta_2 = \inf_{m \ge 1} \left(\alpha_1 \left(1 - \frac{|\beta_1|}{2}\right) \frac{-\lambda_{m,2}^2}{R_{m,2}^2} + \alpha_2 \left(1 - \frac{|\beta_2|}{2}\right) \frac{-\lambda_{m,2}^2 S_{m,2}^2}{R_{m,2}^2} \right).$$

Now, using the asymptotic expansions (4.30) and (4.31), we can easily conclude the following.

1. When $p = q$, (H3) holds even if either α_1 or α_2 is zero. Therefore, by Theorem 2.1, $S_\varepsilon(t)$ is exponentially stable if ε is small enough. This means that when the two wave speeds are the same, only one displacement or rotation angle damping is sufficient for the exponential energy decay in the Timoshenko beam.
2. When $p \neq q$, (H3) holds if both α_1 and α_2 are positive. Therefore, $S_\varepsilon(t)$ is exponentially stable if ε is small enough. On the other hand, for $d_1(x) \equiv 0$ or $d_2(x) \equiv 0$, it is easy to see that $S_\varepsilon(t)$ is not exponentially stable for any $\varepsilon > 0$. This means that when the two wave speeds are different, the displacement and rotation angle dampings are necessary for the exponential energy decay in the Timoshenko beam.

5. Technical supplements.

SUPPLEMENT 1. *Let* $0 < C_\gamma \leq N_b$, *and*

$$g(x) = \frac{\sqrt{N_b^2 + C_\gamma(C_\gamma + x)} - N_b}{\sqrt{N_b^2 - x^2}}.$$

Then g *is a monotonically increasing function on* $(-C_\gamma, N_b)$.

Proof. Write the function g in the following form:

$$g(x) = \frac{C_\gamma}{\sqrt{N_b - x}}\sqrt{1 - \frac{N_b - C_\gamma}{N_b + x}}\left(\sqrt{\frac{N_b^2}{C_\gamma + x} + C_\gamma} + \frac{N_b}{\sqrt{C_\gamma + x}}\right)^{-1}.$$

Then the monotony of g follows from the fact that all the factors are positive monotonically increasing functions on $(-C_\gamma, N_b)$. □

SUPPLEMENT 2. $-\lambda_{n,i}^2$, $i = 1, 2$, *are monotonic increasing functions of* n^2.

Proof. It is obvious that the conclusion holds for $-\lambda_{n,2}^2$. Let

$$\begin{aligned} f(x) &= (p+q)x + p - \sqrt{[(p-q)x + p]^2 + 4pqx} \\ &= (p+q)x + p - \sqrt{[(p+q)x + p]^2 - 4pqx^2}. \end{aligned}$$

Then $2f(n^2) = -\lambda_{n,2}^2$. Since

$$\begin{aligned} f'(x) &= (p+q) - \frac{[(p+q)x + p](p+q) - 4pqx}{\sqrt{[(p+q)x + p]^2 - 4pqx^2}} \\ &= (p+q)\left[1 - \frac{(p+q)x + p - 4pq(p+q)^{-1}x}{\sqrt{[(p+q)x + p]^2 - 4pqx^2}}\right] \end{aligned}$$

and the fraction in the bracket is strictly less than one, we know that $f'(x) > 0$ for all $x > 0$. □

SUPPLEMENT 3. *For any* $1 \leq m < n$, *there exist* $p, q > 0$ *such that* $\lambda_{n,1}^2 = \lambda_{m,2}^2$.

Proof. Let $r = p/q$, and

$$\begin{aligned} f(r) &= \frac{1}{q}(\lambda_{n,1}^2 - \lambda_{m,2}^2) \\ &= (r+1)(m^2 - n^2) + \sqrt{[(r-1)n^2 + r]^2 + 4rn^2} + \sqrt{[(r-1)m^2 + r]^2 + 4rm^2}, \end{aligned}$$

which is continuous for $r \in (0, +\infty)$. Moreover, we have

$$\lim_{r\to 0^+} f(r) = m^2 - n^2 < 0, \qquad \lim_{r\to+\infty} f(r) = +\infty. \tag{5.1}$$

This proves that for a certain ratio of p and q, there exist double eigenvalues. □

SUPPLEMENT 4. *Let $\sqrt{p/q}$ be an irrational number; then there exist sequences of integers $k_j \to +\infty$, $l_j \to +\infty$, such that $qk_j^2 - pl_j^2 = O(1)$.*

Proof. By a result in number theory [5, p. 140], for any $j \geq 1$, there exists a rational number $\frac{k_j}{l_j}$ with $l_j > j$ such that

$$\left|\sqrt{\frac{p}{q}} - \frac{k_j}{l_j}\right| \leq \frac{1}{l_j^2}, \quad \text{i.e.,} \quad \left|l_j\sqrt{\frac{p}{q}} - k_j\right| \leq \frac{1}{l_j}.$$

Thus

$$\begin{aligned} |qk_j^2 - pl_j^2| &= q\left(l_j\sqrt{\frac{p}{q}} + k_j\right)\left|l_j\sqrt{\frac{p}{q}} - k_j\right| \\ &\leq ql_j\left(2\sqrt{\frac{p}{q}} + 1\right)\frac{1}{l_j} = 2\sqrt{pq} + q. \end{aligned}$$ □

REFERENCES

[1] A. BENADDI AND B. RAO, *Energy decay rate of wave equations with indefinite damping*, J. Differential Equations, 161 (2000), pp. 337–357.

[2] G. CHEN, S. A. FULLING, F. J. NARCOWICH, AND S. SUN, *Exponential decay of energy of evolution equations with locally distributed damping*, SIAM J. Appl. Math., 51 (1991), pp. 266–301.

[3] P. FREITAS, *On some eigenvalue problems related to the wave equation with indefinite damping*, J. Differential Equations, 127 (1996), pp. 320–335.

[4] P. FREITAS AND E. ZUAZUA, *Stability results for the wave equation with indefinite damping*, J. Differential Equations, 132 (1996), pp. 338–352.

[5] G. H. HARDY AND E. M. WRIGHT, *An Introduction to the Theory of Numbers*, 5th ed., Clarendon Press, Oxford, UK, 1979.

[6] M. L. J. HAUTUS, *Controllability and observability conditions for linear autonomous systems*, Indag. Math. (N.S.), 72 (1969), pp. 443–448.

[7] F. L. HUANG, *Characteristic condition for exponential stability of linear dynamical systems in Hilbert spaces*, Ann. Differential Equations, 1 (1985), pp. 43–56.

[8] K. LIU, *Locally distributed control and damping for the conservative systems*, SIAM J. Control Optim., 35 (1997), pp. 1574–1590.

[9] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer, New York, 1983.

[10] J. PRÜSS, *On the spectrum of C_0-semigroups*, Trans. Amer. Math. Soc., 284 (1984), pp. 847–857.

[11] Q. ZHOU AND M. YAMAMOTO, *Hautus condition on the exact controllability of conservative systems*, Internat. J. Control, 67 (1997), pp. 371–379.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 1, pp. 166–190

SERIES EXPANSIONS FOR THE EVOLUTION OF MECHANICAL CONTROL SYSTEMS*

FRANCESCO BULLO†

Abstract. This paper presents a series expansion that describes the evolution of a mechanical system starting at rest and subject to a time-varying external force. Mechanical systems are presented as second-order systems on a configuration manifold via the notion of affine connections. The series expansion is derived by exploiting the homogeneity property of mechanical systems and the variations of constant formula. A convergence analysis is obtained using some analytic functions and combinatorial analysis results. This expansion provides a rigorous means of analyzing locomotion gaits in robotics and lays the foundation for the design of motion control algorithms for a large class of underactuated mechanical systems.

Key words. series expansions, control of mechanical systems, nonlinear controllability

AMS subject classifications. 30B99, 70Q05, 93B05, 93B29

PII. S0363012999364796

1. Introduction.

The general purpose of this work is to develop an innovative and powerful control and analysis method for underactuated mechanical control systems. This paper introduces a series expansion that characterizes the evolution of a mechanical system starting at rest and subject to an open loop time-varying force. This tool should prove useful in the study of robotic locomotion and in the design of motion control algorithms.

1.1. Series expansions and their control applications.

Original works on perturbation methods and series expansions in mechanics go back to Poincaré and Lagrange. Magnus [33] describes the evolution of systems on a Lie group. Chen [14], Fliess [17], and Sussmann [42] develop a general framework to describe the evolution of a nonlinear system via the so-called Chen–Fliess series and its factorization. Related work on the "chronological calculus" formalism was developed by Agrachev and Gamkrelidze [2].

Within the context of modern geometric control theory, series expansions play a key role in the study of nonlinear controllability. Small-time local controllability (STLC) was studied, for example, by Sussmann [41, 43], Agrachev and Gamkrelidze [3], and Kawski [22, 24]. Controllability along trajectories was investigated by Bianchini and Stefani in [8]. Finally, the work by Lewis and Murray [32] on configuration controllability for mechanical control systems is much related to this work.

Motion planning problems provide a second important use of series expansions. A rich literature is available on the motion planning problem for kinematic systems, that is, systems without drift. Numerous approaches include algorithms for chained systems by Murray and Sastry [36], for systems on Lie groups by Leonard and Krishnaprasad [30] and Kolmanovsky and McClamroch [26], and the very general solution

*Received by the editors December 4, 1999; accepted for publication (in revised form) January 11, 2001; published electronically May 31, 2001. A preliminary short version of this work appeared in the Proceedings of the IFAC World Congress, Vol. E, Beijing, China, 1999, pp. 479–485. This research was supported by the Campus Research Board at the University of Illinois at Urbana-Champaign.
http://www.siam.org/journals/sicon/40-1/36479.html

†Coordinated Science Laboratory and General Engineering Department, University of Illinois at Urbana-Champaign, 1308 W. Main St, Urbana, IL 61801 (bullo@uiuc.edu, http://motion.csl.uiuc.edu).

proposed by Lafferriere and Sussmann [28]. These works rely on the following observation: an explicit expression for the "input history to final displacement" map simplifies dramatically the two-point boundary value problem that defines the motion planning task. In other words, whenever an explicit expression (provided by a series expansion) for the evolution of the control system is available, the two-point boundary value problem is reduced to a low dimensional nonlinear program. Accordingly, motion control algorithms are designed by inverting this "input history to final displacement" map.

Finally, series expansions and the techniques developed in this paper have potential relevance in several areas including averaging and vibrational stabilization [6, 10], high-order variations for use in optimal control [25], digital multirate sampling of nonlinear systems [18], and model reduction [19].

Series expansions that specifically exploit the structure of mechanical systems have so far not been computed. However, some preliminary progress in this direction has been obtained by Bullo, Leonard, and Lewis [12, 13] via a perturbation analysis. Under the assumption of small amplitude forcing, the authors compute the initial terms of a Taylor series describing the forced evolution. The results are then found to be in agreement with the controllability analysis in [32]. A different but related research direction has focused on open loop vibrational control and the recent progress we described in [10] is related to this paper. A preliminary short version of this work appeared in [11].

1.2. Summary of results. The main contribution of this paper is a series that describes the evolution of a forced mechanical system starting from rest. Mechanical systems are characterized as second-order systems on a configuration manifold using the theory of affine connections. By exploiting the problem's structure, the system's evolution is described as a flow on the configuration space (n-dimensional) instead of a series on the full phase space ($2n$-dimensional).

The treatment relies on some differential geometric tools to describe the homogeneity properties of nonlinear mechanical systems and the variations of constants formula; see [2] and [23]. The homogeneous structure of nonlinear mechanical systems leads to a recursive procedure to compute the forced solution to a mechanical system. The terms in the series are computed recursively via time integrals and certain Lie brackets called symmetric products [32].

The series is guaranteed to convergence in a strong sense for small amplitude inputs and bounded final time. The convergence analysis is sophisticated and relies on various concepts from complex and combinatorial analysis. Following the analysis by Agrachev and Gamkrelidze in [2, Proposition 2.1], a bound is computed for every term of the series so that a notion of order is established. However, as opposed to [2], only a recursive expression for the series terms is available, and this much complicates the treatment. The key idea is to obtain a recursive bound not only on the terms of the expansion but also on their partial derivative.

The series expansion can be computed in simplified fashion in two settings. For simple Hamiltonian systems with integrable forces, the main theorem can be interpreted as a statement on gradient and Hamilton flows: the flow of a Hamiltonian system forced from rest can be written as a (time-varying) gradient flow. For invariant systems on groups, the series can be computed via algebraic manipulations (no differentiations). In other words, the computations are performed on the corresponding Lie algebra, and the theorem reduces to a statement on the flow of polynomial control systems. These results agree with and supersede the preliminary results in

in [12, 13].

Finally, some numerical simulations of a three degree of freedom robotic manipulator are performed. Truncating the series expansion at increasingly higher order, various approximations are obtained, and their accuracy is illustrated via some numerical data.

1.3. Organization. The paper is organized as follows. In section 2 we present the model and the homogeneity properties of a large class of mechanical control systems. Most ideas are common in the literature; some are not. In section 3 we present the main result of the paper, that is, a convergent series describing the evolution of a forced mechanical system. Section 4 contains some applications and extensions, including the simple Hamiltonian and the invariant system settings, as well as some simulations. We present our conclusions in section 5.

2. Some geometric and analytical properties of mechanical systems. We present a geometric definition of mechanical control systems, study their homogeneous properties, and provide bounds using analytic function theory.

2.1. Natural objects on manifolds. We review some basic definitions to fix some notation; see [1]. All the objects we consider are smooth in the sense of analytic. Let Q be a finite dimensional, Hausdorff, second countable manifold, let q be a point on it, let v_q be a point on TQ, let $I \subset \mathbb{R}$ be a real interval, and let $\gamma : I \to Q$ be a curve on Q. We let 0_q denote the zero velocity tangent vector on the tangent space T_qQ. Let $\pi : TQ \to Q$ denote the usual projection on the tangent bundle, that is, $\pi(v_q) = q$. On the manifold Q, we will define scalar functions $q \mapsto f(q) \in \mathbb{R}$ and vector fields $q \mapsto X(q) \in T_qQ$. Lie derivatives of functions and Lie brackets of vector fields are denoted by

$$\mathcal{L}_X f \qquad \text{and} \qquad \mathcal{L}_X Y = [X, Y].$$

2.2. Variation of constants formula in geometric terms. This section presents a quick review of the variation of constants formula within the chronological calculus formalism introduced in [2]; see also [38]. Given a vector field Y and a diffeomorphism ϕ, the *pull-back of* Y *along* ϕ, denoted ϕ^*Y, is a vector field defined by

$$(\phi^*Y)(q) \triangleq T_q\phi^{-1} \circ Y \circ \phi,$$

where $T_q\phi^{-1}$ is the tangent map to ϕ^{-1}; see [1]. In a system of local coordinates $(q^1, \ldots, q^n)$, a vector field is written as $Y(q) = Y^i(q)\ \partial/\partial q^i$, and the pull-back of Y along ϕ is

$$(\phi^*Y)^i(q) = \frac{\partial(\phi^{-1})^i}{\partial q^j} Y^j(\phi(q)),$$

where the summation convention is enforced here and in what follows.

A time-varying vector field $(q, t) \mapsto X(q, t)$ gives rise to the initial value problem

$$\dot{q}(t) = X(q, t), \qquad q(0) = q_0,$$

and its solution at time T, which we refer to as the flow of X, is denoted by $q(T) = \Phi^X_{0,T}(q_0)$. We shall usually assume time-varying quantities to be integrable with respect to time. Given a time-varying vector field $X(q, t)$, we denote its definite time

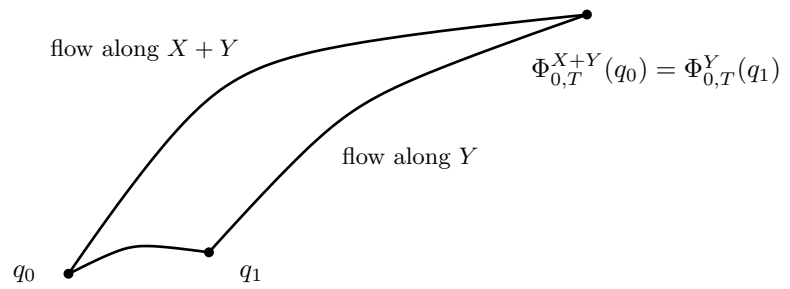


FIG. 2.1. *The flow along $X+Y$ is written as the flow along Y with initial condition q_q. The "variation" q_1 is computed via the variation of constants formula as the flow along $(\Phi_{0,t}^{Y})^* X$ for time $[0,T]$ with initial condition q_0.*

integral from time 0 to time T by

$$\overline{X}(q,T) = \int_0^T X(q,\tau)d\tau. \tag{2.1}$$

The integral takes place over the linear space T_qQ at fixed $q \in Q$. This operation can be defined in two ways. Given a coordinates chart about q, the integral is well defined in the coordinate system. (This definition suffices for the purpose of this paper, since the analysis is local.) A global coordinate-free definition is obtained, providing sufficient conditions in order for T_qQ to be a Banach space and introducing the Cauchy–Bochner integral; see [1, p. 61].

Next, consider the initial value problem

$$\dot{q}(t) = X(q,t) + Y(q,t), \qquad q(0) = q_0, \tag{2.2}$$

where X and Y are analytic time-varying vector fields. If we regard X as a perturbation to the vector field Y, we can describe the flow of $X+Y$ in terms of a nominal and perturbed flow. The following relationship is referred to as the *variation of constants* formula and describes the perturbed flow:

$$\Phi_{0,t}^{X+Y} = \Phi_{0,t}^{Y} \circ \Phi_{0,t}^{(\Phi_{0,t}^{Y})^* X}. \tag{2.3}$$

The result is illustrated in Figure 2.1 and proven in [2, equation (3.15)]; see also [10, Appendix A.1]. The result can be alternatively stated as follows. For all $T \geq 0$, the final value $q(T)$ of the curve $q : [0,T] \to M$ solution to the initial value problem (2.2) is also the final value of the curve solution to

$$\dot{q}(s) = Y(q,s), \qquad q(0) = z(T), \tag{2.4}$$

where $z : [0,T] \to M$ is the solution to the initial value problem

$$\dot{z}(s) = \left((\Phi_{0,s}^{Y})^* X\right)(z), \qquad z(0) = q_0. \tag{2.5}$$

The differential equation (2.5) is referred to as the "pulled back" or the "adjoint" system in [20]. If both X and Y are time invariant, then the classic infinitesimal Campbell–Backer–Hausdorff formula (see [21]) provides a means of computing the

pull-back:

$$(\Phi^Y_{0,t})^* X = \sum_{k=0}^{\infty} \mathrm{ad}^k_Y X \, \frac{t^k}{k!}.$$

If instead X and Y are time-varying, a generalized expression is (see [2])

$$(\Phi^Y_{0,t})^* X(q,t) = X(q,t) + \sum_{k=1}^{\infty} \int_0^t \cdots \int_0^{s_{k-1}} \left(\mathrm{ad}_{Y(q,s_k)} \ldots \mathrm{ad}_{Y(q,s_1)} X(q,t)\right) \, ds_k \ldots ds_1. \tag{2.6}$$

Just like in the classic Campbell–Backer–Hausdorff formula (see [44]), the convergence of the series expansion in the previous equation is a delicate manner. Sufficient conditions for local convergence are given in [2, Propositions 2.1 and 3.1]. For our analysis, the following straightforward statement suffices. If all the Lie brackets $\mathrm{ad}_{Y(s_k)} \ldots \mathrm{ad}_{Y(s_1)} X$ vanish for all k greater than a given N, then the series in (2.6) becomes a finite sum, and it readily converges.

2.3. Affine connections. We refer to [16, 29] for a comprehensive treatment on affine connections and Riemannian geometry. An *affine connection* on Q is a smooth map that assigns to a pair of vector fields X, Y a vector field $\nabla_X Y$ such that for any function f and for any third vector field Z

(i) $\nabla_{fX+Y} Z = f \nabla_X Z + \nabla_Y Z$,
(ii) $\nabla_X (fY + Z) = (\mathcal{L}_X f) Y + f \nabla_X Y + \nabla_X Z$.

We also say that $\nabla_X Y$ is the covariant derivative of Y with respect to X. Vector fields can also be covariantly differentiated along curves, and this concept will be instrumental in writing the Euler–Lagrange equations. Consider a smooth curve $\gamma : [0,1] \to Q$ and a vector field along γ, that is, a map $v : [0,1] \to TQ$ such that $\pi(v(t)) = \gamma(t)$ for all $t \in [0,1]$. Let V be a smooth vector field satisfying $V(\gamma(t)) = v(t)$. The covariant derivative of the vector field v along γ is defined by

$$\frac{Dv(t)}{dt} \triangleq \nabla_{\dot\gamma(t)} v(t) = \nabla_{\dot\gamma(t)} V(q)\big|_{q=\gamma(t)}.$$

It can be shown that this definition is independent of the choice of V. In a system of local coordinates $(q^1, \ldots, q^n)$, an affine connection is uniquely determined by its Christoffel symbols[1] Γ^i_{ij},

$$\nabla_{\frac{\partial}{\partial q^i}} \left(\frac{\partial}{\partial q^j} \right) = \Gamma^k_{ij} \frac{\partial}{\partial q^k},$$

and, accordingly, the covariant derivative of a vector field is written as

$$\nabla_X Y = \left(\frac{\partial Y^i}{\partial q^j} X^j + \Gamma^i_{jk} X^j Y^k \right) \frac{\partial}{\partial q^i}.$$

[1] We here refer to the Γ^k_{ij} functions as Christoffel symbols, even without requiring ∇ to be a Levi–Civita connection.

2.4. Control systems described by affine connections. We introduce a class of control systems that is a generalization of Lagrangian control systems. This approach to the modeling of vehicles and robotic manipulators is common to a number of recent works; see [9, 32, 31, 10]. A *control system described by an affine connection* is defined by the following objects:

(i) an n-dimensional configuration manifold Q, with $q \in Q$ being the configuration of the system and $v_q \in T_qQ$ being the system's velocity,

(ii) an affine connection ∇ on Q, whose Christoffel symbols are $\{\Gamma^i_{jk} : i, j, k \in \{1, \dots, n\}\}$,

(iii) a time-varying vector field Y on Q defining the input force.

The corresponding equations of motion are written as

$$\frac{Dv_q}{dt} = Y(q, t) \tag{2.7}$$

or, equivalently, in coordinates as

$$\dot{q}^i = v^i, \qquad \dot{v}^i + \Gamma^i_{jk}(q) v^j v^k = Y^i(q, t), \tag{2.8}$$

where the indices i, j, k run from 1 to n and where $v_q = v^i \frac{\partial}{\partial q^i}$. These equations are a generalized form of the Euler–Lagrange equations.

Remark 2.1. This definition of control systems described by an affine connection provides a convenient means of treating various classes of Lagrangian mechanical systems. For example, systems with nonholonomic constraints are described within this framework in [31]. We will treat in more details "simple Hamiltonian systems" in section 4.2 and "invariant systems on Lie groups" in section 4.3. A more detailed exposition is presented in [10].

The second-order system in (2.7) can be written as a first-order differential equation on the tangent bundle TQ. Using $\{\frac{\partial}{\partial q^i}, \frac{\partial}{\partial v^i}\}$ as a basis for the tangent bundle to TQ, we define

$$Z(v_q) = v^i \frac{\partial}{\partial q^i} - \Gamma(q)^i_{jk} v^j v^k \frac{\partial}{\partial v^i} \quad \text{and} \quad Y^{\text{lift}}(v_q, t) = Y^i(q, t) \frac{\partial}{\partial v^i},$$

so that the control system is rewritten as

$$\dot{v}_q = Z(v_q) + Y^{\text{lift}}(v_q, t). \tag{2.9}$$

We refer to [32, 29] for coordinate independent definitions of the lifting operation $Y \to Y^{\text{lift}}$ and of the drift vector field Z.

2.5. Homogeneity and Lie algebraic structure. One fundamental structure of the control system in (2.7) is the polynomial dependence of the vector fields Z and Y^{lift} on the velocity variables v^i_q. This structure is reflected in the Lie brackets computations involving Z and Y^{lift}; see related ideas in [40, 10].

We here rely on the notion of *geometric homogeneity*[2] as described in [23]. Given two vector fields X and X_E, we say that the vector field X is homogeneous with degree m with respect to X_E if

$$[X_E, X] = mX.$$

[2]Geometric homogeneity corresponds to the existence of an (infinitesimal) symmetry in the equations of motion. For control systems described by an affine connection the symmetry is invariance under affine time-scaling transformations.

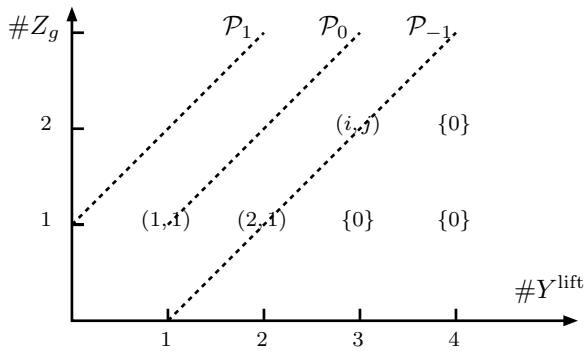


FIG. 2.2. *Table of Lie brackets between the drift vector field Z and the input vector field Y^{lift}. The (i,j)th position contains Lie brackets with i copies of Y^{lift} and j copies of Z. The corresponding homogeneous degree is $j-i$. All Lie brackets to the right of $\mathcal{P}_{-1}$ exactly vanish. All Lie brackets to the left of $\mathcal{P}_{-1}$ vanish when evaluated at $v_q = 0_q$.*

For control systems described by an affine connection, we introduce the Liouville vector field on TQ (see [7, pp. 19 and 29]) as

$$L(v_q) = v^i \frac{\partial}{\partial v^i},$$

where we recall $v_q = v^i \frac{\partial}{\partial q^i}$. The key mathematical relationships between vector fields on TQ are

$$[L, Z] = (+1)Z \quad \text{and} \quad [L, Y^{\text{lift}}] = (-1)Y^{\text{lift}}.$$

Hence the vector field Z is homogeneous of degree $+1$, and the vector field Y^{lift} is homogeneous of degree -1 with respect to the Liouville vector field. Let $\mathcal{P}_j$ be the set of vector fields on TQ of homogeneous degree j, so that

$$Z \in \mathcal{P}_1 \quad \text{and} \quad Y^{\text{lift}} \in \mathcal{P}_{-1}.$$

The sets $\mathcal{P}_j$ enjoy various interesting properties. Figure 2.2 illustrates them, their proof is via direct computation, and they are listed as follows:

(i) $[\mathcal{P}_i, \mathcal{P}_j] \subset \mathcal{P}_{i+j}$, that is, the Lie bracket between a vector field in $\mathcal{P}_i$ and a vector field in $\mathcal{P}_j$ belongs to $\mathcal{P}_{i+j}$.
(ii) $\mathcal{P}_k = \{0\}$ for all $k \leq -2$.
(iii) For all $X \in \mathcal{P}_k$ with $k \geq 1$, $X(0_q) = 0_q$.
(iv) Every $X \in \mathcal{P}_{-1}$ is the lift of a vector field on Q.

It is helpful to provide an interpretation of $\mathcal{P}_i$ in coordinates. In a system of local coordinates, let $\mathcal{H}_i(q, v_q)$ be the set of scalar functions on $TQ = \mathbb{R}^{2n}$, which are arbitrary functions of q and which are homogeneous polynomials in $\{v^1, \ldots, v^n\}$ of degree i. $\mathcal{P}_i$ is the set of vector fields on $\mathbb{R}^{2n}$ with the first n components in $\mathcal{H}_i$ and the second n components in $\mathcal{H}_{i+1}$.

Finally, it is of interest to focus on the Lie bracket $[Y_b^{\text{lift}}, [Z, Y_a^{\text{lift}}]]$, where Y_a, Y_b are two vector fields on Q. This operation will play an important role in later computations. Since this Lie bracket belongs to $\mathcal{P}_{-1}$, there must exist a vector field on Q, which we denote $\langle Y_a : Y_b \rangle$, such that

$$\langle Y_a : Y_b \rangle^{\text{lift}} = [Y_b^{\text{lift}}, [Z, Y_a^{\text{lift}}]].$$

Such a vector field is called the *symmetric product* between Y_b and Y_a, and a direct computation shows that it satisfies

$$\langle Y_b : Y_a \rangle = \nabla_{Y_a} Y_b + \nabla_{Y_b} Y_a,$$

or, equivalently, in coordinates

$$\langle Y_b : Y_a \rangle^i = \frac{\partial Y_a^i}{\partial q^j} Y_b^j + \frac{\partial Y_b^i}{\partial q^j} Y_a^j + \Gamma_{jk}^i \left(Y_a^j Y_b^k + Y_a^k Y_b^j \right).$$

The adjective "symmetric" comes from the equality $\langle Y_a : Y_b \rangle = \langle Y_b : Y_a \rangle$.

2.6. Integrable flows. Here we compute solutions to a few differential equations defined by certain homogeneous vector fields. In particular, significant simplifications take place in the following two cases. First, let $(q,t) \mapsto X(q,t)$ be a time-varying vector field on Q, and consider the differential equation on TQ

$$\dot{v}_q = X^{\text{lift}}(v_q, t) \tag{2.10}$$

with initial condition $v_q(0) = v_0 \in T_{q_0}Q$. It can be seen that

$$\Phi_{0,t}^{X^{\text{lift}}}(v_0) = v_0 + \int_0^t X(q_0, s) ds, \tag{2.11}$$

that is, in coordinates

$$\Phi_{0,t}^{X^{\text{lift}}} \left(\begin{bmatrix} q_0 \\ v_0 \end{bmatrix} \right) = \begin{bmatrix} q_0 \\ v_0 + \int_0^t X(q_0, s) ds \end{bmatrix}.$$

Next, let $X_0 \in \mathcal{P}_0$ and $X_1 \in \mathcal{P}_1$, and consider the differential equation

$$\dot{v}_q = X_0(v_q, t) + X_1(v_q, t) \tag{2.12}$$

with initial condition $v_q(0) = 0_{q_0} \in T_{q_0}Q$. Define the vector field $X_{0,1}$ on Q and its flow $\zeta : [0,T] \mapsto Q$ via

$$\begin{aligned} X_{0,1} &= T\pi \circ X_0, \\ \zeta(t) &= \Phi_{0,t}^{X_{0,1}}(q_0), \end{aligned}$$

where $T\pi : TTQ \to TQ$ is the tangent map to the projection map $\pi : TQ \to Q$. In coordinates, this vector field consists of the first n components of the vector field $X_0 = \begin{bmatrix} X_{0,1}(q,t)', & X_{0,2}(q,v,t)' \end{bmatrix}'$ on TQ. It can be seen that

$$\Phi_{0,t}^{X_0 + X_1}(0_{q_0}) = 0_{\zeta(t)},$$

that is, in coordinates

$$\Phi_{0,t}^{X_0 + X_1} \left(\begin{bmatrix} q_0 \\ 0 \end{bmatrix} \right) = \begin{bmatrix} \zeta(t) \\ 0 \end{bmatrix}.$$

The key observation in proving this statement is that the components of $X_{0,2}$ and X_1 are polynomials in $\{v^1, \ldots, v^n\}$ of degree at least 1. Since the initial velocity is assumed to be zero, v_q remains zero for all time.

2.7. Analyticity and bounds over complex neighborhoods. In this section we introduce a norm on the set of analytic vector fields over a compact subset of Q. We also provide bounds to partial derivatives of analytic functions. The bounds are *not* coordinate-free, i.e., they depend on the specific selection of a coordinate system. Accordingly, the treatment here assumes $Q = \mathbb{R}^n$.

Let q_0 be a point on $\mathbb{R}^n$, let σ be a positive scalar, and define the complex σ-neighborhood of q_0 in $\mathbb{C}^n$ as

$$B_\sigma(q_0) = \{z \in \mathbb{C}^n : \|z - q_0\| < \sigma\}.$$

Let f be a real analytic function on $\mathbb{R}^n$ that admits a bounded analytic continuation over $B_\sigma(q_0)$. The norm of f is defined as

$$\|f\|_\sigma \triangleq \max_{z \in B_\sigma(q_0)} |f(z)|,$$

where f denotes both the function over $\mathbb{R}^n$ and its analytic continuation. Given a time-varying vector field $(q,t) \mapsto Y(q,t) = Y_t(q)$, let Y_t^i be its ith component with respect to the usual basis on $\mathbb{R}^n$. Assuming $t \in [0,T]$ and assuming that every component function Y_t^i is analytic over $B_\sigma(q_0)$, we define the norm of Y as

$$\|Y\|_{\sigma,T} \triangleq \max_{t\in[0,T]} \max_{i\in\{1,\dots,n\}} \|Y_t^i\|_\sigma.$$

In what follows, we will often simplify notation by neglecting the subscript T in the norm of a time-varying vector field. Finally, given an affine connection ∇ with Christoffel symbols $\{\Gamma_{jk}^i : i,j,k \in \{1,\dots,n\}\}$, we introduce the notation

$$\|\Gamma\|_\sigma \triangleq \max_{ijk} \left\|\Gamma_{jk}^i\right\|_\sigma .$$

Next, we examine the norm of partial derivatives of these objects. Recall that the Cauchy integral representation of analytic functions leads to bounds on high-order derivatives of analytic functions in terms of the norm of the functions themselves; see the so-called Cauchy estimates in [27, section 2.3] and [37]. Let $(i_1,\dots,i_m)$ be a collection of integers belonging to $\{1,\dots,n\}$, and let σ' be a positive real strictly less than σ. It is known that

$$\|\partial^m f\|_{\sigma'} \triangleq \max_{i_1,\dots,i_m} \left\| \frac{\partial^m f}{\partial q_{i_1} \cdots \partial q_{i_m}} \right\|_{\sigma'} \leq m!\, \delta^m \|f\|_\sigma,$$

where $\delta = n/(\sigma - \sigma')$. The quantity $\partial^m f / \partial q_{i_1} \cdots \partial q_{i_m}$ is a real function; it is bounded by bounding its analytic continuation over $B_\sigma(q_0)$. Similarly, for vector fields

$$\|\partial^m Y\|_{\sigma'} \triangleq \max_{t\in[0,T]} \max_{i,i_1,\dots,i_m} \left\| \frac{\partial^m Y_t^i}{\partial q_{i_1} \cdots \partial q_{i_m}} \right\|_{\sigma'} \leq m!\, \delta^m \|Y\|_\sigma,$$

and for the Christoffel symbols

$$\|\partial^m \Gamma\|_{\sigma'} \triangleq \max_{i,j,k,i_1,\dots,i_m} \left\| \frac{\partial^m \Gamma_{jk}^i}{\partial q_{i_1} \cdots \partial q_{i_m}} \right\|_{\sigma'} \leq m!\, \delta^m \|\Gamma\|_\sigma.$$

3. A series expansion for mechanical control systems. This section describes first a preliminary bound, and then the main result of the paper, that is, a series expansion describing the evolution of a forced control system starting at rest.

Problem 3.1. Assume that the functions $q \mapsto \Gamma^i_{jk}(q)$ and the vector field $(q,t) \mapsto Y(q,t)$ are analytic in $q \in Q$ and integrable in $t \in [0,T]$ for some positive time T. Let $\gamma : [0,T] \mapsto Q$ be the solution to the differential equation (2.7) with initial condition $\dot\gamma(0) = 0_{q_0}$. Characterize γ as a series expansion containing iterated symmetric products and time integrals of Y.

We start with a conservative bound.

LEMMA 3.2 (bound on evolution). *Consider the system as described in Problem* 3.1. *Select a coordinate system about the point $q_0 \in Q$, and let σ be a positive constant. A sufficient condition for $\gamma([0,T])$ to be a subset of $B_\sigma(q_0)$ is that*

$$\|Y\|_\sigma T^2 < \frac{\eta^2(\sigma n^2 \|\Gamma\|_\sigma)}{n^2\|\Gamma\|_\sigma}, \tag{3.1}$$

where the function $\eta : x \in \mathbb{R}_+ \to [0,\pi/2]$ is the unique solution to $\eta\tan(\eta) = x$.

Proof. Let $T_0 < T$ be the smallest time at which the solution γ reaches the distance $\|\gamma(T_0) - q_0\| = \sigma$. If the solution never reaches this distance, then $\gamma([0,T])$ is obviously a subset of $B_\sigma(q_0)$. Since $\gamma([0,T_0]) \subset B_\sigma(q_0)$ for all $t \in [0,T_0]$, we have the bound $\|\dot\gamma(t)\| \le y(t)$, where

$$\dot y = n^2\|\Gamma\|_\sigma y^2 + \|Y\|_\sigma, \qquad y(0) = 0.$$

The solution to this initial value problem is

$$y(t) = \sqrt{\frac{\|Y\|_\sigma}{n^2\|\Gamma\|_\sigma}}\tan\left(\sqrt{\|Y\|_\sigma n^2\|\Gamma\|_\sigma}\, t\right).$$

Straightforward manipulations show that the condition in (3.1) is equivalent to $Ty(T) < \sigma$. But since y is a monotone function, $T_0 y(T_0) < \sigma$ also. Note that $\|\gamma(0) - q_0\| = 0$ and

$$\frac{d}{dt}\|\gamma(t) - q_0\| \le \|\dot\gamma\| \le y(t) < \sigma/T_0$$

for all $t \in [0,T_0]$. Therefore, $\|\gamma(T_0) - q_0\| < T_0\sigma/T_0$, and the contradiction is now immediate. □

We are now ready to present the main theorem.

THEOREM 3.3 (evolution of a forced mechanical system starting at rest). *Consider the system as described in Problem* 3.1. *Define recursively the time-varying vector fields V_k:*

$$V_1(q,t) = \int_0^t Y(q,s)ds, \tag{3.2}$$

$$V_k(q,t) = -\frac{1}{2}\sum_{j=1}^{k-1}\int_0^t \langle V_j(q,s)\; V_{k-j}(q,s)\rangle ds, \qquad k \ge 2. \tag{3.3}$$

Select a coordinate system about the point $q_0 \in Q$, let $\sigma > \sigma'$ be two positive constants, and assume that

$$\|Y\|_\sigma T^2 < L \triangleq \min\left\{\frac{\sigma - \sigma'}{2^4 n^2(n+1)}, \frac{1}{2^4 n(n+1)\|\Gamma\|_\sigma}, \frac{\eta^2(\sigma' n^2\|\Gamma\|_{\sigma'})}{n^2\|\Gamma\|_{\sigma'}}\right\}. \tag{3.4}$$

Then the solution $\gamma : [0,T] \to Q$ satisfies

$$\dot\gamma(t) = \sum_{k=1}^{+\infty} V_k(\gamma(t),t), \tag{3.5}$$

where V_k satisfies the bound

$$\|V_k\|_{\sigma'} \leq L^{1-k}\, \|Y\|_\sigma^k\, t^{2k-1}, \tag{3.6}$$

and the series $(q,t) \mapsto \sum_{k=1}^\infty V_k(q,t)$ converges absolutely and uniformly for $q \in B_{\sigma'}(q_0)$ and for $t \in [0,T]$.

A few comments on the various steps of the proof are appropriate. First, we investigate how to write the flow of a mechanical control system as the composition of more elementary flows. Two observations play a key role: the homogeneity of system (2.7) renders the computations tractable, and the simplifying procedure can be easily repeated giving rise to an iterative procedure. Second, we prove absolute and uniform convergence of the series expansion resulting from the first formal part of the proof. The proof of the bounds is inspired by the treatment in [2, Proposition 2.1], but it is considerably more complicated here by the recursive nature of the series expansion. Once the series is formally derived and it is proven to be convergent, a limiting argument leads to the final statement in (3.5).

Proof. Part I. *Here we write the solution to* (2.7) *as composition of the flow of two separate vector fields, one of which is defined recursively.*

Let k be a strictly positive integer, let X_k, Y_k, W_k be time-varying vector fields on Q, and let $v_{q,k}$ be a smooth curve on TQ that satisfies the differential equation

$$\begin{aligned} \dot v_{q,k} &= \left(Z + [X_k^{\mathrm{lift}}, Z] + Y_k^{\mathrm{lift}} + W_k^{\mathrm{lift}}\right)(v_{q,k},t), \\ v_{q,k}(0) &= 0_{q_0}. \end{aligned} \tag{3.7}$$

The mechanical system in (2.7) corresponds to setting $k=1$, $X_1 = W_1 = 0$, $Y_1 = Y(q,t)$, and, accordingly, $\dot\gamma(t) = v_{q,1}(t)$. Using the formula in (2.4) and (2.5) discussed in section 2.2, we set

$$v_{q,k}(t) = \Phi_{0,t}^{Y_k^{\mathrm{lift}}}(v_{q,k+1}(t)) \tag{3.8}$$

and

$$\begin{aligned} \dot v_{q,k+1} &= \left(\left(\Phi_{0,t}^{Y_k^{\mathrm{lift}}}\right)^* (Z + [X_k^{\mathrm{lift}}, Z] + W_k^{\mathrm{lift}})\right)(v_{q,k+1}), \\ v_{q,k+1}(0) &= 0_{q_0}, \end{aligned} \tag{3.9}$$

where we compute the pull-back along the flow by means of the infinite series in (2.6). Remarkably, this series reduces to a finite sum. From the discussion in section 2.5 on the Lie algebraic structure of the various vector fields, we have

$$\begin{gathered} \mathrm{ad}_{Y_k^{\mathrm{lift}}}^{m+2} Z = 0, \\ \mathrm{ad}_{Y_k^{\mathrm{lift}}}^{m+1} [X_k^{\mathrm{lift}}, Z] = 0, \quad \mathrm{ad}_{Y_k^{\mathrm{lift}}}^{m} W_k^{\mathrm{lift}} = 0 \end{gathered}$$

for all $m \geq 1$. With a little bookkeeping we can exploit these equalities and compute

$$\begin{aligned}
&\left(\Phi_{0,t}^{Y_k^{\text{lift}}}\right)^* (Z+[X_k^{\text{lift}}, Z] + W_k^{\text{lift}}) \\
&\qquad = Z + [X_k^{\text{lift}}, Z] + W_k^{\text{lift}} + \int_0^t [Y_k^{\text{lift}}(s), \big(Z + [X_k^{\text{lift}}, Z]\big)]ds \\
&\qquad\qquad + \int_0^t \int_0^{s_1} [Y_k^{\text{lift}}(s_2), [Y_k^{\text{lift}}(s_1), Z]]ds_2 ds_1 \\
&\qquad = Z + [X_k^{\text{lift}} + \overline{Y}_k^{\text{lift}}, Z] + [\overline{Y}_k^{\text{lift}}(s), [X_k^{\text{lift}}, Z]] + W_k^{\text{lift}} \\
&\qquad\qquad + \int_0^t \int_0^{s_1} [Y_k^{\text{lift}}(s_2), [Y_k^{\text{lift}}(s_1), Z]]ds_2 ds_1 \\
&\qquad = Z + [X_k^{\text{lift}} + \overline{Y}_k^{\text{lift}}, Z] - \left\langle \overline{Y}_k : X_k \right\rangle^{\text{lift}} + W_k^{\text{lift}} - \frac{1}{2} \left\langle \overline{Y}_k : \overline{Y}_k \right\rangle^{\text{lift}},
\end{aligned}$$

where we have used the $\bar{\cdot}$ notation introduced in (2.1). The last equality also relies on

$$\int_0^t \int_0^{s_1} [Y_k^{\text{lift}}(s_2), [Y_k^{\text{lift}}(s_1), Z]]ds_2 ds_1 = -\frac{1}{2} \left\langle \overline{Y}_k : \overline{Y}_k \right\rangle^{\text{lift}},$$

which follows from an integration by parts and the symmetry of the symmetric product. Remarkably, the differential equation describing the evolution of $v_{k+1}(t)$ is of the same form as (3.7) describing the evolution of $v_{q,k}(t)$, where

$$\begin{aligned}
X_{k+1} &= X_k + \overline{Y}_k, \\
Y_{k+1} + W_{k+1} &= -\left\langle \overline{Y}_k \ : \ X_k + \tfrac{1}{2}\overline{Y}_k \right\rangle + W_k.
\end{aligned}$$

The vector field X_k can be computed and substituted in as

$$\begin{aligned}
X_k &= \sum_{j=1}^{k-1} \overline{Y}_j, \\
Y_{k+1} + W_{k+1} &= -\left\langle \overline{Y}_k \ : \ \sum_{j=1}^{k-1} \overline{Y}_j + \frac{1}{2}\overline{Y}_k \right\rangle + W_k. \qquad (3.10)
\end{aligned}$$

Notice that the quantities Y_k and W_k are not yet uniquely determined. Equation (3.10) is verified for all k if and only if for all m

$$(Y_2 + Y_3 + \cdots + Y_{m+1}) + W_{m+1} = -\sum_{k=1}^{m} \left\langle \overline{Y}_k : \sum_{j=1}^{k-1} \overline{Y}_j + \frac{1}{2}\overline{Y}_k \right\rangle, \qquad (3.11)$$

where we used $W_1 = 0$. Some further manipulation leads to

$$\begin{aligned}
\sum_{k=1}^{m} \left\langle \overline{Y}_k : \sum_{j=1}^{k-1} \overline{Y}_j + \frac{1}{2}\overline{Y}_k \right\rangle &= \sum_{k=1}^{m} \sum_{j=1}^{k-1} \left\langle \overline{Y}_k : \overline{Y}_j \right\rangle + \frac{1}{2} \sum_{k=1}^{m} \left\langle \overline{Y}_k : \overline{Y}_k \right\rangle \\
&= \frac{1}{2} \sum_{j,k=1, j\neq k}^{m} \left\langle \overline{Y}_k : \overline{Y}_j \right\rangle + \frac{1}{2} \sum_{k=1}^{m} \left\langle \overline{Y}_k : \overline{Y}_k \right\rangle \\
&= \frac{1}{2} \sum_{j,k=1}^{m} \left\langle \overline{Y}_k : \overline{Y}_j \right\rangle.
\end{aligned}$$

A selection of $\{Y_i : i \in \{1,\dots,m\}\}$, and W_{m+1} that satisfies (3.11) is

$$Y_i = -\frac{1}{2} \sum_{j,k=1, j+k=i}^{m} \left\langle \overline{Y}_k : \overline{Y}_j \right\rangle = -\frac{1}{2} \sum_{j=1}^{i-1} \left\langle \overline{Y}_j : \overline{Y}_{i-j} \right\rangle, \tag{3.12}$$

$$W_{m+1} = -\frac{1}{2} \sum_{j,k=1, j+k>m}^{m} \left\langle \overline{Y}_k : \overline{Y}_j \right\rangle.$$

Note that (3.12) is a well-defined recursive relationship, and note that the recursive definition of V_k in (3.3) and (3.2) inside the theorem statement corresponds to setting $V_k(q,t) = \overline{Y}_k(q,t)$. The iteration procedure proves that, for any $k \geq 2$, the solution to the original mechanical system $\dot\gamma = v_{q,1} : [0,T] \mapsto TQ$ satisfies

$$\dot\gamma(t) = \left(\Phi_{0,t}^{Y_1^{\text{lift}}} \circ \Phi_{0,t}^{Y_2^{\text{lift}}} \circ \cdots \circ \Phi_{0,t}^{Y_{k-1}^{\text{lift}}} \right) (v_{q,k}(t)),$$

where $v_{q,k} : [0,T] \mapsto TQ$ is the solution to (3.7). The flow $\dot\gamma$ is now written as the composition of k flows, and a first simplification is immediate. For all integers i, j and for all times s_1, s_2 the vector fields Y_i^{lift} and Y_j^{lift} commute, that is,

$$[Y_i^{\text{lift}}(v_q, s_1), Y_j^{\text{lift}}(v_q, s_2)] = 0,$$

so that γ is the solution to

$$\dot\gamma(t) = \Phi_{0,t}^{\sum_{j=1}^{k-1} Y_j^{\text{lift}}} (v_{q,k}(t)). \tag{3.13}$$

A second simplification is also straightforward. The vector field in (3.13) is homogeneous of degree 0, i.e., it is in the form of (2.10). According to the result in (2.11), we have for all $t \in [0,T]$

$$\dot\gamma(t) = v_{q,k}(t) + \sum_{j=1}^{k-1} \overline{Y}_j(\pi(v_{q,k}(t)), t), \tag{3.14}$$

where the sequence of vector fields Y_j is defined via (3.12) and where the curve $v_{q,k} : [0,T] \mapsto TQ$ is the solution to

$$\frac{dv_{q,k}}{dt} = \left(Z + \left[\sum_{j=1}^{k-1} \overline{Y}_j^{\text{lift}},\, Z \right] + Y_k^{\text{lift}} + W_k^{\text{lift}} \right) (v_{q,k}, t), \tag{3.15}$$

$$v_{q,k}(0) = 0_{q_0}.$$

Part II. *Here we show absolute and uniform convergence of the series* $\sum_{k=1}^{\infty} Y_k(q,t)$ *over all* q *in a compact neighborhood of* q_0 *and for all* $t \leq T$.

Given the vector field Y, let $\Omega_1 = \{Y\}$, and define recursively the set Ω_k to be the collection of vector fields $-\frac{1}{2} \left\langle \overline{B}_i : \overline{B}_{k-i} \right\rangle$ for all $B_i \in \Omega_i$ and $B_{k-i} \in \Omega_{k-i}$. The first few sets are

$$\Omega_1 = \{Y\}, \quad \Omega_2 = \left\{ -\frac{1}{2} \left\langle \overline{Y} : \overline{Y} \right\rangle \right\}, \quad \Omega_3 = \left\{ \frac{1}{4} \left\langle \overline{Y} : \overline{\left\langle \overline{Y} : \overline{Y} \right\rangle} \right\rangle \right\}, \tag{3.16}$$

$$\Omega_4 = \left\{ -\frac{1}{8} \left\langle \overline{Y} : \overline{\left\langle \overline{Y} : \overline{\left\langle \overline{Y} : \overline{Y} \right\rangle} \right\rangle} \right\rangle, -\frac{1}{8} \left\langle \overline{\left\langle \overline{Y} : \overline{Y} \right\rangle} : \overline{\left\langle \overline{Y} : \overline{Y} \right\rangle} \right\rangle \right\}.$$

Next, we prove by induction that, for all k, the vector field Y_k is the sum of N_k vector fields belonging to Ω_k. The statement is true at $k = 1$ with $N_1 = 1$. We assume it is true for all $j < k$ and prove it for k. Because of the induction assumption, we write $Y_j = \sum_{a=1}^{N_j} B_{j,a}$, where the $B_{j,a}$ are elements in Ω_j. We compute

$$\begin{aligned} Y_k &= -\frac{1}{2}\sum_{j=1}^{k-1} \left\langle \overline{Y}_j : \overline{Y}_{k-j} \right\rangle \\ &= -\frac{1}{2}\sum_{j=1}^{k-1} \left\langle \sum_{a=1}^{N_j} \overline{B_{j,a}} : \sum_{b=1}^{N_{k-j}} \overline{B_{k-j,b}} \right\rangle \\ &= \sum_{j=1}^{k-1}\sum_{a=1}^{N_j}\sum_{b=1}^{N_{k-j}} \underbrace{-\frac{1}{2}\left\langle \overline{B_{j,a}} : \overline{B_{k-j,b}} \right\rangle}_{\in\ \Omega_k}. \end{aligned}$$

This concludes the proof by induction, and the recursive relation on N_k is

$$N_1 = 1, \qquad N_k = \sum_{j=1}^{k-1} N_j N_{k-j}, \quad k \geq 2. \tag{3.17}$$

As we discuss in the appendix, the sequence N_k can be explicitly computed and bounded as

$$N_k = \frac{1}{k}\binom{2k-2}{k-1} \leq \frac{2^{2(k-1)}}{k-\frac{1}{2}}. \tag{3.18}$$

We now focus our attention on bounding the generic time-varying vector field $(q, s) \mapsto B_k(q, s)$ in Ω_k. Recall the symbols $\delta, \|\cdot\|_\sigma, \|\partial^m \cdot\|_\sigma$ introduced in section 2.7. We claim that there exist sequences of real and integer coefficients $\{c_k : k \in \mathbb{R}\}$ and $\{d_k : k \in \mathbb{N}\}$ such that

$$\|\partial^m B_k\|_{\sigma'} \leq c_k \frac{(m+d_k)!}{d_k!}\, \delta^{m+k-1} \|Y\|_\sigma^k t^{2(k-1)}. \tag{3.19}$$

For convenience, we redefine δ to $\delta = \max\{\frac{n}{\sigma-\sigma'}, \|\Gamma\|_\sigma\}$, so that $\|\partial^m \Gamma\|_{\sigma'} \leq m!\, \delta^{m+1}$. As discussed in that section, the bound in (3.19) is satisfied at $k = 1$ for all $m \in \mathbb{N}$, with $c_1 = 1$, $d_1 = 0$. In what follows we provide a proof by induction on $k \geq 2$.

Any time-varying vector field B_k at $k \geq 2$ can be written as $B_k = -\frac{1}{2}\left\langle \overline{B_a} : \overline{B_b} \right\rangle$ for some $1 \leq a, b \leq k-1, a + b = k$ and $B_a \in \Omega_a$, $B_b \in \Omega_b$. Accordingly, we compute

$$\begin{aligned} &\|\partial^m \left\langle \overline{B_a} : \overline{B_b} \right\rangle\|_{\sigma'} \\ &\quad = \max_{i, i_1, \dots, i_m} \left\| \frac{\partial^m}{\partial q_{i_1} \cdots \partial q_{i_m}} \left(\frac{\partial \overline{B_a^i}}{\partial q^j} \overline{B_b^j} + \frac{\partial \overline{B_b^i}}{\partial q^j} \overline{B_a^j} + \Gamma_{jl}^i (\overline{B_a^j}\, \overline{B_b^l} + \overline{B_b^j}\, \overline{B_a^l}) \right) \right\|_{\sigma'} \\ &\quad \leq \max_i \left(\left\| \partial^m \left(\frac{\partial \overline{B_a^i}}{\partial q^j} \overline{B_b^j} \right) \right\|_{\sigma'} + \left\| \partial^m \left(\frac{\partial \overline{B_b^i}}{\partial q^j} \overline{B_a^j} \right) \right\|_{\sigma'} + 2 \left\| \partial^m \left(\Gamma_{jl}^i \overline{B_a^j}\, \overline{B_b^l} \right) \right\|_{\sigma'} \right). \end{aligned}$$

Relying on the equality

$$\frac{d^m}{dx^m} f(x) g(x) = \sum_{\alpha=0}^{m} \binom{m}{\alpha} \frac{d^\alpha f(x)}{dx^\alpha} \frac{d^{m-\alpha} g(x)}{dx^{m-\alpha}},$$

the first term is bounded according to

$$\begin{aligned}
\left\| \partial^m \left(\frac{\partial \overline{B_a^i}}{\partial q^j} \overline{B_b^j} \right) \right\|_{\sigma'} &\leq n \sum_{\alpha=0}^{m} \frac{m!}{\alpha!(m-\alpha)!} \left\| \partial^{\alpha+1} \overline{B_a} \right\|_{\sigma'} \left\| \partial^{m-\alpha} \overline{B_b} \right\|_{\sigma'} \\
&\leq n \sum_{\alpha=0}^{m} \frac{m!}{\alpha!(m-\alpha)!} \left(c_a \frac{(\alpha+1+d_a)!}{d_a!} \delta^{\alpha+a} \|Y\|_\sigma^a \frac{t^{2a-1}}{2a-1} \right) \\
&\qquad \cdot \left(c_b \frac{(m-\alpha+d_b)!}{d_b!} \delta^{m-\alpha+b-1} \|Y\|_\sigma^b \frac{t^{2b-1}}{2b-1} \right) \\
&= \frac{n c_a c_b}{(2a-1)(2b-1)} \left(\frac{m!}{d_a! d_b!} \sum_{\alpha=0}^{m} \frac{(\alpha+1+d_a)!(m-\alpha+d_b)!}{\alpha!(m-\alpha)!} \right) \\
&\qquad \cdot \delta^{m+a+b-1} \|Y\|_\sigma^{a+b} t^{2(a+b-1)}.
\end{aligned}$$

The third term is bounded according to

$$\begin{aligned}
\left\| \partial^m \left(\Gamma_{jl}^i \overline{B_a^j} \, \overline{B_b^l} \right) \right\|_{\sigma'} &\leq n^2 \sum_{\alpha=0}^{m} \sum_{\beta=0}^{\alpha} \frac{m! \left\| \partial^{m-\alpha} \Gamma \right\|_{\sigma'}}{(m-\alpha)!\beta!(\alpha-\beta)!} \left\| \partial^\beta \overline{B_a} \right\|_{\sigma'} \left\| \partial^{\alpha-\beta} \overline{B_b} \right\|_{\sigma'} \\
&\leq n^2 \sum_{\alpha=0}^{m} \sum_{\beta=0}^{\alpha} \frac{m!}{(m-\alpha)!\beta!(\alpha-\beta)!} \left((m-\alpha)! \, \delta^{m-\alpha+1} \right) \\
&\qquad \cdot \left(c_a \frac{(\beta+d_a)!}{d_a!} \delta^{\beta+a-1} \|Y\|_\sigma^a \frac{t^{2a-1}}{2a-1} \right) \\
&\qquad \cdot \left(c_b \frac{(\alpha-\beta+d_b)!}{d_b!} \delta^{\alpha-\beta+b-1} \|Y\|_\sigma^b \frac{t^{2b-1}}{2b-1} \right) \\
&\leq \frac{n^2 c_a c_b}{(2a-1)(2b-1)} \left(\frac{m!}{d_a! d_b!} \sum_{\alpha=0}^{m} \sum_{\beta=0}^{\alpha} \frac{(\beta+d_a)!(\alpha-\beta+d_b)!}{\beta!(\alpha-\beta)!} \right) \\
&\qquad \cdot \delta^{m+a+b-1} \|Y\|_\sigma^{a+b} t^{2(a+b-1)}.
\end{aligned}$$

To simplify notation, let us define

$$S(l, d_1, d_2) \triangleq \sum_{a=0}^{l} \frac{(a+d_1)!(l-a+d_2)!}{a!(l-a)!}.$$

Putting it all together,

$$\begin{aligned}
\left\| \partial^m \left\langle \overline{B_a} : \overline{B_b} \right\rangle \right\|_{\sigma'} &\leq \frac{n c_a c_b}{(2a-1)(2b-1)} \frac{m!}{d_a! d_b!} \, \delta^{m+a+b-1} \|Y\|_\sigma^{a+b} t^{2(a+b-1)} \\
&\qquad \cdot \left(S(m, d_a+1, d_b) + S(m, d_a, d_b+1) + 2n \sum_{\alpha=0}^{m} S(\alpha, d_a, d_b) \right).
\end{aligned}$$

Equation (A.2) in the appendix implies that

$$S(l, d_1, d_2) = \frac{d_1! d_2! (l+1+d_1+d_2)!}{l!(1+d_1+d_2)!}, \tag{3.20}$$

so that we compute

$$\begin{aligned}
&\frac{m!}{d_a!d_b!}\left(S(m,d_a+1,d_b)+S(m,d_a,d_b+1)+2n\sum_{\alpha=0}^{m}S(\alpha,d_a,d_b)\right)\\
&\quad=\frac{m!}{d_a!d_b!}\left(\Big((d_a+1)!d_b!+d_a!(d_b+1)!\Big)\frac{(m+2+d_a+d_b)!}{m!(2+d_a+d_b)!}\right.\\
&\qquad\left.+2n\sum_{\alpha=0}^{m}\frac{d_a!d_b!(\alpha+1+d_a+d_b)!}{\alpha!(1+d_a+d_b)!}\right)\\
&\quad=\frac{(m+2+d_a+d_b)!}{(1+d_a+d_b)!}+\frac{2nm!}{(1+d_a+d_b)!}\underbrace{\sum_{\alpha=0}^{m}\frac{(\alpha+1+d_a+d_b)!}{\alpha!}}_{S(m,1+d_a+d_b,0)},
\end{aligned}$$

and again applying (3.20) with $(l,d_1,d_2)=(m,1+d_a+d_b,0)$

$$\begin{aligned}
&=\frac{(m+2+d_a+d_b)!}{(1+d_a+d_b)!}+\frac{2nm!}{(1+d_a+d_b)!}\frac{(1+d_a+d_b)!(m+2+d_a+d_b)!}{m!(2+d_a+d_b)!}\\
&=\frac{(m+2+d_a+d_b)!}{(2+d_a+d_b)!}\left(2+2n+d_a+d_b\right).
\end{aligned}$$

Substitute in

$$\begin{aligned}
&\|\partial^m\left\langle\overline{B_a}:\overline{B_b}\right\rangle\|_{\sigma'}\\
&\qquad\le\frac{nc_ac_b(2+2n+d_a+d_b)(m+2+d_a+d_b)!}{(2a-1)(2b-1)(2+d_a+d_b)!}\,\delta^{m+a+b-1}\,\|Y\|_\sigma^{a+b}\,t^{2(a+b-1)}.
\end{aligned}$$

Next, we express everything back in terms of $k=a+b$ and $B_k=-\frac{1}{2}\left\langle\overline{B_a}:\overline{B_b}\right\rangle$. We have that

$$\|\partial^m B_k\|_{\sigma'}\le\max_{a+b=k}\left(\frac{nc_ac_b(2+2n+d_a+d_b)}{2(2a-1)(2b-1)}\,\frac{(m+2+d_a+d_b)!}{(2+d_a+d_b)!}\right)\delta^{m+k-1}\,\|Y\|_\sigma^{k}\,t^{2(k-1)}.$$

Equation (3.19) is proven by defining sequences c_k and d_k such that $c_1=1$, $d_1=0$, together with

$$\begin{aligned}
d_k&\ge\max_{a+b=k}2+d_a+d_b,\\
c_k&\ge\max_{a+b=k}\frac{nc_ac_b(2+2n+d_a+d_b)}{2(2a-1)(2b-1)}.
\end{aligned}$$

It is immediate to see that $d_k=2(k-1)$ satisfies the recursive requirement, so that we require c_k to satisfy $c_1=1$, together with the requirement

$$c_k\ge\max_{a+b=k}\frac{n(k+n-1)c_ac_b}{(2a-1)(2b-1)}=\max_{a\in\{1,\dots k-1\}}\frac{n(k+n-1)c_ac_{k-a}}{(2a-1)(2k-2a-1)}.$$

Consider the polynomial $p(a)=(2a-1)(2k-2a-1)$ in $a\in[1,k-1]$; it assumes its minimum value $(2k-3)$ at $a=1$, or, equivalently, $a=k-1$. Accordingly, a stricter requirement on c_k is

$$c_k\ge\max_{a\in\{1,\dots k-1\}}\frac{n(k+n-1)}{2k-3}c_ac_{k-a}.$$

Since $(k-1)/(2k-3) \leq 1$ and $n/(2k-3) \leq n$ for all $k \geq 2$, a conservative selection of c_k that satisfies this requirement is provided by the sequence

$$c_1 = 1, \qquad c_k = n(1+n) \sum_{a=1}^{k-1} c_a c_{k-a}, \quad k \geq 2.$$

Recalling the definition in (3.17), one can show that $c_k = (n(1+n))^{k-1} N_k$.

Finally, we summarize all the analysis in Part II and prove convergence. Evaluating at $m = 0$ the bound in (3.19), we have

$$\|B_k\|_{\sigma'} \leq (n(1+n))^{k-1} N_k \, \delta^{k-1} \|Y\|_\sigma^k \, t^{2(k-1)},$$

and recalling the bound in (3.18), we compute

$$\begin{aligned} \|Y_k\|_{\sigma'} &\leq N_k \, \|B_k\|_{\sigma'} \leq (n(1+n))^{k-1} N_k^2 \, \delta^{k-1} \, \|Y\|_\sigma^k \, t^{2(k-1)} \\ &\leq \frac{(2^4 n(1+n)\,\delta)^{k-1}}{(k-1/2)^2} \, \|Y\|_\sigma^k \, t^{2(k-1)}. \end{aligned}$$

An immediate consequence is that for $(2^4 n(n+1)\delta) \, \|Y\|_\sigma T^2 < 1$, the series

$$Y_\infty(q,t) \triangleq \lim_{K \to \infty} \sum_{k=1}^{K} Y_k(q,t)$$

converges absolutely and uniformly in $t \in [0,T]$ and $q \in B_{\sigma'}(q_0)$.

Part III. *Here we provide the final limiting argument by collecting various results in Parts* I *and* II *and in Lemma* 3.2.

We start by studying the behavior as $k \to \infty$ of (3.14) and of the initial value problem (3.15) from Part I. We shall exploit a variation of a standard result on the continuous dependence of solutions of differential equations with respect to parameter changes; see [15, chapter I, section 3]. Uniform convergence of the vector field describing a differential equation, say, for example, $\sum_{k=1}^{K} Y_k$, implies the uniform convergence of the solution to the Kth differential equation to the solution of the limiting differential equation. In order to apply this result to the differential equation (3.15), we need to ensure that the vector field on right-hand side converges uniformly and absolutely.

Assume that the time length T and input vector field Y satisfy the bound in (3.4) inside the theorem statement. Then Lemma 3.2 guarantees that $\gamma([0,T]) \subset B_{\sigma'}(q_0)$, and the analysis in Part II guarantees that series $\sum_{k=1}^{\infty} Y_k$ converges absolutely and uniformly over $q \in B_{\sigma'}(q_0)$. Therefore, the series converges uniformly and absolutely along the curve γ. From (3.14) one can deduce that $\gamma(t) = \pi\,(v_{q,k}(t))$, so that the series $\sum_{k=1}^{\infty} Y_k$ converges also along $\pi \circ v_{q,k} : [0,T] \mapsto Q$. Accordingly, we can take the limit as $k \to \infty$ in (3.15).

Notice that uniformly in $t \in [0,T]$ and $q \in B_{\sigma'}(q_0)$

$$\lim_{k \to \infty} Y_k(q,t) = 0 \qquad \text{and} \qquad \lim_{k \to \infty} W_k(q,t) = 0,$$

and define the time-varying vector field

$$V_\infty = \sum_{k=1}^{\infty} V_k = \sum_{k=1}^{\infty} \overline{Y}_k.$$

Taking the limit as $k \to \infty$ in (3.14) and (3.15), one obtains

$$\dot\gamma(t) = v_{q,\infty}(t) + V_\infty(\pi(v_{q,\infty}(t)), t),$$

where the curve $v_{q,\infty} : [0, T] \mapsto TQ$ is the solution to

$$\begin{aligned} \frac{dv_{q,\infty}}{dt} &= \left(Z + \left[V_\infty^{\text{lift}}, Z\right]\right)(v_{q,\infty}, t), \\ v_{q,\infty}(0) &= 0_{q_0}. \end{aligned} \tag{3.21}$$

According to the discussion in section 2.6, the initial value problem in (3.21) can be explicitly integrated. Because $Z \in \mathcal{P}_1$, $[V_\infty^{\text{lift}}, Z] \in \mathcal{P}_0$, and because of the equality

$$T\pi \circ \left[V_\infty^{\text{lift}}, Z\right] = V_\infty,$$

the curve $v_{q,\infty}$ satisfies

$$v_{q,\infty}(t) = 0_{\zeta(t)}, \text{ where } \zeta(t) = \Phi_{0,t}^{V_\infty}(q_0),$$

and therefore

$$\dot\gamma(t) = 0_{\zeta(t)} + V_\infty(\pi(0_{\zeta(t)}), t) = V_\infty(\zeta(t), t).$$

The last two statements imply $\gamma = \zeta$ and are equivalent to (3.5). □

Two brief comments are appropriate. First, it is interesting to emphasize an intermediate result proved in Part II: the V_k term in the series is the sum of the known number of vector fields belonging to the set Ω_k; see the definition preceding (3.16). This additional structure might be useful in controllability or motion planning studies. Second, it is unpleasant to remark that while the series expansion is stated in a coordinate-free context, its convergence properties rely on the introduction of a coordinate system.

4. Applications and extensions. We present a few diverse comments in order to relate the theorem to various earlier works as well as to obtain stronger results under specific additional assumptions on the system.

4.1. The first few order terms and small amplitude forcing. Equation (3.5) is well defined in the sense that, at fixed q, the integration is performed with respect to the time variable. Using the abbreviated notation introduced in (2.1), the first few terms of the sequence $\{V_k : k \in \mathbb{N}\}$ are computed as

$$\begin{aligned} V_1 &= \overline{Y}, \\ V_2 &= -\frac{1}{2}\overline{\langle \overline{Y} : \overline{Y} \rangle}, \\ V_3 &= \frac{1}{2}\overline{\left\langle \overline{\langle \overline{Y} : \overline{Y} \rangle} : \overline{Y} \right\rangle}, \\ V_4 &= -\frac{1}{2}\overline{\left\langle \left\langle \overline{\langle \overline{Y} : \overline{Y} \rangle} : \overline{Y} \right\rangle : \overline{Y} \right\rangle} - \frac{1}{8}\overline{\left\langle \overline{\langle \overline{Y} : \overline{Y} \rangle} : \overline{\langle \overline{Y} : \overline{Y} \rangle} \right\rangle}, \end{aligned}$$

so that we can write

$$\begin{aligned} \dot\gamma(t) = \overline{Y}(\gamma, t) &- \frac{1}{2}\overline{\langle \overline{Y} : \overline{Y} \rangle}(\gamma, t) + \frac{1}{2}\overline{\left\langle \overline{\langle \overline{Y} : \overline{Y} \rangle} : \overline{Y} \right\rangle}(\gamma, t) \\ &- \frac{1}{2}\overline{\left\langle \left\langle \overline{\langle \overline{Y} : \overline{Y} \rangle} : \overline{Y} \right\rangle : \overline{Y} \right\rangle}(\gamma, t) - \frac{1}{8}\overline{\left\langle \overline{\langle \overline{Y} : \overline{Y} \rangle} : \overline{\langle \overline{Y} : \overline{Y} \rangle} \right\rangle}(\gamma, t) + O(\|Y\|^5 t^9). \end{aligned}$$

This series converges under the assumption that the product of final time T and input magnitude $\|Y\|_\sigma$ is small. Typically, in controllability studies [41], it is the final time that is assumed to be small (the famous acronym STLC stands for STLC). Within the context of motion planning problems [30, 13], it is instead convenient to study the small input magnitude case. Motivated by the treatment in [13], let ϵ be a small positive constant, and consider a total acceleration of the form

$$Y(q,t,\epsilon) = \epsilon X_1(q,t) + \epsilon^2 X_2(q,t) + \epsilon^3 X_3(q,t), \qquad t \in [0,1].$$

Accordingly, (3.5) is equivalent to

$$\begin{aligned}\dot\gamma(t) = \epsilon\overline{X}_1(\gamma,t) + \epsilon^2\left(\overline{X}_2 - \frac{1}{2}\overline{\langle \overline{X}_1 : \overline{X}_1\rangle}\right)(\gamma,t) \\ + \epsilon^3\left(\overline{X}_3 - \frac{1}{2}\overline{\langle \overline{X}_1 : \overline{X}_2\rangle} + \frac{1}{2}\overline{\left\langle \overline{\langle \overline{X}_1 : \overline{X}_1\rangle} : \overline{X}_1\right\rangle}\right)(\gamma,t) + O(\epsilon^4).\end{aligned}$$

This expression generalizes the results presented in Proposition 4.1 in [13]. Note that those results were proven via a perturbation theory argument that is not as general and powerful as the treatment in Theorem 3.3.

4.2. Simple Hamiltonian systems with integrable forces. In this and the following section we analyze systems with more structure both in the affine connections ∇ as well as in the input forces Y. Here we consider systems with Lagrangian equal to "kinetic minus potential" and with integrable forces. In the interest of brevity, we refer to the textbooks [16, 34] for a detailed presentation and review here only the necessary notation. The affine connection of a simple system is the Levi–Civita connection associated with the kinetic energy matrix M; that is, the Christoffel symbols are defined according to the usual relationship

$$\Gamma_{ij}^k = \frac{1}{2}M^{mk}\left(\frac{\partial M_{mj}}{\partial q^i} + \frac{\partial M_{mi}}{\partial q^j} - \frac{\partial M_{ij}}{\partial q^m}\right), \tag{4.1}$$

where M_{ij} and M^{mk} are the components of the matrix representation of M and of its inverse. An integrable time-varying force is written as

$$Y(q,t) = \operatorname{grad}\varphi(q,t), \qquad \text{where} \qquad (\operatorname{grad}\varphi)^i = M^{ij}\frac{\partial\varphi}{\partial q_j}, \tag{4.2}$$

and where φ is a scalar function on $\mathbb{R}^n \times \mathbb{R}$.

One remarkable simplification takes place for a simple system described by a Levi–Civita connection: the set of gradient vector fields is closed under the operation of a symmetric product. Let φ_1, φ_2 be scalar functions on $\mathbb{R}^n$, and define a symmetric product between functions according to

$$\langle \varphi_1 : \varphi_2\rangle \triangleq \frac{\partial\varphi_1}{\partial q}M^{-1}\frac{\partial\varphi_2}{\partial q}. \tag{4.3}$$

Then the symmetric product of the corresponding gradient vector fields equals the gradient of the symmetric product of the functions. In equations,

$$\langle \operatorname{grad}\varphi_1 : \operatorname{grad}\varphi_2\rangle = \operatorname{grad}\langle \varphi_1 : \varphi_2\rangle .$$

We refer to [10] for the proof. Accordingly, the main theorem can be restated as follows.

THEOREM 4.1. *Consider the system as described in Problem* 3.1. *Additionally, let the Christoffel symbols and the input vector field be defined as in* (4.1) *and* (4.2). *Define recursively the time-varying functions*

$$\varphi_1(q,t) = \int_0^t \varphi(q,s)ds,$$
$$\varphi_k(q,t) = -\frac{1}{2}\sum_{j=1}^{k-1}\int_0^t \Big\langle \varphi_j(q,s) :\ \varphi_{k-j}(q,s)\Big\rangle ds, \qquad k \geq 2.$$

Then the solution $\gamma : [0,T] \to Q$ *satisfies*

$$\dot\gamma(t) = \operatorname{grad} \sum_{k=1}^{+\infty} \varphi_k(\gamma(t),t). \tag{4.4}$$

In other words, the flow of a simple Hamiltonian system forced from rest is written as a (time-varying) gradient flow. For completeness, we include a convergence treatment derived from the one in the main theorem.

Remark 4.2. Given $0 < \sigma'' < \sigma' < \sigma$, we assume M and φ to be analytic in a neighborhood $B_\sigma(q_0)$ of q_0 and uniformly integrable in $t \in [0,T]$. Two immediate bounds are

$$\|\operatorname{grad}\varphi\|_{\sigma'} \leq n\|M^{-1}\|_{\sigma'}\left\|\frac{\partial\varphi}{\partial q}\right\|_{\sigma'},$$
$$\|\Gamma\|_{\sigma'} \leq A \triangleq \frac{3n^2}{2(\sigma-\sigma')}\|M^{-1}\|_{\sigma'}\|M\|_\sigma.$$

Accordingly, the bounds in the main theorem can be restated (in a more conservative manner) as follows. If

$$\left\|\frac{\partial\varphi}{\partial q}\right\|_{\sigma'} T^2 < \frac{1}{n\|M^{-1}\|_{\sigma'}}\min\left\{\frac{\sigma'-\sigma''}{2^4n^2(n+1)}, \frac{1}{2^4n(n+1)A}, \frac{\eta^2(n^2\sigma''A)}{n^2A}\right\},$$

the series $\sum_{k=1}^\infty \varphi_k(q,t)$ converges absolutely and uniformly in t and q for all $t \in [0,T]$ and for all q in a neighborhood $B_{\sigma''}(q_0)$ of q_0.

4.3. Invariant systems on Lie groups. In this section we briefly investigate systems with kinetic energy and input forces invariant under a certain group action. These systems have a configuration space G with the structure of an n-dimensional matrix Lie group. Systems in this class include satellites, hovercraft, and underwater vehicles.

The equation of motion (2.7) decouples into a kinematic and dynamic equation in the configuration variable $g \in G$ and the body velocity[3] $v \in \mathbb{R}^n$. The kinematic equation can be written as a matrix differential equation[4] using matrix group notation

[3] More precisely, the body velocity v lives in the Lie algebra of the group G.

[4] Alternatively, the kinematic equation can be written in a system of local coordinates q (e.g., Euler angles in the case of rotation matrices) as $\dot q = J(q)v$, where $J(q)$ is an appropriate Jacobian matrix.

TABLE 4.1
Numerical comparison of various degrees of approximations. The entries in the table are the error values $e_{\epsilon,N}$ that provide a measure of the accuracy of the Nth-order truncated approximation.

ϵ	1	.1	.01
$N=1$	$5.3\cdot 10^{-3}$	$4.6\cdot 10^{-5}$	$4.5\cdot 10^{-7}$
$N=2$	$2.4\cdot 10^{-4}$	$4.2\cdot 10^{-7}$	$4.2\cdot 10^{-10}$
$N=3$	$1.4\cdot 10^{-4}$	$3.0\cdot 10^{-9}$	$2.3\cdot 10^{-13}$
$N=4$	$5.2\cdot 10^{-5}$	$2.4\cdot 10^{-10}$	$3.5\cdot 10^{-15}$

$\dot g = g\widehat v$; we refer to [35] for the details. The dynamic equation, sometimes referred to as Euler–Poincaré, is

$$\dot v^i + \gamma^i_{jk} v^j v^k = y^i(t), \tag{4.5}$$

where the coefficients γ^i_{jk} are determined by the group and metric structure. The curve $y : [0,T] \mapsto \mathbb{R}^n$ denotes the time-varying forcing.

Within this setting, the result in Theorem 3.3 is summarized as follows. The solution to (4.5) with initial condition $v(0) = 0$ is $v(t) = \sum_{k=1}^{\infty} v_k(t)$, where

$$v_1(t) = \int_0^t y(s)ds,$$
$$v_k(t) = -\frac{1}{2}\sum_{j=1}^{k-1}\int_0^t \Big\langle v_j(s) :\ v_{k-j}(s)\Big\rangle ds, \qquad k \geq 2,$$

and where the symmetric product between velocity vectors is $\langle x : y\rangle^i = -2\gamma^i_{jk} x^j y^k$. Local convergence for the series expansion can be easily established in this setting.

This result agrees with and indeed supersedes the ones presented in [13] obtained via the perturbation method. The relationship of this case to the more general setting studied in Theorem 3.3 is clarified via the notion of invariant connection; see [5, Appendix B] and [39, section 27, "Variations on a theme by Euler"] for more details.

4.4. Simulations for a three degree of freedom manipulator. In this section we illustrate the approximations derived in Theorem 3.3 by applying them to an example system. We consider a three-link planar manipulator. The configuration is described by three angles $(\theta_1, \theta_2, \theta_3)$. A constant (integrable) force is applied to the first variable. Specifically, we set $\varphi(q,t) = \epsilon\theta_1$, and we let the parameter ϵ vary in the range 10^{-2} to 1. The integration time is $T = 1$ seconds. Setting all lengths, masses, and moments of inertia to unity, the kinetic energy matrix is

$$M = \frac{1}{16}\begin{bmatrix} 25 & 6\cos(\theta_1-\theta_2) & 2\cos(\theta_1-\theta_3) \\ 6\cos(\theta_1-\theta_2) & 21 & 2\cos(\theta_2-\theta_3) \\ 2\cos(\theta_1-\theta_3) & 2\cos(\theta_2-\theta_3) & 17 \end{bmatrix}.$$

The initial condition is assumed to be $q(0) = (0, \pi/4, 0)$. We investigate the error value $e_{\epsilon,N} = \|\gamma(T) - \gamma_N(T)\|$, where γ_N is the solution to the Nth order truncation: $\dot\gamma_N(t) = \operatorname{grad}\sum_{k=1}^{N}\varphi_k(\gamma_N, t)$. An empirical forecast of the $e_{\epsilon,N}$ is computed as follows. Since $T = 1$ and $\|Y\| = O(\epsilon)$, there exist two constants c, d such that the kth term in the series is bounded by $c(d\epsilon)^k$. Summing the neglected contributions from $k = N+1$ to infinity and assuming that $d\epsilon \ll 1$, one can compute $e_{\epsilon,N} \approx c(d\epsilon)^{N+1}$.

We summarize the results of the numerical investigation[5] in Table 4.1. The results are in qualitative agreement with the theoretical forecasts.

5. Conclusions. We have presented a series expansion that describes the evolution of a forced mechanical system. Our result provides a first-order description to the solutions of a second-order initial value problem. Both the series and the proof method provide insight into the geometry of mechanical control systems. The treatment expands on our previous work [10] on high-amplitude high-frequency averaging and vibrational stabilization.

Series expansions are the underlying technique for controllability and motion planning. For mechanical systems moving in the low-velocity regime, these two problems have been tackled with various degrees of success in [32, 13]. Future research will rely on the contributions in this work to develop more general motion planning algorithms than the ones in [13] and sharper sufficient controllability tests than the ones in [32].

Appendix. Some basic identities in combinatorial analysis. We here present a basic result and derive a useful expression that is needed in the proof of the main theorem. The main reference is the method of generating functions as described in section 3.4 in [4]. The first identity is explicitly proven in the reference. If $N_1 = 1$ and

$$N_k = \sum_{j=1}^{k-1} N_j N_{k-j}, \qquad k \geq 2,$$

then

$$N_k = \frac{1}{k}\binom{2k-2}{k-1} = \frac{2^k}{(4k-2)}\frac{1\cdot 3\cdot 5\cdots(2k-1)}{1\cdot 2\cdot 3\cdots k} \leq \frac{4^k}{4k-2}. \tag{A.1}$$

The second equality needed in the proof of Theorem 3.3 is

$$\sum_{a=0}^{k}\binom{a+d_1}{d_1}\binom{k-a+d_2}{d_2} = \binom{k+1+d_1+d_2}{k}. \tag{A.2}$$

To prove it we use the method of generating functions; see [4]. We claim that, for all real x with $|x| < 1$,

$$\sum_{k=0}^{\infty}\left(\sum_{a=0}^{k}\binom{a+d_1}{d_1}\binom{k-a+d_2}{d_2}\right)x^k = \sum_{k=0}^{\infty}\binom{k+1+d_1+d_2}{k}x^k. \tag{A.3}$$

The first step is to notice that

$$\begin{aligned}\sum_{k=0}^{\infty}\left(\sum_{a=0}^{k}\binom{a+d_1}{d_1}\binom{k-a+d_2}{d_2}\right)x^k &= \sum_{k=0}^{\infty}\sum_{m+n=k}\binom{n+d_1}{d_1}\binom{m+d_2}{d_2}x^{n+m}\\ &= \left(\sum_{n=0}^{\infty}\binom{n+d_1}{d_1}x^n\right)\left(\sum_{m=0}^{\infty}\binom{m+d_2}{d_2}x^m\right).\end{aligned}$$

[5] The numerical integration is performed inside the Mathematica environment, specifying 16 digits of accuracy and 32 digits of working precision.

Accordingly, we define

$$f_a(x) = \sum_{m=0}^{\infty} \binom{m+a}{a} x^m, \tag{A.4}$$

and the thesis in (A.3) is equivalent to proving that

$$f_{d_1}(x) f_{d_2}(x) = f_{d_1+d_2+1}(x). \tag{A.5}$$

In passing, we also note that the convergence radius of f is $|x| < 1$.

The second step is to study the properties of f. First of all,

$$\begin{aligned} f_a(x) &= \sum_{m=0}^{\infty} \frac{(m+a)!}{m!a!} x^m = \frac{1}{a!} \sum_{m=0}^{\infty} (m+a) \cdots (m+1) x^m \\ &= \frac{1}{a!} \frac{d^a}{dx^a} \sum_{m=0}^{\infty} x^{m+a} = \frac{1}{a!} \frac{d^a}{dx^a} \left(x^a \sum_{m=0}^{\infty} x^m \right) = \frac{1}{a!} \frac{d^a}{dx^a} \frac{x^a}{1-x}. \end{aligned}$$

Additionally, it is immediate to see that

$$f_0(x) = \frac{1}{1-x}, \qquad x f_0(x) = f_0(x) - 1,$$

and, consequently,

$$f_a(x) = \frac{1}{a!} \frac{d^a}{dx^a} \frac{1}{1-x} = \frac{1}{a!} \frac{d^a}{dx^a} f_0(x).$$

Finally, we prove by induction that

$$f_a(x) = f_0(x)^{a+1}. \tag{A.6}$$

At $a = 0$ the statement is obvious. We assume it is true up to a and compute

$$\begin{aligned} f_{a+1}(x) &= \frac{1}{(a+1)!} \frac{d^{a+1}}{dx^{a+1}} \frac{1}{1-x} = \frac{1}{(a+1)!} \frac{d^a}{dx^a} \left(\frac{1}{1-x} \right)^2 \\ &= \frac{1}{(a+1)!} \sum_{b=0}^{a} \binom{a}{b} \left(\frac{d^b}{dx^b} \frac{1}{1-x} \right) \left(\frac{d^{a-b}}{dx^{a-b}} \frac{1}{1-x} \right) \\ &= \frac{a!}{(a+1)!} \sum_{b=0}^{a} \left(\frac{1}{b!} \frac{d^b}{dx^b} \frac{1}{1-x} \right) \left(\frac{1}{(a-b)!} \frac{d^{a-b}}{dx^{a-b}} \frac{1}{1-x} \right) \\ &= \frac{1}{a+1} \sum_{b=0}^{a} \left(\frac{1}{1-x} \right)^{b+1} \left(\frac{1}{1-x} \right)^{a-b+1} = \left(\frac{1}{1-x} \right)^{a+2}. \end{aligned}$$

This concludes the proof of (A.6), which immediately implies (A.5) and the main thesis in (A.3).

Acknowledgments. The author thanks Jim Radford and Andrew D. Lewis for helpful and stimulating discussions.

REFERENCES

[1] R. Abraham, J. E. Marsden, and T. S. Ratiu, *Manifolds, Tensor Analysis, and Applications*, 2nd ed., Appl. Math. Sci. 75, Springer-Verlag, New York, 1988.

[2] A. A. Agrachev and R. V. Gamkrelidze, *The exponential representation of flows and the chronological calculus*, Math. USSR Sbornik, 35 (1978), pp. 727–785.

[3] A. A. Agrachev and R. V. Gamkrelidze, *Local controllability and semigroups of diffeomorphisms*, Acta Appl. Math., 32 (1993), pp. 1–57.

[4] G. E. Andrews, *Number Theory*, Dover Publications, New York, 1994.

[5] V. I. Arnol'd, *Mathematical Methods of Classical Mechanics*, 2nd ed., Grad. Texts in Math. 60, Springer-Verlag, New York, 1989.

[6] J. Baillieul, *Stable average motions of mechanical systems subject to periodic forcing*, in Dynamics and Control of Mechanical Systems: The Falling Cat and Related Problems, Fields Inst. Commun. 1, M. J. Enos, ed., AMS, Providence, RI, 1993, pp. 1–23.

[7] A. L. Besse, *Manifolds All of Whose Geodesics are Closed*, Springer-Verlag, New York, 1978.

[8] R. M. Bianchini and G. Stefani, *Graded approximations and controllability along a trajectory*, SIAM J. Control Optim., 28 (1990), pp. 903–924.

[9] A. M. Bloch and P. E. Crouch, *Nonholonomic control systems on Riemannian manifolds*, SIAM J. Control Optim., 33 (1995), pp. 126–148.

[10] F. Bullo, *Vibrational control of mechanical systems*, SIAM J. Control Optim., to appear.

[11] F. Bullo, *A series describing the evolution of mechanical control systems*, in Proceedings of the IFAC World Conference, Vol. E, Beijing, China, 1999, pp. 479–485.

[12] F. Bullo and N. E. Leonard, *Motion control for underactuated mechanical systems on Lie groups*, in Proceedings of the European Control Conference, Brussels, Belgium, 1997, p. 480.

[13] F. Bullo, N. E. Leonard, and A. D. Lewis, *Controllability and motion algorithms for underactuated Lagrangian systems on Lie groups*, IEEE Trans. Automat. Control, 45 (2000), pp. 1437–1454.

[14] K. T. Chen, *Integration of paths, geometric invariants, and a generalized Baker–Hausdorff formula*, Ann. of Math (2), 67 (1957), pp. 164–178.

[15] W. A. Coppel, *Stability and Asymptotic Behavior of Differential Equations*, D.C. Heath, Boston, 1965.

[16] M. P. Do Carmo, *Riemannian Geometry*, Birkhäuser, Boston, 1992.

[17] M. Fliess, *Fonctionelles causales non linéares et indeterminées non commutatives*, Bull. Soc. Math. France, 109 (1981), pp. 3–40.

[18] P. D. Giamberardino, M. Djemai, S. Monaco, and D. Normand-Cyrot, *Exact steering and stabilization of a PVTOL aircraft*, in Proceedings of the IEEE Conference on Decision and Control, San Diego, CA, 1997, pp. 2049–2054.

[19] W. S. Gray and J. M. A. Scherpen, *Hankel operators and Gramians for nonlinear systems*, in Proceedings of the IEEE Conference on Decision and Control, Tampa, FL, 1998, pp. 1416–1421.

[20] H. Hermes, *Nilpotent and high-order approximations of vector field systems*, SIAM Rev., 33 (1991), pp. 238–264.

[21] A. Isidori, *Nonlinear Control Systems*, 3rd ed., Springer-Verlag, New York, 1995.

[22] M. Kawski, *High-order small-time local controllability*, in Nonlinear Controllability and Optimal Control, H. J. Sussmann, ed., Dekker, New York, 1990, pp. 441–477.

[23] M. Kawski, *Geometric homogeneity and applications to stabilization*, in Proceedings of the Nonlinear Control Systems Design Symposium (NOLCOS), Pergamon, Tahoe City, CA, 1995, pp. 251–256.

[24] M. Kawski, *Nonlinear control and combinatorics of words*, in Geometry of Feedback and Optimal Control, B. Jakubczyk and W. Respondek, eds., Dekker, New York, 1998, pp. 305–346.

[25] M. Kawski and H. J. Sussmann, *Noncommutative power series and formal Lie-algebraic techniques in nonlinear control theory*, in Operators, Systems, and Linear Algebra, U. Helmke, D. Pratzel-Wolters, and E. Zerz, eds., Teubner, Stuttgart, Germany, 1997, pp. 111–128.

[26] I. Kolmanovsky and N. H. McClamroch, *Stabilizing feedback laws for internally actuated multibody systems in-space*, Nonlinear Anal. Theory Methods Appl., 26 (1996), pp. 1461–1479.

[27] S. G. Krantz, *Function Theory of Several Complex Variables*, Pure Appl. Math., John Wiley and Sons, New York, 1982.

[28] G. Lafferriere and H. J. Sussmann, *A differential geometric approach to motion planning*, in Nonholonomic Motion Planning, Z. Li and J. F. Canny, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1993, pp. 235–270.
[29] S. Lang, *Differentiable and Riemannian Manifolds*, 3rd ed., Springer-Verlag, New York, 1995.
[30] N. E. Leonard and P. S. Krishnaprasad, *Motion control of drift-free, left-invariant systems on Lie groups*, IEEE Trans. Automat. Control, 40 (1995), pp. 1539–1554.
[31] A. D. Lewis, *Simple mechanical control systems with constraints*, IEEE Trans. Automat. Control, 45 (2000), pp. 1420–1436.
[32] A. D. Lewis and R. M. Murray, *Configuration controllability of simple mechanical control systems*, SIAM J. Control Optim., 35 (1997), pp. 766–790.
[33] W. Magnus, *On the exponential solution of differential equations for a linear operator*, Comm. Pure Appl. Math., 7 (1954), pp. 649–673.
[34] J. E. Marsden and T. S. Ratiu, *Introduction to Mechanics and Symmetry*, 2nd ed., Springer-Verlag, New York, 1999.
[35] R. M. Murray, Z. X. Li, and S. S. Sastry, *A Mathematical Introduction to Robotic Manipulation*, CRC Press, Boca Raton, FL, 1994.
[36] R. M. Murray and S. S. Sastry, *Nonholonomic motion planning: Steering using sinusoids*, IEEE Trans. Automat. Control, 5 (1993), pp. 700–726.
[37] W. Rudin, *Real and Complex Analysis*, 3rd ed., McGraw-Hill, New York, 1987.
[38] J. A. Sanders and F. Verhulst, *Averaging Methods in Nonlinear Dynamical Systems*, Springer-Verlag, New York, 1985.
[39] D. H. Sattinger and O. L. Weaver, *Lie Groups and Algebras, with Applications to Physics, Geometry and Mechanics*, Appl. Math. Sci. 61, Springer-Verlag, New York, 1986.
[40] E. D. Sontag and H. J. Sussmann, *Time-optimal control of manipulators*, in Proceedings of the IEEE International Conference on Robotics and Automation, San Francisco, CA, 1986, pp. 1692–1697.
[41] H. J. Sussmann, *A sufficient condition for local controllability*, SIAM J. Control Optim., 16 (1978), pp. 790–802.
[42] H. J. Sussmann, *A product expansion of the Chen series*, in Theory and Applications of Nonlinear Control Systems, C. I. Byrnes and A. Lindquist, eds., Elsevier, Oxford, UK, 1986, pp. 323–335.
[43] H. J. Sussmann, *A general theorem on local controllability*, SIAM J. Control Optim., 25 (1987), pp. 158–194.
[44] V. S. Varadarajan, *Lie Groups, Lie Algebras, and Their Representations*, Grad. Texts in Math. 102, Springer-Verlag, New York, 1984.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 1, pp. 191–210

SMOOTH FEEDBACK, GLOBAL STABILIZATION, AND DISTURBANCE ATTENUATION OF NONLINEAR SYSTEMS WITH UNCONTROLLABLE LINEARIZATION*

CHUNJIANG QIAN†, WEI LIN†, AND W. P. DAYAWANSA‡

Abstract. Problems of global asymptotic stabilization and disturbance attenuation are addressed for a class of highly nonlinear systems that are comprised of a lower dimensional zero-dynamics subsystem and a chain of power integrators perturbed by a *nontriangular* vector field. It is shown in this paper that global stabilization and disturbance attenuation are solvable by *smooth* state feedback if one takes full advantage of the characteristics of the system in the feedback design to dominate the nonlinearity rather than to cancel it. A systematic design procedure which is based upon, but generalizes, the recent technique of *adding a power integrator* is developed for the explicit construction of the smooth controllers. Several examples are presented to demonstrate the key features of the proposed nonlinear control schemes.

Key words. adding a power integrator, smooth state feedback, global asymptotic stabilization, disturbance attenuation, high-order nonlinear systems, uncontrollable linearization

AMS subject classifications. 93C10, 93D15, 93D05

PII. S0363012900370090

1. Introduction.

Most of the practical control systems are usually nonlinear with uncertainties or disturbances. One of the important problems in nonlinear control theory is to design a smooth state feedback control law that globally stabilizes the system in the absence of additive disturbances and attenuates the effect of the disturbances on the system output to an arbitrary degree of accuracy in the presence of disturbances. To address this issue, new ideas and powerful concepts have been developed over the last decade, leading to the development of a number of systematic design methods for global synthesis of *affine* systems. Among the proposed approaches, a Lyapunov-like design technique called adding a linear integrator [3, 30], also known as backstepping [16, 24], has been proved to be one of the effective control methods in studying the problems such as global stabilization and disturbance attenuation for a class of nonlinear systems in the so-called normal form [11, 13, 23, 12]

$$\begin{aligned}
&\dot{z} = f_0(z, x_1) + \phi_0(z, x_1)w,\\
&\dot{x}_i = x_{i+1} + f_i(z, x_1, \cdots, x_i) + \phi_i(z, x_1, \cdots, x_i)w, \quad i = 1, \cdots, r, \quad x_{r+1} := v,\\
&y = x_1,
\end{aligned} \tag{1.1}$$

where v, y, and w are the system input, output, and disturbance, respectively.

There are three key features exhibited in system (1.1): (i) The Jacobian linearization of the x-subsystem of (1.1) at the origin is *controllable*. Hence, when $w = 0$ and dim $z = 0$, (1.1) is *feedback linearizable*. (ii) The system is *affine* in the control

*Received by the editors March 29, 2000; accepted for publication (in revised form) February 14, 2001; published electronically May 31, 2001.

http://www.siam.org/journals/sicon/40-1/37009.html

†Department of Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, OH 44106 (qian@nonlinear.cwru.edu, linwei@nonlinear.cwru.edu). The research of these authors was supported in part by the U.S. NSF under grants ECS-9875273 and DMS-9972045.

‡Department of Mathematics, Texas Tech University, Lubbock, TX 79409 (daya@math.ttu.edu). The research of this author was supported by the U.S. NSF under grants ECS 9707927 and ECS 9720357 and by the State of Texas Advanced Research Program under grant 003644-0402-1999.

input v. (iii) (1.1) is of a *lower-triangular* form. These assumptions have been commonly used in most of the existing control schemes that were based on the integrator backstepping design. They guarantee that for system (1.1) the problem of global stabilization when $w = 0$ [11, 17, 18, 16, 24] or the problem of global disturbance attenuation with internal stability when $w \neq 0$ [11, 13, 23, 12, 31] is solvable by smooth state feedback. A globally stabilizing controller can be constructed by using the technique of adding a linear integrator [3, 30] and dealing with a design problem for a "one dimension system" every time. In fact, it was proved in [3, 30] that the cascade system

$$\begin{aligned} \dot{z} &= f(z) + y f_1(z, y), \\ \dot{y} &= u + g_1(z, y) \end{aligned} \tag{1.2}$$

is globally asymptotically stabilizable by smooth state feedback if there exists a positive definite and proper Lyapunov function V such that $L_f V(z) < 0 \; \forall z \neq 0$. The proof was based on the design of a Lyapunov function $U(z, y) = V(z) + \frac{y^2}{2}$ and a smooth controller $u(z, y) = -g_1(z, y) - y - L_{f_1} V$, making $\dot{U} = L_f V(z) - y^2 < 0 \quad \forall (z, y) \neq (0, 0)$. The crucial point here is to use $u(z, y)$ to cancel the nonlinear terms $g_1(z, y)$ and $L_{f_1} V$ in the equation $\dot{U} = L_f V + y L_{f_1} V + y(u + g_1(z, y))$. As illustrated by Theorem 9.2.3 in [11], global stabilization of minimum-phase systems (1.1) with $w = 0$ can then be achieved by repeatedly using this backstepping construct r times.

Many existing design methodologies on global stabilization and disturbance attenuation use this aforementioned construct [11, 3, 30, 16, 23, 1]; hence it becomes clear that they rely essentially on *feedback cancelation* and therefore can only be applied to a class of lower-triangular systems of the form (1.1).

An important question that remains largely open is to what extent the three structural conditions of (1.1)—*feedback linearizability, affine structure, and lower-triangularity*—can be significantly relaxed so that global stabilization and disturbance attenuation are still solvable by *smooth* state feedback for a larger class of highly nonlinear systems (even possibly nonaffine systems) with *uncontrollable* Jacobian linearization. In this paper we shall address this question and provide a partial answer. Specifically, we shall first consider the problem of global stabilization via *smooth* state feedback for a class of nonlinear systems of the form

$$\begin{aligned} \dot{z} &= f_0(z, x_1), \\ \dot{x}_i &= x_{i+1}^{p_i} + f_i(z, x_1, \ldots, x_i, x_{i+1}), \qquad i = 1, \ldots, r, \qquad x_{r+1} := u, \end{aligned} \tag{1.3}$$

where $z \in \mathbb{R}^{n-r}$ and $x = (x_1, \ldots, x_r)^T$ are the states, $u \in \mathbb{R}$ and $y \in \mathbb{R}$ are the system input and output, respectively, and p_i, $i = 1, \ldots, r$, are positive integers. The functions $f_i : R^{n-r+i+1} \to \mathbb{R}$ and $f_0 : \mathbb{R}^{n-r+1} \to \mathbb{R}^{n-r}$ are smooth and evaluate to zero at $(z, x_1, \ldots, x_r) = (0, 0, \ldots, 0)$. The system above can be regarded as a lower dimensional zero-dynamics driven by a nonlinear subsystem that consists of a chain of power integrators perturbed by a *nonstrict-triangular* vector field.

The notable characteristics of system (1.3) are that it is neither feedback linearizable (even partially) nor affine in the control input when $p_i > 1$. More significantly, (1.3) is not in a lower-triangular form due to the appearance of x_{i+1} in f_i. All these make global stabilization of system (1.3) challenging and unsolvable via the existing design methods, such as feedback linearization and backstepping. Actually, one may try to apply the adding a linear integrator method to system (1.3). However, a more careful examination indicates that in the high-order case, the conventional

backstepping design results in a possibly *nonsmooth* control law. To make this point clear, consider system (1.2) with a *nonaffine* input, e.g., $u = v^3$. If the adding a linear integrator method were applied, the resulting control law would be given by $v = -(g_1(z,y) + y + L_{f_1}V)^{\frac{1}{3}}$, which, in general, is only C^0. Therefore, no more integrators can be added in the next step of the recursive design. The same difficulty would be faced in the problem of disturbance attenuation. Therefore, a new synthesis tool that goes beyond the adding a linear integrator design must be developed in order to handle inherently nonlinear systems like (1.3).

Motivated by the theory of homogeneous systems [8, 1, 9, 10, 14, 15] and the subsequent works [6, 7, 5, 25, 19, 27], we develop in this paper a *new* design tool which is based upon, but substantially extends, the adding a power integrator technique proposed in [20]. The new design method enables us to relax, in addition to the removal of feedback linearizability and affineness, the *lower-triangularity* condition assumed in [20], hence enlarging the class of nonlinear systems for which global stabilization and disturbance attenuation are still solvable by *smooth* state feedback. There are, however, some major differences from the work in [20]. One of them is that the current paper deals with the *nonstrict-triangular* system (1.3) rather than a lower-triangular system, and several technical issues arise immediately. For example, when the system is not in the lower-triangular form, the adding a power integrator technique developed in [20] encounters two obstacles, namely, that at every step it is necessary to guarantee the existence of a *smooth* virtual controller, and that a method has to be developed to carry out an iterative design for sequentially adding a power integrator. Solving these problems will be one of the main contributions of this paper. The other key difference from [20] is that a constructive solution to the global stabilization of nontriangular systems is sought, which requires subtle but important changes in the stability analysis. Finally, due to the nature of the nontriangular structure, it is more difficult to identify suitable growth conditions for global stabilization and disturbance attenuation to be solvable via smooth state feedback. As we shall see in what follows, the new *adding a power integrator* technique proposed in this paper takes full advantage of the characteristics of the dynamic system in the feedback design, particularly, using feedback to *dominate* the high-order nonlinearities of the system. This is in sharp contrast to the classical adding a linear integrator approach [3, 30], which relies upon exact cancelation of terms. As illustrated above, designs which rely on exact cancelation may result in a *nondifferentiable* controller, which makes an iterative design exceptionally difficult. Our generalized adding a power integrator design overcomes not only this difficulty but also enables us to solve, in a unified manner, the problems of global stabilization and disturbance attenuation via *static smooth* feedback for nonlinear systems (1.3) which are neither feedback linearizable nor transformable to a lower-triangular form.

A byproduct of the new adding a power integrator design is the development of sufficient conditions for the problem of global disturbance attenuation to be solvable by *smooth* state feedback, and the explicit construction of smooth state feedback control laws. The robust control results thus obtained incorporate and significantly generalize the existing disturbance attenuation results [22, 23, 12], which are only applicable to a class of feedback linearizable systems having a lower-triangular structure (1.1).

2. Global asymptotic stabilization via smooth state feedback. For the sake of simplicity, in this section we concentrate on the situation where the nonlinear system (1.3) involves no zero-dynamics, i.e., dim $z = 0$. In this case, system (1.3)

reduces to

$$\begin{aligned}
\dot{x}_1 &= x_2^{p_1} + f_1(x_1, x_2),\\
&\vdots\\
\dot{x}_{n-1} &= x_n^{p_{n-1}} + f_{n-1}(x_1, \cdots, x_n),\\
\dot{x}_n &= u^{p_n} + f_n(x_1, \cdots, x_n, u).
\end{aligned} \tag{2.1}$$

Our objective is to investigate conditions under which the problem of global asymptotic stabilization is solvable by *smooth* state feedback. It is worth noticing that, compared to the system considered in [20], the high-order system (2.1) is not in a triangular form due to the existence of x_{i+1} in $f_i(x_1, \ldots, x_{i+1})$. This new feature, together with the nonaffineness and uncontrollability of the Jacobian linearization, makes the existing nonlinear feedback design methods inapplicable. Moreover, the technique of adding a power integrator, proposed recently in [20], is only applicable to a class of *strict-triangular* systems. In what follows, we shall show how a *generalized adding a power integrator* technique can be further developed, enabling us to construct a smooth state feedback control law that renders system (2.1) globally asymptotically stable (GAS) at the equilibrium $(x_1, \ldots, x_n) = (0, \ldots, 0)$.

We begin with two assumptions that characterize the subclass of nonlinear systems (2.1).

Assumption 2.1. $p_1 \geq p_2 \geq \cdots \geq p_n \geq 1$ are *odd* integers.

Assumption 2.2. For $i = 1, \ldots, n$,

$$f_i(x_1, \ldots, x_{i+1}) = \sum_{l=0}^{p_i - 1} a_{i,l}(x_1, \ldots, x_i) x_{i+1}^l, \tag{2.2}$$

$$|a_{i,l}(x_1, \ldots, x_i)| \leq (|x_1|^{p_i - l} + \cdots + |x_i|^{p_i - l}) \gamma_{i,l}(x_1, \ldots, x_i), \quad l = 0, \ldots, p_i - 1, \tag{2.3}$$

where $x_{n+1} := u$ and $\gamma_{i,l}(\cdot)$ is a smooth and nonnegative function.

Under Assumptions 2.1–2.2, it is possible to prove the following global stabilization theorem, which is one of the main results of the paper.

THEOREM 2.3. *Suppose the nonlinear system* (2.1) *satisfies Assumptions* 2.1–2.2. *Then there exists a* C^∞ *state feedback control law* $u = u(x_1, \ldots, x_n)$ *with* $u(0, \ldots, 0) = 0$, *such that the closed-loop system is GAS at the equilibrium* $x = 0$.

The proof of Theorem 2.3 relies crucially on the following simple but useful lemma.

LEMMA 2.4. *Let* $\mathcal{X}$, $\mathcal{Y}$, *and* $\mathcal{Z}_i$, $i = 1, \ldots, l$, *be real variables. Assume that* $g_1 : \mathbb{R} \to \mathbb{R}$ *and* $g_2 : \mathbb{R}^{l+1} \to \mathbb{R}$ *are smooth mappings. Then, for any positive integers* m, n, *and real number* $N > 0$, *there exist two nonnegative smooth functions* $h_1 : \mathbb{R}^2 \to \mathbb{R}$ *and* $h_2 : \mathbb{R}^{l+1} \to \mathbb{R}$, *such that*

(i) $|\mathcal{X}^m[(\mathcal{Y} + \mathcal{X} g_1(\mathcal{X}))^n - (\mathcal{X} g_1(\mathcal{X}))^n]| \leq \frac{|\mathcal{X}|^{m+n}}{N} + |\mathcal{Y}|^{m+n} h_1(\mathcal{X}, \mathcal{Y})$,

(ii) $|\mathcal{Y}^n(\mathcal{Z}_1^m + \cdots + \mathcal{Z}_l^m + \mathcal{Y}^m) g_2(\mathcal{Z}_1, \ldots, \mathcal{Z}_l, \mathcal{Y})|$
$\leq \frac{|\mathcal{Z}_1|^{m+n} + \cdots + |\mathcal{Z}_l|^{m+n}}{N} + |\mathcal{Y}|^{m+n} h_2(\mathcal{Z}_1, \ldots, \mathcal{Z}_l, \mathcal{Y})$.

Proof. For any positive integers m, n and any real-valued smooth function $\gamma(\mathcal{X}, \mathcal{Y}) > 0$, set

$$\begin{aligned}
a &= |\mathcal{X}|^m \gamma^{\frac{m}{m+n}}, & b &= |\mathcal{Y}|^n \gamma^{-\frac{m}{m+n}},\\
p &= \tfrac{m+n}{m}, & q &= \tfrac{m+n}{n}.
\end{aligned}$$

Then, using Young's inequality

$$|ab| \leq \frac{|a|^p}{p} + \frac{|b|^q}{q}$$

yields

$$|\mathcal{X}|^m|\mathcal{Y}|^n \le \frac{m}{m+n}\gamma(\mathcal{X},\mathcal{Y})|\mathcal{X}|^{m+n} + \frac{n}{m+n}\gamma^{-\frac{m}{n}}(\mathcal{X},\mathcal{Y})|\mathcal{Y}|^{m+n}. \tag{2.4}$$

From (2.4) it is straightforward to deduce the inequality (ii) by choosing appropriate $\gamma(\cdot)$ in (2.4). To deduce (i), first expand the expression $\mathcal{X}^m[(\mathcal{Y}+\mathcal{X}g_1(\mathcal{X}))^n - (\mathcal{X}g_1(\mathcal{X}))^n]$, and apply (2.4) to each term. □

With the aid of Lemma 2.4, a constructive solution to the global stabilization problem can be obtained by repeatedly using a *generalized adding a power integrator* technique, which is based upon, but substantially extends, the design method proposed in [20].

Proof of Theorem 2.3. *Initial step.* Consider the x_1-subsystem of (2.1). Let $V_1(x_1) = \frac{x_1^2}{2}$. Then

$$\begin{aligned}\dot{V}_1(x_1) &= x_1\left[x_2^{p_1} + f_1(x_1,x_2)\right]\\ &= x_1\left[x_2^{p_1} - x_2^{*p_1} + f_1(x_1,x_2) - f_1(x_1,x_2^*)\right] + x_1\left(x_2^{*p_1} + f_1(x_1,x_2^*)\right)\end{aligned}$$

for arbitrary x_2^*. By Assumption 2.2 and Lemma 2.4, there are smooth functions $\gamma_{1,l}(x_1) \ge 0,\ \rho_{1,l}(x_1) \ge 0,\ l = 0,\dots,p_1-1$, such that

$$\begin{aligned}|f_1(x_1,x_2)| &\le \sum_{l=0}^{p_1-1}|x_1|^{p_1-l}\gamma_{1,l}(x_1)|x_2|^l \le \sum_{l=0}^{p_1-1}\left(\frac{|x_2|^{p_1}}{2p_1} + |x_1|^{p_1}\rho_{1,l}(x_1)\right)\\ &\le \frac{|x_2|^{p_1}}{2} + |x_1|^{p_1}\hat{\rho}_1(x_1), \quad \hat{\rho}_1(x_1) := \sum_{l=0}^{p_1-1}\rho_{1,l}(x_1) \ge 0.\end{aligned} \tag{2.5}$$

Using this estimate, we have

$$\dot{V}_1(x_1) \le x_1\left[x_2^{p_1} - x_2^{*p_1} + f_1(x_1,x_2) - f_1(x_1,x_2^*)\right] + x_1x_2^{*p_1} + \frac{|x_1x_2^{*p_1}|}{2} + x_1^{p_1+1}\hat{\rho}_1(x_1).$$

Clearly, the virtual smooth controller

$$x_2^*(x_1) = -x_1\alpha_1(x_1), \quad \alpha_1(x_1) := (2n + 2\hat{\rho}_1(x_1))^{\frac{1}{p_1}} > 0, \tag{2.6}$$

is such that

$$\dot{V}_1(x_1) \le -nx_1^{p_1+1} + x_1\left[x_2^{p_1} - x_2^{*p_1} + f_1(x_1,x_2) - f_1(x_1,x_2^*)\right]. \tag{2.7}$$

Inductive step. Suppose at step k there exist a C^∞ global change of coordinates

$$\xi_1 = x_1 - x_1^*, \quad \xi_2 = x_2 - x_2^*(x_1), \dots, \quad \xi_k = x_k - x_k^*(x_1,\dots,x_{k-1}) \tag{2.8}$$

(where $x_1^* = 0,\ x_m^*(x_1,\dots,x_m)$ with $x_m^*(0,\dots,0) = 0,\ m = 2,\dots,k$, are smooth functions), a Lyapunov function

$$V_k(\xi_1,\dots,\xi_k) = \sum_{j=1}^{k}\frac{\xi_j^{p_1-p_j+2}}{p_1-p_j+2},$$

and a smooth state feedback control law of the form

$$x_{k+1}^*(x_1,\dots,x_k) = -\xi_k\alpha_k(\xi_1,\dots,\xi_k), \qquad \xi_m = x_m - x_m^*,\ m = 1,\dots,k, \tag{2.9}$$

with $\alpha_k(\cdot) > 0$ being *smooth*, such that

$$\begin{aligned}\dot{V}_k(\xi_1,\dots,\xi_k) \le& -(n-k+1)(\xi_1^{p_1+1}+\cdots+\xi_k^{p_1+1})\\ &+\xi_k^{p_1-p_k+1}[x_{k+1}^{p_k}-{x_{k+1}^*}^{p_k}+f_k(x_1,\dots,x_{k+1})-f_k(x_1,\dots,x_k,x_{k+1}^*)].\end{aligned}\tag{2.10}$$

We claim that (2.10) also holds at step $k+1$. To see this, define

$$\xi_{k+1} = x_{k+1} - x_{k+1}^*(x_1,\dots,x_k).$$

Then

$$\dot{\xi}_{k+1} = x_{k+2}^{p_{k+1}} + F_{k+1}(x_1,\dots,x_{k+2}),\tag{2.11}$$

where

$$\begin{aligned}F_{k+1}(x_1,\dots,x_{k+1},x_{k+2}) &= f_{k+1}(x_1,\dots,x_{k+1},x_{k+2})\\ &\quad-\sum_{m=1}^{k}\frac{\partial x_{k+1}^*}{\partial x_m}\left(x_{m+1}^{p_m}+f_m(x_1,\dots,x_{m+1})\right)\\ &= \sum_{l=0}^{p_{k+1}-1}\tilde{a}_{k+1,l}(x_1,\dots,x_{k+1})x_{k+2}^l,\\ \tilde{a}_{k+1,l}(x_1,\dots,x_{k+1}) &= a_{k+1,l}(x_1,\dots,x_{k+1}),\qquad l=1,\dots,p_{k+1}-1,\\ \tilde{a}_{k+1,0}(x_1,\dots,x_{k+1}) &= a_{k+1,0}(x_1,\dots,x_{k+1})\\ &\quad-\sum_{m=1}^{k}\frac{\partial x_{k+1}^*}{\partial x_m}\left(x_{m+1}^{p_m}+f_m(x_1,\dots,x_{m+1})\right).\end{aligned}$$

Assumption 2.2 implies that for $l=1,\dots,p_{k+1}-1$, there are C^∞ functions $\tilde{\gamma}_{k+1,l}(\cdot)\ge 0$ such that

$$\begin{aligned}|\tilde{a}_{k+1,l}(x_1,\dots,x_{k+1})| &\le (|x_1|^{p_{k+1}-l}+\cdots+|x_{k+1}|^{p_{k+1}-l})\gamma_{k+1,l}(x_1,\dots,x_{k+1})\\ &\le (|\xi_1|^{p_{k+1}-l}+\cdots+|\xi_{k+1}|^{p_{k+1}-l})\tilde{\gamma}_{k+1,l}(\xi_1,\dots,\xi_{k+1}).\end{aligned}\tag{2.12}$$

By definition,

$$\begin{aligned}|\tilde{a}_{k+1,0}(x_1,\dots,x_{k+1})| &\le |a_{k+1,0}(x_1,\dots,x_{k+1})|\\ &\quad+\sum_{m=1}^{k}\left|\frac{\partial x_{k+1}^*}{\partial x_m}\right|\left(|x_{m+1}^{p_m}|+|f_m(x_1,\dots,x_{m+1})|\right)\\ &\le (|x_1|^{p_{k+1}}+\cdots+|x_{k+1}|^{p_{k+1}})\gamma_{k+1,0}(x_1,\dots,x_{k+1})\\ &\quad+\sum_{m=1}^{k}\left|\frac{\partial x_{k+1}^*}{\partial x_m}\right|(|x_1|^{p_m}+\cdots+|x_{m+1}|^{p_m})\,r_m(x_1,\dots,x_m).\end{aligned}$$

Note that $p_{k+1}\le p_k\le\cdots\le p_1$ and $x_m^*(0)=0,\ m=1,\dots,k$. Thus we have

$$|\tilde{a}_{k+1,0}(\cdot)| \le (|\xi_1|^{p_{k+1}}+\cdots+|\xi_{k+1}|^{p_{k+1}})\tilde{\gamma}_{k+1,0}(\xi_1,\dots,\xi_{k+1})\tag{2.13}$$

for a C^∞ function $\tilde{\gamma}_{k+1,0}(\cdot)\ge 0$. Putting (2.12) and (2.13) together, it is deduced from Lemma 2.4 that

$$|F_{k+1}(x_1,\dots,x_{k+2})| \le \sum_{l=0}^{p_{k+1}-1}(|\xi_1|^{p_{k+1}-l}+\cdots+|\xi_{k+1}|^{p_{k+1}-l})\tilde{\gamma}_{k+1,l}(\cdot)|x_{k+2}|^l$$

$$\leq \sum_{l=0}^{p_{k+1}-1} \left(\frac{|x_{k+2}|^{p_{k+1}}}{2p_{k+1}} + (|\xi_1|^{p_{k+1}} + \cdots + |\xi_{k+1}|^{p_{k+1}})\rho_{k+1,l}(\xi_1, \ldots, \xi_{k+1}) \right)$$
$$= \frac{|x_{k+2}|^{p_{k+1}}}{2} + (|\xi_1|^{p_{k+1}} + \cdots + |\xi_{k+1}|^{p_{k+1}})\bar{\rho}_{k+1}(\xi_1, \ldots, \xi_{k+1}),$$

where $\rho_{k+1,l}(\xi_1, \ldots, \xi_{k+1}) \geq 0,\ l = 0, \ldots, p_{k+1} - 1$, and $\bar{\rho}_{k+1}(\xi_1, \ldots, \xi_{k+1}) \geq 0$ are smooth functions. Consequently (by Lemma 2.4(ii)),

$$|\xi_{k+1}^{p_1 - p_{k+1}+1} F_{k+1}(x_1, \ldots, x_{k+2})| \leq \frac{|\xi_{k+1}^{p_1 - p_{k+1}+1} x_{k+2}^{p_{k+1}}|}{2} + \frac{\xi_1^{p_1+1} + \cdots + \xi_k^{p_1+1}}{2} + \xi_{k+1}^{p_1+1} \hat{\rho}_{k+1}(\xi_1, \ldots, \xi_{k+1}), \tag{2.14}$$

for a $C^\infty \hat{\rho}_{k+1}(\cdot) \geq 0$.

Now construct the smooth Lyapunov function

$$V_{k+1}(\xi_1, \ldots, \xi_{k+1}) = V_k(\xi_1, \ldots, \xi_k) + \frac{\xi_{k+1}^{p_1 - p_{k+1}+2}}{p_1 - p_{k+1} + 2}, \tag{2.15}$$

which is positive definite and proper. Clearly, it follows from (2.10) that

(2.16)

$$\begin{aligned}
\dot{V}_{k+1}(\xi_1, \ldots, \xi_{k+1}) &= \dot{V}_k(\xi_1, \ldots, \xi_k) + \xi_{k+1}^{p_1 - p_{k+1}+1} \dot{\xi}_{k+1} \\
&\leq \xi_k^{p_1 - p_k + 1}[(\xi_{k+1} + x_{k+1}^*)^{p_k} - {x_{k+1}^*}^{p_k} + f_k(x_1, \ldots, x_k, \xi_{k+1} + x_{k+1}^*) \\
&\quad - f_k(x_1, \ldots, x_k, x_{k+1}^*)] - (n - k + 1)(x_1^{p_1+1} + \cdots + \xi_k^{p_1+1}) \\
&\quad + \xi_{k+1}^{p_1 - p_{k+1}+1}[x_{k+2}^{p_{k+1}} + F_{k+1}(x_1, \ldots, x_{k+2})] \\
&\leq \xi_k^{p_1 - p_k + 1}[(\xi_{k+1} + x_{k+1}^*)^{p_k} \\
&\quad - {x_{k+1}^*}^{p_k} + f_k(x_1, \ldots, x_k, \xi_{k+1} + x_{k+1}^*) - f_k(x_1, \ldots, x_k, x_{k+1}^*)] \\
&\quad - (n - k + 1)(x_1^{p_1+1} + \cdots + \xi_k^{p_1+1}) \\
&\quad + \xi_{k+1}^{p_1 - p_{k+1}+1}[{x_{k+2}^*}^{p_{k+1}} + F_{k+1}(x_1, \ldots, x_{k+2}^*)] \\
&\quad + \xi_{k+1}^{p_1 - p_{k+1}+1}[x_{k+2}^{p_{k+1}} - {x_{k+2}^*}^{p_{k+1}} + f_{k+1}(x_1, \ldots, x_{k+2}) \\
&\quad - f_{k+1}(x_1, \ldots, x_{k+2}^*)].
\end{aligned}$$

Since $x_{k+1}^* = -\xi_k \alpha_k(\xi_1, \ldots, \xi_k)$, it can be shown that

$$|(\xi_{k+1} + x_{k+1}^*)^l - {x_{k+1}^*}^l| \leq |\xi_{k+1}|(|\xi_{k+1}|^{l-1} + |\xi_k|^{l-1})\beta_{k+1,l}(\xi_1, \ldots, \xi_{k+1}) \tag{2.17}$$

for a $C^\infty\ \beta_{k+1,l}(\cdot) \geq 0$. Hence

$$\begin{aligned}
&\xi_k^{p_1 - p_k + 1}\ [\ f_k(x_1, \ldots, x_k, \xi_{k+1} + x_{k+1}^*) - f_k(x_1, \ldots, x_k, x_{k+1}^*)] \\
&\quad \leq |\xi_{k+1}|(|\xi_1|^{p_1} + \cdots + |\xi_{k+1}|^{p_1}) \sum_{l=1}^{p_k - 1} 2k^2 \beta_{k+1,l}(\cdot)\gamma_{k+1,l}(\cdot) \\
&\quad \leq \frac{\xi_1^{p_1+1} + \cdots + \xi_k^{p_1+1}}{4} + \xi_{k+1}^{p_1+1} \tilde{\rho}_{k+1,2}(\xi_1, \ldots, \xi_{k+1})
\end{aligned} \tag{2.18}$$

for a $C^\infty\ \tilde{\rho}_{k+1,2}(\cdot) \geq 0$. This, together with Lemma 2.4, implies that there exists a *smooth* function $\tilde{\rho}_{k+1}(\cdot) \geq 0$, such that

$$\begin{aligned}(2.19)\qquad \xi_k^{p_1-p_k+1}\ &\big[(\xi_{k+1}+x_{k+1}^*)^{p_k}-x_{k+1}^{*p_k}+f_k(x_1,\dots,x_k,\xi_{k+1}+x_{k+1}^*)\\ &\quad -f_k(x_1,\dots,x_k,x_{k+1}^*)\big]\\ &\le \frac{\xi_1^{p_1+1}+\cdots+\xi_k^{p_1+1}}{2}+\xi_{k+1}^{p_1+1}\tilde\rho_{k+1}(\xi_1,\dots,\xi_{k+1}).\end{aligned}$$

Substituting (2.14) and (2.19) into (2.16) yields

$$\begin{aligned}(2.20)\quad \dot V_{k+1}(\xi_1,\dots,\xi_{k+1})\le\ &-(n-k)(\xi_1^{p_1+1}+\cdots+\xi_k^{p_1+1})\\ &+\xi_{k+1}^{p_1-p_{k+1}+1}[x_{k+2}^{p_k}-{x_{k+2}^*}^{p_k}+f_{k+1}(x_1,\dots,x_{k+2})\\ &-f_{k+1}(x_1,\dots,x_{k+1},x_{k+2}^*)]\\ &+\xi_{k+1}^{p_1-p_{k+1}+1}{x_{k+2}^*}^{p_{k+1}}+\frac{|\xi_{k+1}^{p_1-p_{k+1}+1}{x_{k+2}^*}^{p_{k+1}}|}{2}\\ &+\xi_{k+1}^{p_1+1}\left(\tilde\rho_{k+1}(\cdot)+\hat\rho_{k+1}(\cdot)\right).\end{aligned}$$

Clearly, the *smooth state* feedback control law

$$x_{k+2}^*(x_1,\dots,x_{k+1})=-\xi_{k+1}\alpha_{k+1}(\xi_1,\dots,\xi_{k+1}),\qquad \xi_m=x_m-x_m^*,\ m=1,\dots,k+1,\tag{2.21}$$

with $\alpha_{k+1}(\xi_1,\dots,\xi_{k+1})=[2n-2k+2\tilde\rho_{k+1}(\cdot)+2\hat\rho_{k+1}(\cdot)]^{\frac{1}{p_{k+1}}}>0$, is such that

$$\begin{aligned}\dot V_{k+1}(\xi_1,\dots,\xi_{k+1})\le\ &-(n-k)(\xi_1^{p_1+1}+\cdots+\xi_{k+1}^{p_1+1})\\ &+\xi_{k+1}^{p_1-p_{k+1}+1}[x_{k+2}^{p_{k+1}}-x_{k+2}^{*p_{k+1}}+f_{k+1}(x_1,\dots,x_{k+2})\\ &-f_{k+1}(x_1,\dots,x_{k+1},x_{k+2}^*)].\end{aligned}$$

This completes the inductive proof.

Repeatedly using the above inductive argument, it is easy to prove, at the nth step, that one can explicitly construct a change of coordinates $(\xi_1,\dots,\xi_n)$ of the form (2.8), a *smooth* state feedback law

$$u=x_{n+1}^*=-\xi_n\alpha_n(\xi_1,\dots,\xi_n),\qquad \alpha_n(\cdot)>0,\tag{2.22}$$

and a positive definite and proper Lyapunov function $V_n(\xi_1,\dots,\xi_n)$ of the form (2.15), such that

$$\dot V_n(\xi_1,\dots,\xi_n)\le-(\xi_1^{p_1+1}+\cdots+\xi_n^{p_1+1}).\tag{2.23}$$

Since the change of coordinates (2.8) is a global diffeomorphism, we conclude from (2.23) that system (2.1) is globally asymptotically stabilizable at the equilibrium $(x_1,\dots,x_n)=(0,\dots,0)$ by the smooth state feedback control law (2.22). □

Remark 2.5. When $a_{i,l}(x_1,\dots,x_i)=0$, $i=1,\dots,n$, and $l=1,\cdots,p_i-1$, system (2.1) reduces to a lower-triangular system considered in [20]. Therefore, Theorem 2.3 includes the previous global stabilization result for a high-order lower-triangular system as a special case. The technique used in the proof of Theorem 2.3 is a generalized version of the adding a power integrator technique proposed in [20].

We conclude this section with an example that demonstrates how a smooth controller can be designed for a two-dimensional nonlinear system in the form (2.1).

Example 2.6. Consider the planar system

$$\begin{aligned}\dot x_1&=x_2^3+x_1^2x_2+x_1^3,\\ \dot x_2&=u^3.\end{aligned}\tag{2.24}$$

Obviously, this system is of the form (2.1) with $p_1 = p_2 = 3$. Hence Assumption 2.1 holds. A simple calculation shows that $f_1(x_1, x_2) = a_{1,0}(x_1) + a_{1,1}(x_1)x_2$ with $a_{1,0}(x_1) = x_1^3$ and $a_{1,1}(x_1) = x_1^2$. Clearly,

$$|a_{1,0}(x_1)| \le |x_1|^3, \qquad |a_{1,1}(x_1)| \le |x_1|^2,$$

which implies that Assumption 2.2 is also fulfilled. By Theorem 2.3, there exists a smooth controller that renders the equilibrium $(x_1, x_2) = (0, 0)$ of (2.24) GAS. The controller can be explicitly constructed by following the design procedure in Theorem 2.3.

First, choose $V_1 = \frac{x_1^2}{2}$. Then the virtual controller

$$x_2^* = -\sqrt[3]{\frac{11}{2}}x_1 \tag{2.25}$$

renders

$$\dot{V}_1 \le -2x_1^4 + x_1(x_2^3 - x_2^{*3}) + x_1^3(x_2 - x_2^*). \tag{2.26}$$

Next consider the Lyapunov function

$$V_2 = \frac{x_1^2}{2} + \frac{\left(x_2 + \sqrt[3]{\frac{11}{2}}x_1\right)^2}{2}.$$

Using an argument similar to the proof of Theorem 2.3, it is easy to see that the smooth controller

$$u = -38\left(x_2 + \sqrt[3]{\frac{11}{2}}x_1\right)$$

is such that

$$\dot{V}_2 \le -x_1^4 - \left(x_2 + \sqrt[3]{\frac{11}{2}}x_1\right)^4,$$

thus stabilizing the system (2.24) in the large.

3. Necessity of Assumptions 2.1–2.2. In this section, we discuss to what extent the sufficient conditions Assumptions 2.1 and 2.2 given in the previous section are *necessary* for global stabilization via *smooth* state feedback. We give examples to illustrate that smooth stabilization of the system (2.1) is not possible if either Assumption 2.1 or Assumption 2.2 fails to be satisfied.

First, we show that Assumption 2.1 is somewhat necessary for solving the stabilization problem of high-order nonlinear systems (2.1).

Example 3.1. Consider the two-dimensional nonlinear system

$$\begin{aligned} \dot{x}_1 &= x_2^3 + x_1^2x_2 + x_1^3, \\ \dot{x}_2 &= u^p. \end{aligned} \tag{3.1}$$

When $p = 3$ or $p = 1$, the system satisfies the hypotheses Assumptions 2.1 and 2.2. By Theorem 2.3, there is a smooth static state feedback control law that globally

asymptotically stabilizes (3.1) at $(x_1, x_2) = (0, 0)$. An explicit smooth controller has been given in Example 2.6.

If $p \geq 5$ is an odd integer, system (3.1) satisfies Assumption 2.2 but not Assumption 2.1. Then one can prove the following claim: system (3.1) with $p = 5$ cannot be stabilized, even locally, by any *smooth static* state feedback.

Proof. The claim is proved by contradiction. Suppose there is a smooth state feedback $\alpha(x)$, with $\alpha(0) = 0$ and $x = (x_1, x_2)^T$, such that system (3.1) is locally asymptotically stable at $(0, 0)$. Then $\forall \varepsilon \in (0, 1)$, there is a $\delta \in (0, \varepsilon)$ such that $\|x(0)\| < \delta \Rightarrow \|x(t, x(0))\| < \varepsilon \ \forall t \geq 0$.

Now consider the domain $\Omega_\varepsilon = \{x| \ \|x\| \leq \varepsilon, x_1 \geq 0, x_1 \geq -2x_2\}$ shown in Figure 3.1. Choose an initial condition $(x_1(0), x_2(0))$ in the interior of Ω_ε and $0 < \|x(0)\| < \delta$. Let $x(t, x(0))$ be a trajectory starting from $x(0)$. In the interior of Ω_ε, $1 > x_1 > -2x_2$ and $x_1 > 0$, which implies $\dot{x}_1 > 0$, $(x_1, x_2) \in$ interior of Ω_ε.

In fact, when $x_2 \geq 0$ it is clear that $\dot{x}_1 > 0$. In the case when $x_2 < 0$, a direct calculation gives

$$\dot{x}_1 = x_2^3 + x_1^2 x_2 + x_1^3 \geq x_2^3 + x_1^3 - \frac{2x_1^3}{3} - \frac{|x_2|^3}{3} = \frac{4x_2^3 + x_1^3}{3} > \frac{(2x_2)^3 + x_1^3}{3} \geq 0.$$

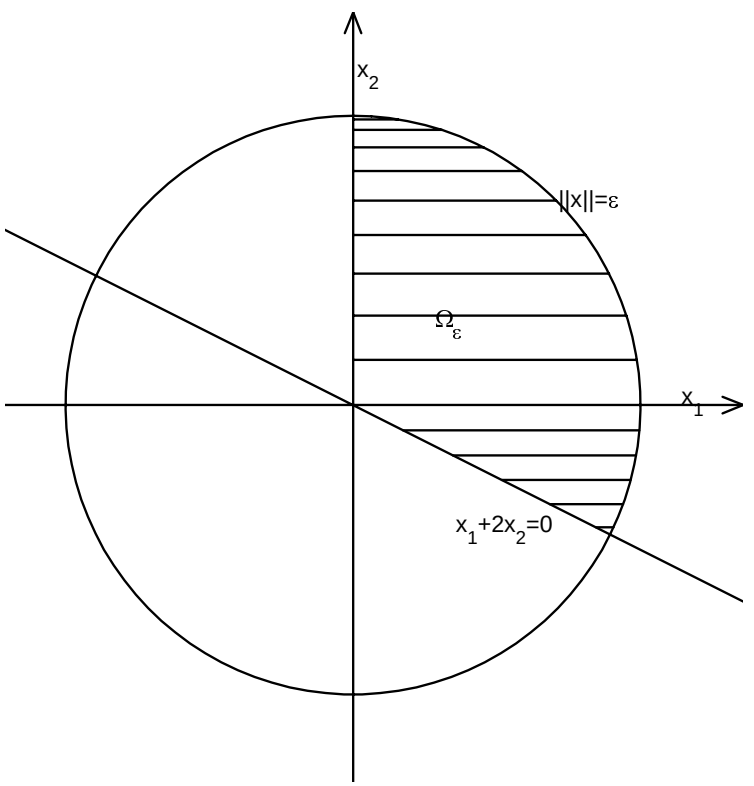


FIG. 3.1. *The domain of* Ω_ε.

Therefore, $x(t)$ cannot remain in Ω_ε forever if the origin is asymptotically stabilizable. In other words, the trajectory of (3.1) must cross the boundaries of Ω_ε in finite time. By assumption, $\|x(t, x(0))\| < \varepsilon \ \ \forall t \geq 0$. Thus the trajectory cannot cross the boundary $\Omega_\varepsilon \bigcap \{x| \ \|x\| = \varepsilon\}$. Note that the trajectory cannot cross

$\Omega_\varepsilon \bigcap \{x|\ x_1 = 0, x_2 \neq 0\}$ as on that boundary $\dot{x}_1 > 0$. In conclusion, the trajectory $x(t, x(0))$ of (3.1) must cross the line $\Omega_\varepsilon \bigcap \{x|\ x_1 + 2x_2 = 0\}$. In other words, there exists a point $x^* = (x_1^*, x_2^*)$ on $\Omega_\varepsilon \bigcap \{x|\ x_1 + 2x_2 = 0\}$ such that

$$-2\alpha^5(x^*) \geq {x_1^*}^3 + {x_1^*}^2 x_2^* + {x_2^*}^3, \qquad x_1^* = -2x_2^*, \qquad \|x^*\| < \varepsilon. \tag{3.2}$$

Since $x_1(0) > 0$ and $\dot{x}_1(t) > 0$ in the interior of Ω_ε, $x_1^* \geq x_1(0) > 0$.

From the argument above it is concluded that for every $\varepsilon_i \in (0, 1)$, there exists a $x_1^* = \theta_i > 0$ such that (3.2) holds. Choose $\varepsilon_1 = \varepsilon$ and $\varepsilon_i = \frac{\theta_{i-1}}{2}$, $i = 2, 3, \ldots$. In this way, we have generated a monotone, strictly decreasing sequence $\{\theta_i > 0\}_{i=1}^{\infty}$, having the properties that $\theta_i \to 0$ as $i \to \infty$ and

$$-2\alpha^5\left(\theta_i, \frac{-\theta_i}{2}\right) \geq \theta_i^3 + \theta_i^2\left(\frac{-\theta_i}{2}\right) + \left(\frac{-\theta_i}{2}\right)^3 = \frac{3\theta_i^3}{8} > 0, \quad i = 1, 2, \ldots.$$

This implies that

$$\frac{|\alpha(\theta_i, \frac{-\theta_i}{2}) - \alpha(0,0)|}{\|(\theta_i, \frac{-\theta_i}{2}) - (0,0)\|} = \frac{|\alpha(\theta_i, \frac{-\theta_i}{2})|}{\sqrt{\theta_i^2 + \frac{\theta_i^2}{4}}} \geq \frac{\sqrt[5]{6}}{\sqrt{5}}\theta_i^{-\frac{2}{5}} \to \infty \quad \text{as} \quad i \to \infty.$$

Therefore, the controller $u = \alpha(x)$ is not differentiable at $(0, 0)$, which is a contradiction. □

The next example demonstrates that the condition (2.3) is also crucial for the problem of smooth stabilization to be solvable.

Example 3.2. Consider the scalar system

$$\dot{x} = u^5 + u^2 x^p, \tag{3.3}$$

where p is a nonnegative integer.

This system always satisfies Assumption 2.1 and (2.2). However, if $p < 3$, $|a_{1,2}(x)| = |x^p|$ cannot be bounded by $|x|^3\rho(x)$, where $\rho(x) \geq 0$ is a smooth function. Thus (2.3) is not satisfied. In this case, it can be shown that there is no smooth controller, making the equilibrium $x = 0$ of system (3.3) locally asymptotically stable. To this end, suppose that there exists a stabilizing smooth controller $u(x)$ with $u(0) = 0$ and $u(x) \neq 0\ \forall x \in N - \{0\}$, where N is an open neighborhood of $x = 0$. By smoothness, $u(x) = xh(x)$ for a smooth function $h(x)$. Let δ be a real number in $(0, 1)$. Then, for $0 < x \leq \delta$ and sufficiently small δ, $x^p + u^3 = x^p + x^3h^3(x) > 0$ and $u(x) \neq 0$. This implies that for $V(x) = x$,

$$\dot{V} = u^2(x^p + u^3) > 0, \qquad x \in \{x \mid V(x) > 0, |x| \leq \delta\}.$$

By the Lyapunov instability theorem, the closed-loop system is unstable at the equilibrium $x = 0$.

When $p \geq 3$, (2.3) holds. By Theorem 2.3, there does exist a smooth state stabilizer. A smooth controller can be constructed as

$$u = -x\left(\frac{5}{3} + (1 + x^2)^{\frac{5(p-3)}{6}}\right)^{1/5}. \tag{3.4}$$

It is straightforward to verify that the time derivative of $V = x^2/2$ along the trajectory of the closed-loop system (3.3)–(3.4) satisfies

$$\dot{V} \leq -x^6.$$

Finally, the necessity of hypothesis (2.2) can be illustrated by the simple example below.

Example 3.3. Consider the planar system

$$\begin{aligned} \dot{x}_1 &= x_2^3 + x_2^3 x_1, \\ \dot{x}_2 &= u^3, \end{aligned} \tag{3.5}$$

for which Assumption 2.1 and (2.3) are fulfilled. However, (2.2) is violated as the order of x_2 in the decomposition term is equal to $p_1 = 3$. We claim that the system cannot be GAS by smooth state feedback, for observe that $x_1 = -1$ makes the first component of the vector field vanish. Thus, regardless of the choice of $u = u(x_1, x_2)$, the vertical line $x_1 = -1$ in the phase-plane is an invariant set which cannot be crossed by any trajectory. Indeed, for any $x_1(0) \leq -1$ there exists no feedback controller that can steer $x_1(t)$ to zero as $t \to \infty$.

4. Synthesis of cascade systems.

A. Global stabilization by smooth feedback. Now we show how the smooth feedback stabilization result developed in section 2 can be extended to the class of nonlinear systems of the form (1.3). We begin by listing three assumptions.

Assumption 4.1. $p_1 \geq p_2 \geq \cdots \geq p_r \geq 1$ are odd integers.

Assumption 4.2. For $i = 1, \ldots, r$,

$$f_i(z, x_1, \ldots, x_i, x_{i+1}) = \sum_{j=0}^{p_i - 1} C_{i,j}(z, x_1, \ldots, x_i) x_{i+1}^j,$$

$$|C_{i,j}(z, x_1, \ldots, x_i)| \leq (\|z\|^{p_i - j} + |x_1|^{p_i - j} + \cdots + |x_i|^{p_i - j}) \rho_{i,j}(z, x_1, \ldots, x_i), \tag{4.1}$$

where $x_{r+1} := u$ and $\rho_{i,j}(\cdot) \geq 0$ is a known smooth function.

Assumption 4.3. Assume that there is a real-valued, nonnegative smooth function $\gamma_0(z, x_1)$ such that

$$\|f_0(z, x_1)\| \leq (\|z\|^{p_1} + |x_1|^{p_1}) \gamma_0(z, x_1).$$

Using Assumptions 4.1–4.3, it is possible to prove the following global stabilization result.

THEOREM 4.4. *Consider the nonlinear system* (1.3) *satisfying Assumptions* 4.1–4.3. *Suppose there exist a smooth function* $x_1 = v^*(z)$ *with* $v^*(0) = 0$, *and a smooth Lyapunov function* $V(z)$, *which is positive definite and proper, such that*

$$\frac{\partial V}{\partial z} f_0(z, v^*(z)) \leq -\|z\|^{p_0 + 1} W(z), \qquad W(z) > 0, \quad \forall z, \tag{4.2}$$

$$\left\| \frac{\partial V}{\partial z} \frac{\partial f_0(z, x_1)}{\partial x_1} \right\| \leq (\|z\|^{p_0} + |x_1|^{p_0}) \gamma_1(z, x_1), \quad \gamma_1(\cdot) \geq 0, \tag{4.3}$$

where $p_0 \geq p_1$ *is an odd integer and* $W(z)$ *and* $\gamma_1(\cdot)$ *are smooth functions. Then system* (1.3) *is globally asymptotically stabilizable at* $(z, x_1, \ldots, x_r) = 0$ *by a smooth static controller* $u = u(z, x_1, \ldots, x_r)$ *with* $u(0, 0, \ldots, 0) = 0$.

Proof. The theorem can be proved in a fashion similar to the proof of Theorem 2.3 by using the generalized adding a power integrator technique. The major difference lies in the first step. For this reason, in what follows we consider only the case where $r = 1$ in system (1.3), i.e.,

$$\begin{aligned} \dot{z} &= f_0(z, x_1), \\ \dot{x}_1 &= u^{p_1} + f_1(z, x_1, u). \end{aligned} \tag{4.4}$$

This system can be rewritten as

$$\begin{aligned} \dot{z} &= f_0(z, v^*(z)) + \Delta(z, x_1), \\ \dot{x}_1 &= u^{p_1} + f_1(z, x_1, u) \end{aligned} \tag{4.5}$$

with

$$\Delta(z, x_1) = f_0(z, x_1) - f_0(z, v^*(z)). \tag{4.6}$$

For system (4.5), consider the Lyapunov function

$$U(z, x_1) = V(z) + \frac{(x_1 - v^*(z))^{p_0 - p_1 + 2}}{p_0 - p_1 + 2},$$

which is positive definite and proper. Then a direct calculation gives

$$\begin{aligned} \dot{U}(z, x_1)|_{(4.5)} \le & -\|z\|^{p_0+1} W(z) + \frac{\partial V(z)}{\partial z} \Delta(z, x_1) + \xi^{p_0 - p_1 + 1} u^{p_1} \\ & + \xi^{p_0 - p_1 + 1} \phi(z, \xi, u), \end{aligned} \tag{4.7}$$

where

$$\xi = x_1 - v^*(z) \equiv x_1 - z^T \alpha(z) \ \text{ for a } C^\infty, \quad \alpha(z) \in R^m,$$

$$\phi(z, \xi, u) = f_1(z, \xi + v^*(z), u) - \frac{\partial v^*}{\partial z} f_0(z, \xi + v^*(z)).$$

Using Assumption 4.2 and Lemma 2.4(i), we have

$$\begin{aligned} |\phi(z, \xi, u)| \le & \sum_{j=0}^{p_1 - 1} (\|z\|^{p_1 - j} + |\xi + v^*|^{p_1 - j}) \rho_{1,j}(z, \xi + v^*(z)) |u^j| \\ & + (\|z\|^{p_1} + |\xi + v^*|^{p_1}) \left| \frac{\partial v^*}{\partial z} \right| \gamma_0(z, \xi + v^*) \\ \le & \sum_{j=0}^{p_1 - 1} \left\| \begin{matrix} z \\ \xi \end{matrix} \right\|^{p_1 - j} |u^j| \hat{\rho}_{1,j}(z, \xi) + (\|z\|^{p_1} + |\xi + v^*|^{p_1}) \left| \frac{\partial v^*}{\partial z} \right| \gamma_0(z, \xi + v^*) \\ \le & \frac{|u|^{p_1}}{2} + (\|z\|^{p_1} + |\xi|^{p_1}) \gamma(z, \xi) \end{aligned} \tag{4.8}$$

for a smooth $\gamma(z, \xi) \ge 0$. By (4.8) and Lemma 2.4, there exists a smooth nonnegative function $\rho_1(z, \xi)$ such that

$$|\xi^{p_0 - p_1 + 1} \phi(z, \xi, u)| \le \frac{\|z\|^{p_0+1} W(z)}{4} + \xi^{p_0+1} \rho_1(z, \xi) + \frac{1}{2} |\xi|^{p_0 - p_1 + 1} |u|^{p_1}. \tag{4.9}$$

Using (4.3), (4.6), and the Taylor expansion formula, one has

$$\begin{aligned} \left| \frac{\partial V(z)}{\partial z} \Delta(z, x_1) \right| &\equiv \left| \frac{\partial V}{\partial z} \left(\int_0^1 \left. \frac{\partial f_0}{\partial x_1} \right|_{x_1 = v^* + \lambda\xi} d\lambda \right) \xi \right| \\ &\le |\xi| \int_0^1 (\|z\|^{p_0} + |v^* + \lambda\xi|^{p_0}) \gamma_1(\cdot) d\lambda \\ &\le |\xi| \int_0^1 (\|z\|^{p_0} + |\lambda\xi|^{p_0}) \tilde{\gamma}(z, v^* + \lambda\xi) d\lambda \\ &\le |\xi| \, (\|z\|^{p_0} + |\xi|^{p_0}) \int_0^1 \tilde{\gamma}(z, v^* + \lambda\xi) d\lambda, \end{aligned}$$

where $\tilde{\gamma}(\cdot) \geq 0$ is a smooth function.

The last inequality, together with Lemma 2.4(i), implies

$$\left|\frac{\partial V(z)}{\partial z}\Delta(z,x_1)\right| \leq \frac{\|z\|^{p_0+1}W(z)}{4} + \xi^{p_0+1}\rho_2(z,\xi) \tag{4.10}$$

for an appropriate nonnegative smooth function $\rho_2(z,\xi)$.

Substituting (4.9) and (4.10) into (4.7), we arrive at

$$\begin{aligned}\dot{U}(z,x_1)|_{(4.5)} \leq & -\frac{\|z\|^{p_0+1}W(z)}{2} + \xi^{p_0-p_1+1}u^{p_1} \\ & +\frac{1}{2}|\xi|^{p_0-p_1+1}|u|^{p_1} + \xi^{p_0+1}[\rho_1(z,\xi)+\rho_2(z,\xi)].\end{aligned} \tag{4.11}$$

Choose a smooth controller of the form

$$u = -2\xi\left[\frac{W(z)}{2} + \rho_1(z,\xi) + \rho_2(z,\xi)\right]^{\frac{1}{p_1}}. \tag{4.12}$$

Clearly, the proposed controller renders

$$\dot{U}(z,x_1)|_{(4.5)} \leq -\left(\|z\|^{p_0+1} + \xi^{p_0+1}\right)\frac{W(z)}{2}.$$

Thus Theorem 4.4 is true when $r = 1$.

If $r > 1$, Theorem 4.4 can be proved by repeatedly using the technique of adding a power integrator, as done in the proof of Theorem 2.3. □

In the remainder of this section, we use several examples to illustrate some of the interesting features of Theorem 4.4 and its applications to the smooth feedback stabilization problem. First, we illustrate how Theorem 4.4 can be used to deal with a general system which is not in a lower-triangular form.

Example 4.5. Consider the nonlinear system

$$\begin{aligned}\dot{z}_1 &= z_2^3 - 2z_1x_1^2, \\ \dot{z}_2 &= x_1^5 - z_2^5, \\ \dot{x}_1 &= u^3 + u^2\sin x_1 + uz_1z_2,\end{aligned} \tag{4.13}$$

which is not feedback linearizable, as the linearized system is not controllable. Moreover, the system is neither in a lower-triangular form nor affine in u; thus it cannot be handled by [20].

On the other hand, the system is of the form (1.3). The zero-dynamics (z_1, z_2) of (4.13) is GAS by $x_1 = v^*(z_1,z_2) = -z_1$. In fact, let $V(z_1,z_2) = \frac{z_1^4+2z_2^2}{4}$. A direct calculation gives

$$\dot{V}|_{x_1=v^*(z_1,z_2)} \leq -\frac{1}{3}(z_1^6 + z_2^6).$$

For $p_0 = 5$, $p_1 = 3$, it is easy to see that (4.2) and (4.3) are satisfied. Since $f_1(z_1,z_2,x_1,u) = u^3 + u^2\sin x_1 + uz_1z_2$, Assumption 4.2 also holds. Therefore, by Theorem 4.4 system (4.13) is GAS by smooth state feedback.

It has been known that achieving global stabilizability for a nonlinear system whose zero-dynamics depend on more than one component of x (i.e., x_1) is usually

difficult [29, 4]. It is of interest to point out that this may not be the case when dealing with high-order nonlinear systems. In fact, one can extend Theorem 4.4 to a more general class of nonlinear systems (than (1.3)), in which the zero-dynamics $\dot{z} = f_0(z, x_1)$ are replaced by

$$\dot{z} = f_0(z, x_1) + \sum_{j=1}^{p_1-1} a_j(z, x_1) x_2^j. \tag{4.14}$$

COROLLARY 4.6. *Consider the nonlinear system* (1.3)*, where the z-equation is replaced by* (4.14)*. Suppose all the hypotheses of Theorem* 4.4 *are fulfilled. In addition, assume the following:*

$$a_j(z, v^*(z)) = 0, \quad \left|\frac{\partial a_j(z, x_1)}{\partial x_1}\right| \le (\|z\|^{p_1-1-j} + |x_1|^{p_1-1-j})\hat{r}_j(z, x_1), \quad j = 1, \dots, p_1 - 1, \tag{4.15}$$

where $\hat{r}_j(z, x_1) \ge 0$ *is smooth. Then the system is globally asymptotically stabilizable by smooth state feedback.*

The proof of this result is strongly reminiscent of the proof of Theorem 4.4 and is therefore left to the reader as an exercise. The example below demonstrates an appealing application of Corollary 4.6 to a high-order nonlinear system that is not in the form (1.3).

Example 4.7. The nonlinear system

$$\begin{aligned} \dot{z} &= x_2^2 z + x_2^2 x_1 + x_1^3, \\ \dot{x}_1 &= x_2^3, \\ \dot{x}_2 &= u^3 \end{aligned} \tag{4.16}$$

satisfies the assumptions of Corollary 4.6, and therefore a globally stabilizing *smooth* controller exists.

To construct such a smooth feedback law, let $\xi = x_1 + z$ and $V = \frac{z^2+\xi^2}{2}$. Then the time derivative of V along the trajectory of (z, x_1)-subsystem of (4.16) is

$$\begin{aligned} \dot{V} &= -z^4 + z((\xi - z)^3 + z^3) + z\xi x_2^2 + \xi(x_2^3 + (\xi - z)^3 + x_2^2\xi) \\ &= -z^4 - 2\xi^3 z + 2\xi z^3 + \xi x_2^3 + z\xi x_2^2 + x_2^2\xi^2 + \xi^4. \end{aligned}$$

By Lemma 2.4(i), it is clear that

$$\begin{aligned} x_2^2\xi^2 &\le |\xi| \left(\frac{4}{3}|\xi|^3 + \frac{1}{3}|x_2|^3\right), \\ |\xi z x_2^2| &\le |\xi| \left(\frac{4}{3}|z|^3 + \frac{1}{3}|x_2|^3\right). \end{aligned}$$

Hence

$$\dot{V} \le -z^4 + 2|z\xi^3| + \frac{10}{3}|\xi z^3| + \xi x_2^3 + \frac{2}{3}|\xi x_2^3| + \frac{7}{3}\xi^4. \tag{4.17}$$

Note that

$$\begin{aligned} 2|z\xi^3| &\le \frac{1}{8}z^4 + 3\xi^4, \\ \frac{10}{3}|\xi z^3| &\le \frac{5}{8}z^4 + \frac{160}{3}\xi^4. \end{aligned}$$

Substituting the estimates above into (4.17), we have

$$\dot{V} \leq -\frac{1}{4}z^4 + \xi x_2^3 + \frac{2}{3}|\xi x_2^3| + \xi^4\left(\frac{7}{3} + 3 + \frac{160}{3}\right).$$

Obviously, the virtual smooth controller

$$x_2 = x_2^* = -(177)^{1/3}\xi$$

renders

$$\dot{V}|_{x_2=x_2^*} \leq -\frac{1}{4}(z^4 + \xi^4).$$

Therefore, a globally stabilizing smooth controller for (4.16) can be derived directly by adding one more power integrator $\dot{x}_2 = u^3$.

B. Disturbance attenuation with stability. So far we have studied the global stabilization problem for nonlinear systems in the absence of disturbances. When a nonlinear system under consideration involves an undesired input or disturbance that is additive to the system, it is no longer possible nor meaningful to asymptotically stabilize the system, due to the shift of the equilibrium of the system caused by additive external disturbances. Therefore, in the presence of external disturbances, a more realistic problem to be addressed is the so-called problem of disturbance attenuation. That is, the problem of seeking a *smooth* state feedback control law so that the influence of the disturbance on the output is as small as possible. Of course, such a feedback law should also guarantee global asymptotic stability of the system in the absence of disturbances.

The problem of disturbance attenuation of this type was first formulated, in terms of L_2-gain, for linear systems in [32]. For nonlinear systems, the problem has been one of the important subjects in nonlinear control theory over the last decade; see, for instance, the books [11, 26, 31]. To the best of our knowledge, the first work on the disturbance attenuation problem was reported in [22]. The solution presented in [22] is characterized in terms of the L_∞ induced norm from the disturbance inputs to the outputs. A drawback of [22] is that it did not consider the *internal stability* which is crucial for a meaningful application or a practical implementation. The stability issue was addressed later in [23], where a novel design technique based on adding a linear integrator [3, 30] was presented, resulting in an elegant solution to the global disturbance attenuation problem with internal stability, for minimum-phase nonlinear systems of the form (1.1). The disturbance attenuation result obtained in [23] has been generalized to a larger class of minimum-phase and nonminimum-phase nonlinear systems [3, 12, 13]. Notice that all the above-mentioned papers assume that the controlled plants are feedback linearizable and have a lower-triangular structure. As a result, most of the disturbance attenuation results in the literature can only be applied to affine systems that are globally diffeomorphic to (1.1).

In a recent work [28], the disturbance attenuation problem has been studied for a class of lower-triangular systems which are neither feedback linearizable nor affine in the control input. In what follows, we show how the result of [28] can be extended, with the aid of the *generalized adding a power integrator* technique, to the *nonstrict-triangular* cascade system

$$
\begin{aligned}
\dot{z} &= f_0(z, x_1) + \phi_0(z, x_1)w, \\
\dot{x}_1 &= x_2^{p_1} + f_1(z, x_1, x_2) + \phi_1(z, x_1)w, \\
&\vdots \\
\dot{x}_r &= u^{p_r} + f_r(z, x_1, \ldots, x_r, u) + \phi_r(z, x_1, \ldots, x_r)w, \\
y &= h(z, x_1),
\end{aligned} \tag{4.18}
$$

where $w \in \mathbb{R}^s$ represents a disturbance signal, the functions $f_i(\cdot)$, $i = 0, 1, \ldots, r$, are defined as in (1.3), $\phi_0 : \mathbb{R}^{n-r+1} \to \mathbb{R}^{1\times s}$, $\phi_i : \mathbb{R}^{n-r+i} \to \mathbb{R}^{1\times s}$, $i = 1, \ldots, r$, are smooth functions which do not necessarily vanish at the origin, and $h : \mathbb{R}^{n-r+1} \to \mathbb{R}$ is a smooth output function with $h(0, 0) = 0$.

Our goal is to construct, under appropriate conditions, a *smooth* state feedback control law

$$u = u(x) \qquad \text{with} \quad u(0) = 0, \tag{4.19}$$

such that for any real number $\gamma > 0$, the closed-loop system (4.18)–(4.19) satisfies the following:

(1) When $w = 0$, the closed-loop system (4.18)–(4.19) is GAS at $x = 0$.
(2) For every disturbance $w(t) \in L_2$, the response of the closed-loop system (4.18)–(4.19) starting from the initial state $x(0) = 0$ is such that

$$\int_0^t |y(s)|^{2p} ds \le \gamma^2 \int_0^t \|w(s)\|^2 ds \quad \forall t \ge 0, \quad \text{for some integer } p \ge 1. \tag{4.20}$$

□

Here, the problem of disturbance attenuation with internal stability is formulated in terms of an L_2–L_{2p}-gain (rather than a conventional L_2-gain) for cascade nonlinear systems (4.18). This is due to the consideration that the standard L_2-gain formulation is usually *not well-posed* in the case of high-order nonlinear systems. As a matter of fact, for cascade nonlinear system (4.18) an L_2 input signal may not necessarily produce an L_2 output signal. As shown in [28], an L_2 input signal is likely to yield an L_{2p} output signal, where p may be varying and depending on a structure of the system or the output of the system. In the case of feedback linearizable systems, $p = 1$. Then, the formulation above reduces to the standard L_2-gain characterization.

The following theorem gives a constructive solution to the disturbance attenuation problem characterized by an L_2–L_{2p}-gain for the cascade nonlinear system (4.18).

THEOREM 4.8. *Consider the nonlinear system* (4.18). *Suppose there exists a smooth Lyapunov function $V(z)$, which is positive definite and proper, such that*

$$
\begin{aligned}
&\left\| \frac{\partial V}{\partial z} \phi_0(z, x_1) \right\| \le (\|z\|^{p_0} + |x_1|^{p_0}) r_0(z, x_1), \qquad r_0(z, x_1) \ge 0, \\
&\left\| \frac{\partial V}{\partial z} \frac{\partial f_0(z, x_1)}{\partial x_1} \right\| \le (\|z\|^{2p_0 - 1} + |x_1|^{2p_0 - 1}) \tilde{r}_0(z, x_1), \qquad \tilde{r}_0(z, x_1) \ge 0,
\end{aligned} \tag{4.21}
$$

and

$$\frac{\partial V}{\partial z} f_0(z, v^*(z)) + \frac{1}{4\gamma^2} \left(\frac{\partial V}{\partial z} \phi_0(z, v^*(z)) \right)^2 + h^{2p_0}(z, v^*(z)) \le -\|z\|^{2p_0} W(z) \tag{4.22}$$

for a C^∞ $W(z) > 0$, where $p_0 \ge p_1$ is an odd integer, $v^*(z)$ is a smooth real-valued function with $v^*(0) = 0$, and the functions $r_0(\cdot)$, $\tilde{r}_0(\cdot)$, and $W(z)$ are smooth. Then

the disturbance attenuation problem with internal stability is solvable via smooth state feedback, if the hypotheses in Assumptions 4.1– 4.3 *hold.*

The proof of this result again relies on the *generalized adding a power integrator* technique. To begin with, we first introduce an important technical lemma which shows that the Hamilton–Jacobi–Isaacs (HJI) partial differential inequality (4.22) arising from the disturbance attenuation problem implies a dissipation inequality at the first step and can be propagated through adding a power integrator at each step.

LEMMA 4.9. *Consider a nonlinear system described by equations of the form*

$$\begin{cases} \dot{z} = f_0(z,\zeta) + \phi_0(z,\zeta)w, \\ \dot{\zeta} = u^{p_1} + f_1(z,\zeta,u) + \phi_1(z,\zeta)w, \\ y = h(z,\zeta), \end{cases} \tag{4.23}$$

in which $z \in \mathbb{R}^{n-r}$ *and* $\zeta \in \mathbb{R}$. *Suppose there exists a smooth Lyapunov function* $V(z)$, *which is positive definite and proper, such that*

$$\left\| \frac{\partial V}{\partial z}\phi_0(z,\zeta) \right\| \le (\|z\|^{p_0} + |\zeta|^{p_0}) r_0(z,\zeta), \qquad r_0(z,\zeta) \ge 0, \tag{4.24}$$

$$\left\| \frac{\partial V}{\partial z}\frac{\partial f_0(z,\zeta)}{\partial \zeta} \right\| \le (\|z\|^{2p_0-1} + |\zeta|^{2p_0-1}) \tilde{r}_0(z,\zeta), \qquad \tilde{r}_0(z,\zeta) \ge 0, \tag{4.25}$$

and

$$\frac{\partial V}{\partial z} f_0(z, v^*(z)) + \frac{1}{4\gamma^2}\left(\frac{\partial V}{\partial z}\phi_0(z, v^*(z))\right)^2 + h^{2p_0}(z, v^*(z)) \le -\|z\|^{2p_0} W(z), \;\; W(z) > 0, \tag{4.26}$$

where $p_0 \ge p_1$ *is an odd integer,* $v^*(z)$ *is a smooth function with* $v^*(0) = 0$, *and the functions* $r_0(\cdot)$, $\tilde{r}_0(\cdot)$, *and* $W(z)$ *are smooth. Under Assumptions* 4.2 *and* 4.3 *(with* $r = 1$, $\zeta = x_1$*), there are a smooth state feedback law* $u(z,\zeta)$ *with* $u(0,0) = 0$ *and a smooth Lyapunov function* $U(z,\zeta)$, *which is positive definite and proper, such that*

$$L_F U(z,\zeta) + \frac{1}{4\gamma^2}\left(L_\Phi U(z,\zeta)\right)^2 + h^{2p_0}(z,\zeta) \le -\left(\|z\|^{2p_0} + \|\zeta\|^{2p_0}\right)\mathcal{W}(z,\zeta)$$

for a C^∞ $\mathcal{W}(z,\zeta) > 0$, *where*

$$F(z,\zeta) = \begin{pmatrix} f_0(z,\zeta) \\ u^{p_1}(z,\zeta) + f_1(z,\zeta,u(z,\zeta)) \end{pmatrix}, \qquad \Phi(z,\zeta) = \begin{pmatrix} \phi_0(z,\zeta) \\ \phi_1(z,\zeta) \end{pmatrix}.$$

The proof of Lemma 4.9 is analogous to that of Theorem 4.4 with an appropriate modification. However, a different Lyapunov function $U(z,\zeta) = V(z) + \frac{(\zeta - v^*(z))^{2p_0-p_1+1}}{2p_0-p_1+1}$ must be used to carry out the proof.

Proof of Theorem 4.8. By Lemma 4.9, it is clear that Theorem 4.8 holds when $r = 1$. In the case where $r > 1$, Theorem 4.8 can be proved by repeatedly using Lemma 4.9. In fact, it can be easily verified that all the conditions of Lemma 4.9 are fulfilled when adding a power integrator each time. At the last step, one can prove that there are a smooth Lyapunov function $\tilde{U}(z,x)$, which is positive definite and proper, and a smooth state feedback law $u(z,x)$ with $u(0,0) = 0$, such that system (4.18) satisfies

$$L_{\tilde{F}}\tilde{U}(z,x) + \frac{1}{4\gamma^2}\left(L_{\tilde{\Phi}}\tilde{U}(z,x)\right)^2 + h^{2p_0}(z,x_1) \le -\left(\|z\|^{2p_0} + \|x\|^{2p_0}\right)\tilde{W}(z,x)$$

for a C^∞ $\tilde{W}(z,x) > 0$, where

$$\tilde{F}(z,x) = \begin{pmatrix} f_0(z,x_1) \\ x_2^{p_1} + f_1(z,x_1,x_2) \\ \vdots \\ u^{p_r}(z,x) + f_r(z,x_1,\ldots,x_r,u(z,x)) \end{pmatrix},$$

$$\tilde{\Phi}(z,\zeta) = \begin{pmatrix} \phi_0(z,x_1) \\ \phi_1(z,x_1) \\ \vdots \\ \phi_r(z,x_1,\ldots,x_r) \end{pmatrix}.$$

It is straightforward to prove that the HJI inequality above implies

$$\dot{\tilde{U}} + y^{2p_0} - \gamma^2 w^2 \le -\left(\|z\|^{2p_0} + \|x\|^{2p_0}\right)\tilde{W}(z,x). \tag{4.27}$$

Therefore, Theorem 4.8 follows immediately from the dissipation inequality (4.27). Indeed, (4.27) implies that system (4.18) is GAS by the smooth state feedback $u = u(z,x)$ when $w = 0$. Moreover, in the presence of an L_2 disturbance signal $w(t)$, the disturbance attenuation problem, characterized in term of an $L_2 - L_{2p_0}$-gain, is solved by *smooth* state feedback. □

As a consequence of Theorem 4.8, we arrive at the following important conclusion.

COROLLARY 4.10. *Consider the nonlinear system* (4.18) *with trivial zero-dynamics, i.e.,* $\dim z = 0$. *Under the hypotheses in Assumptions* 4.1–4.2 *with* $z = 0$, *the disturbance attenuation problem with* $p = p_1$ *is solvable by smooth state feedback.*

REFERENCES

[1] A. BACCIOTTI, *Local stabilizability of nonlinear control systems*, World Scientific, River Edge, NJ, 1992.

[2] R. W. BROCKETT, *Asymptotic stability and feedback stabilization*, in Differential Geometric Control Theory, R. W. Brockett, R. S. Millman, and H. J. Sussmann, eds., Birkäuser, Basel, Boston, 1983, pp. 181–191.

[3] C. I. BYRNES AND A. ISIDORI, *New results and examples in nonlinear feedback stabilization*, Systems Control Lett., 12 (1989), pp. 437–442.

[4] C. I. BYRNES AND A. ISIDORI, *Asymptotic stabilization of minimum phase nonlinear systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 1122–1137.

[5] J. M. CORON AND L. PRALY, *Adding an integrator for the stabilization problem*, Systems Control Lett., 17 (1991), pp. 89–104.

[6] W. P. DAYAWANSA, *Recent advances in the stabilization problem for low dimensional systems*, in Proceedings of the 2nd IFAC Symposium on Nonlinear Control Systems Design, Bordeaux, France, 1992, pp. 1–8.

[7] W. P. DAYAWANSA, C. F. MARTIN, AND G. KNOWLES, *Asymptotic stabilization of a class of smooth two-dimensional systems*, SIAM J. Control Optim., 28 (1990), pp. 1321–1349.

[8] W. HAHN, *Stability of Motion*, Springer-Verlag, New York, 1967.

[9] H. HERMES, *Homogeneous coordinates and continuous asymptotically stabilizing feedback controls*, in Differential Equations Stability and Control, S. Elaydi, ed., Lecture Notes in Appl. Math. 109, Marcel Dekker, New York, 1991, pp. 249–260.

[10] H. HERMES, *Nilpotent and high-order approximations of vector field systems*, SIAM Rev., 33 (1991), pp. 238–264.

[11] A. ISIDORI, *Nonlinear Control Systems*, 3rd ed., Springer-Verlag, New York, 1995.

[12] A. ISIDORI, *Global almost disturbance decoupling with stability for non minimum-phase single-input single-output nonlinear systems*, Systems Control Lett., 28 (1996), pp. 115–122.

[13] A. Isidori and W. Lin, *Global L_2-gain design for a class of nonlinear systems*, Systems Control Lett., 34 (1998), pp. 295–302.
[14] M. Kawski, *Homogeneous stabilizing feedback laws*, Control Theory Adv. Tech., 6 (1990), pp. 497–516.
[15] M. Kawski, *Stabilization of nonlinear systems in the plane*, Systems Control Lett., 12 (1989), pp. 169–175.
[16] M. Krstić, I. Kanellakopoulos, and P. V. Kokotović, *Nonlinear and Adaptive Control Design*, John Wiley, New York, 1995.
[17] W. Lin, *Global robust stabilization of minimum-phase nonlinear systems with uncertainty*, Automatica J. IFAC, 33 (1997), pp. 453–462.
[18] W. Lin and T. Shen, *Robust passivity and feedback design for minimum-phase nonlinear systems with structural uncertainty*, Automatica J. IFAC, 35 (1999), pp. 35–47.
[19] W. Lin and C. Qian, *New results on global stabilization of feedforward nonlinear systems via small feedback*, in Proceedings of the 37th IEEE Conference on Decision and Control, Tampa, FL, 1998, pp. 879–884.
[20] W. Lin and C. Qian, *Adding one power integrator: A tool for global stabilization of high-order lower-triangular systems*, Systems Control Lett., 39 (2000), pp. 339–351.
[21] Y. Lin, E. D. Sontag, and Y. Wang, *Input to state stabilizability for parameterized family of systems*, Internat. J. Robust Nonlinear Control, 5 (1995), pp. 187–205.
[22] R. Marino, W. Respondek, and A. J. Van der Schaft, *Almost disturbance decoupling for single-input single output nonlinear systems*, IEEE Trans. Automat. Control, 34 (1989), pp. 1013–1017.
[23] R. Marino, W. Respondek, A. J. Van der Schaft, and P. Tomei, *Nonlinear H_∞ almost disturbance decoupling*, Systems Control Lett., 23 (1994), pp. 159–168.
[24] R. Marino and P. Tomei, *Nonlinear Control Design*, Prentice-Hall, London, UK, 1995.
[25] F. Mazenc, *Stabilization of feedforward systems approximated by a nonlinear chain of integrators*, Systems Control Lett, 32 (1997), pp. 223–229.
[26] H. Nijmeijer and A. J. van der Schaft, *Nonlinear Dynamical Control Systems*, Springer-Verlag, New York, 1990.
[27] C. Qian and W. Lin, *Using small feedback to stabilize a wider class of feedforward systems*, in Proceedings of the 14th IFAC World Congress, Vol. E, Beijing, China, 1999, pp. 309–314.
[28] C. Qian and W. Lin, *Almost disturbance decoupling for a chain of power integrators perturbed by a lower-triangular vector field*, IEEE Trans. Automat. Control, 45 (2000), pp. 1208–1214.
[29] H. Sussmann, *Limitations on the stabilizability of globally minimum phase systems*, IEEE Trans. Automat. Control, 35 (1990), pp. 117–119.
[30] J. Tsinias, *Sufficient Lyapunov-like conditions for stabilization*, Math. Control Signals Systems, 2 (1989), pp. 343–357.
[31] A. J. van der Schaft, *L_2-gain and Passivity Techniques in Nonlinear Control*, Lecture Notes in Control and Inform. Sci. 218, Springer-Verlag, New York, 1996.
[32] J. C. Willems, *Almost invariant subspace: An approach to high gain feedback design. Part* I: *Almost controlled invariant subspaces*, IEEE Trans. Automat. Control, 26 (1981), pp. 235–252.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 1, pp. 211–230

OBSERVABILITY AND CONTROL OF SCHRÖDINGER EQUATIONS*

K.-D. PHUNG†

Abstract. We propose an exact controllability result for Schrödinger equations in bounded domains under the Bardos–Lebeau–Rauch geometric control condition with an estimate of the control which is explicit with respect to the time of controllability. Also, we prove an explicit in time logarithmic observability estimate for the Schrödinger equation, where no geometrical conditions are supposed on the domain.

Key words. observability, controllability

AMS subject classifications. 93B07, 93B05, 93C05

PII. S0363012900368405

1. Introduction.

Let Ω be a bounded domain of $\mathbb{R}^n$, $n \geq 1$, with a smooth boundary $\partial\Omega$. We consider a nonempty open subset ω of Ω. The question we wish to address is that of controllability for Schrödinger equations with an explicit in time bound of the cost of the following control function. Given a time $\varepsilon > 0$ and considering initial data w_o in some appropriate space X, can we find a control $\vartheta \in L^1(0,\varepsilon;X)$ such that the solution of the system

$$\left\{\begin{array}{r} i\partial_t w + \Delta w = \vartheta_{|\omega} \quad \text{in } \Omega \times]0,\varepsilon[\,, \\ w = 0 \quad \text{on } \partial\Omega \times]0,\varepsilon[\,, \\ w(\cdot,0) = w_o \quad \text{in } \Omega \end{array}\right. \tag{1.1}$$

satisfies $w(\cdot,\varepsilon) \equiv 0$ in Ω, with an estimate of the control ϑ

$$\frac{1}{\sqrt{\varepsilon}} \left\| \vartheta_{|\omega} \right\|_{L^1(0,\varepsilon;X)} \leq \mathcal{C}(\varepsilon) \left\| w_o \right\|_X , \tag{1.2}$$

where $\mathcal{C}$ is an explicit function of ε?

From the work of Lions [Li] on the control for distributed systems, such a result can be obtained with the Hilbert uniqueness method (HUM) by solving the dual observability problem: in the case where $X = L^2(\Omega)$, under which hypothesis the solution of the homogenous Schrödinger equation

$$\left\{\begin{array}{r} i\partial_t u + \Delta u = 0 \quad \text{in } \Omega \times \mathbb{R}_t, \\ u = 0 \quad \text{on } \partial\Omega \times \mathbb{R}_t, \\ u(\cdot,0) = u_o \quad \text{in } \Omega \end{array}\right. \tag{1.3}$$

satisfies

$$\forall \varepsilon > 0, \quad \forall u_o \in L^2(\Omega), \quad \left\| u_o \right\|_{L^2(\Omega)} \leq \mathcal{C}(\varepsilon) \left\| u \right\|_{L^2(\omega\times]0,\varepsilon[)} . \tag{1.4}$$

The relation (1.4) concerns any initial data but may need suitable geometric conditions on ω. Also, we can establish such an observability estimate which is true for

*Received by the editors February 21, 2000; accepted for publication (in revised form) December 14, 2000; published electronically May 31, 2001.
http://www.siam.org/journals/sicon/40-1/36840.html
†17 rue Léonard Mafrand, 92320 Chatillon, France (phung@cmla.ens-cachan.fr).

any geometric situation but carries information only when $\|u_o\|_{H^s(\Omega)} = O(\|u_o\|_{L^2(\Omega)})$ for some $s \geq 1$. The problem becomes the following. For all $\omega \subset \Omega$, can we find a positive continuous and strictly increasing function $\mathcal{F} : \mathbb{R}_+^* \to \mathbb{R}_+^*$ which satisfies the relation $\lim_{x\to 0}\mathcal{F}(x) = 0$ such that one has the assertion

(1.5)

$$\forall \varepsilon > 0, \quad \forall u_o \in H^s(\Omega) \cap H_0^1(\Omega), \quad \|u_o\|^2_{L^2(\Omega)} \leq \mathcal{D}(\varepsilon)\ \mathcal{F}\left(\frac{\|u\|^2_{L^2(\omega\times]0,\varepsilon[)}}{\|u_o\|^2_{L^2(\Omega)}}\right)\|u_o\|^2_{H^s(\Omega)},$$

where $\mathcal{D}$ is an explicit positive function of ε?

These exact controllability and observability problems were already investigated in [M], [F], and in [LT], [M], [Le], [B], [T] if the control acts on a part of the boundary, but $\mathcal{C}(\varepsilon)$ was not calculated explicitly. The observability problem arises also in the context of parabolic [FI], [F-CZ], [LR1] or hyperbolic [Li], [BLR] systems. In [FI] and [LR1], an exact null controllability result for parabolic problems is established with no restriction on the time of control or on the support of the control function. Also, Fernandez-Cara and Zuazua [F-CZ] proved an explicit in time observability estimate for the heat equation. In [Ru], Russell used a transformation to show that a null controllability result for the heat equation for any time can be obtained from the exact controllability result for the wave equation in some time. Concerning hyperbolic systems, Bardos, Lebeau, and Rauch [BLR] show a link between the propagation of rays of geometric optics and the problem of exact controllability for hyperbolic problems. They give a geometrical control condition which is sufficient and almost necessary to obtain the observability for hyperbolic problems. Without this geometrical control condition, Robbiano [Ro] proved a logarithmic observability estimate for hyperbolic problems (but where ε must be large enough because of the finite speed of propagation) and showed how to use it to obtain an approximate control result with an estimation of the cost of the control.

In this paper, we give simple techniques and results which try to answer the three previous questions in different geometrical situations. Our strategy is to obtain results for the Schrödinger equation from well-known works on observation and controllability for parabolic and hyperbolic problems. Also, we will describe a method to have an exact control result for Schrödinger equations in bounded domain in $\mathbb{R}^n$, $n > 1$, from an observability result for the Schrödinger equation in one space dimension. Even if our approach does not give optimal results, we hope it can be used in other control problems. Let us now state the different results of this paper in the next section. The first result concerns the problem of observability for the Schrödinger equation when no geometrical conditions are required on ω. We give a logarithmic observability estimate. Then we study the case where the Bardos–Lebeau–Rauch geometric control hypothesis [BLR] holds. The second result is about the particular one-dimensional situation. The third result concerns the problem of exact control for the Schrödinger equation in a bounded domain of $\mathbb{R}^n$, $n > 1$.

2. The main results and some remarks. When no geometrical condition is assumed, we propose a logarithmic explicit in time observability estimate.

THEOREM 2.1. *Suppose $\Omega \subset \mathbb{R}^n$, $n \geq 1$, is of class C^∞. Let ω be a nonempty open subset of Ω. Then there exists $C > 0$ such that for all $\varepsilon > 0$, for all initial data $u_o \in H^2(\Omega) \cap H_0^1(\Omega)$, the solution of the homogenous Schrödinger equation* (1.3)

satisfies

$$\|u_o\|^2_{L^2(\Omega)} \leq \frac{C\,(1+1/\varepsilon)}{\ln\left(2+\frac{\|u_o\|^2_{L^2(\Omega)}}{\|u\|^2_{L^2(\omega\times]0,\varepsilon[)}}\right)}\ \|\Delta u_o\|^2_{L^2(\Omega)}\,. \tag{2.1}$$

The following second result is about the Schrödinger equation in one space dimension. We give an explicit in time observability estimate.

THEOREM 2.2. *Suppose $\Omega \subset \mathbb{R}^n$, $n = 1$. Let ω be a nonempty open subset of Ω. Then there exists $C > 0$ such that for all $\varepsilon > 0$, for all initial data $u_o \in L^2(\Omega)$, the solution of the homogenous Schrödinger equation* (1.3) *satisfies*

$$\|u_o\|^2_{L^2(\Omega)} \leq e^{C\left(1+1/\varepsilon^2\right)}\,\|u\|^2_{L^2(\omega\times]0,\varepsilon[)}\,. \tag{2.2}$$

The last result concerns the problem of exact control for the Schrödinger equation in a bounded domain of $\mathbb{R}^n$, $n > 1$. We estimate the size of the control.

THEOREM 2.3. *Suppose $\Omega \subset \mathbb{R}^n$, $n > 1$, is of class C^∞, and there is no infinite order of contact between the boundary $\partial\Omega$ and the bicharacteristics of $\partial_t^2 - \Delta$. If all generalized bicharacteristic rays meet $\omega \times]0, T_c[$ for some $0 < T_c < +\infty$, then for all $\varepsilon > 0$, for all initial conditions $w_o \in H_0^1(\Omega)$, there is a control $\vartheta = \vartheta_\varepsilon \in L^1\left(0,\varepsilon; L^2(\Omega)\right)$ such that the solution $w \in C\left([0,\varepsilon]; L^2(\Omega)\right)$ of the Schrödinger problem* (1.1) *satisfies $w(\cdot,\varepsilon) \equiv 0$ in Ω . Furthermore there exists a constant $C > 0$, such that for all $\varepsilon > 0$, we have*

$$\|\nabla w(\cdot,t)\|_{L^2(\Omega)} \leq \left(\frac{C}{\sqrt{t}} + e^{C\left(1+1/\varepsilon^2\right)}\right)\|\nabla w_o\|_{L^2(\Omega)} \qquad \forall t > 0 \tag{2.3}$$

with an estimate of the control ϑ_ε as follows:

$$\frac{1}{\sqrt{\varepsilon}}\,\|\vartheta_\varepsilon\|_{L^1(0,\varepsilon;L^2(\omega))} \leq \left(C+\sqrt{\varepsilon}\right)e^{C\left(1+1/\varepsilon^2\right)}\,\|\nabla w_o\|_{L^2(\Omega)}\,. \tag{2.4}$$

Let us make some comments.

1. Theorem 2.1 expresses a unique continuation property for the Schrödinger equation. Our approach to proving Theorem 2.1 consists of using an explicit in time observability estimate for parabolic equations obtained by Fernandez-Cara and Zuazua [F-CZ]. Next we introduce a Gaussian transformation to return to the solution of the Schrödinger equation. The logarithmic observability estimate (2.1) is equivalent to the following interpolation inequality:

$$\exists C > 0, \quad \forall \varepsilon, \delta > 0, \quad \|u_o\|^2_{L^2(\Omega)} \leq \exp\left(C\left(1+\frac{1}{\varepsilon}\right)\delta\right)\|u\|^2_{L^2(\omega\times]0,\varepsilon[)} + \frac{1}{\delta}\,\|\Delta u_o\|^2_{L^2(\Omega)}\,. \tag{2.5}$$

2. Theorem 2.2 asserts that we have an exact controllability result for the Schrödinger equation in one space dimension, due to the HUM of Lions [Li]. The proof of Theorem 2.2 combines multiplier techniques [M], [F] and interpolation inequalities. The interpolation estimates which are similar to (2.5) allow us to absorb the terms of lower order. The estimate (2.2) of Theorem 2.2 is also true for $n > 1$ if we suppose $\omega \subset \Omega \subset \mathbb{R}^n$ to be a neighborhood of $\overline{\Gamma_o}$, where $\Gamma_o = \{x \in \partial\Omega \,/\, (x - x_o)\cdot\nu(x) > 0\}$ is either equal to $\partial\Omega$ or is such that the boundary $\partial\Omega \setminus \Gamma_o$ is included in a hyperplan (see [F]), when x_o is a fixed point of $\mathbb{R}^n$, and $\nu(x)$ is the unit outward normal vector.

The author is indebted to Professor Zuazua, who called his attention to the papers [MZ], [I] and who pointed out that Theorem 2.2 can also be obtained from [MZ, Thm. 3.4], [I, Thm. 1], and a Fourier analysis.

3. Lebeau [Le] has proved the exact boundary controllability for the Schrödinger equation with an analytic boundary under the geometrical control condition of the work of Bardos, Lebeau, and Rauch on exact controllability for the wave equation [BLR]. Moreover, Burq [B] has proved the existence of open subsets of $\partial\Omega$ which do not geometrically control Ω for the wave equation in which it is possible to construct an exact boundary control for the Schrödinger equation if initial data are more regular than those with finite energy. Here the result of Theorem 2.3 is not optimal in norm in the sense that it should be enough to choose the initial condition in $w_o \in L^2(\Omega)$ to obtain a result of exact controllability for the Schrödinger equation with a control in $L^2(\omega \times]0,\varepsilon[)$ when it satisfies suitable geometric conditions (see the previous comment or the multiplier techniques [M], [F]). Also, Theorem 2.3 only implies the following observability estimate under the geometric control condition:

$$\forall \varepsilon > 0, \quad \forall u_o \in L^2(\Omega), \quad \|u_o\|_{H^{-1}(\Omega)} \leq \mathcal{C}(\varepsilon) \sup_{[0,\varepsilon]} \|u\|_{L^2(\omega)}. \tag{2.6}$$

4. Here our construction of the control given by Theorem 2.3 provides more precise information on the cost of the control. We propose a proof based on the theorem of Bardos, Lebeau, and Rauch [BLR] on exact controllability for hyperbolic equations, and a transformation inspired from the work of Boutet de Monvel [BdM] on the propagation of singularities in Schrödinger-type equations (see also [KS]). We will also use the estimate (2.2) of Theorem 2.2 in one space dimension to establish an explicit estimate on the size of the control function for the problem of exact controllability for the Schrödinger equation in $\Omega \subset \mathbb{R}^n$, $n > 1$, when the control region ω controls geometrically Ω. Furthermore, our method described (in section 5) below also applies to the case of Schrödinger equations with nonconstant principal part [HL]. We have the following control result.

THEOREM 2.4. *Let $\Delta_{\mathcal{A}} = \sum_{i,j=1}^n \frac{\partial}{\partial x_i} a_{ij}(x) \frac{\partial}{\partial x_j} + a_o(x)$, where the coefficients of $\Delta_{\mathcal{A}}$ are real, smooth, and satisfy the following conditions: $a_{ij}(x) = a_{ji}(x)$ and $\sum_{i,j=1}^n a_{ij}(x)\xi_i\xi_j \geq c|\xi|^2$. Suppose $\Omega \subset \mathbb{R}^n$, $n > 1$, of class C^∞, and there is no infinite order of contact between the boundary $\partial\Omega$ and the bicharacteristics of $\partial_t^2 - \Delta_{\mathcal{A}}$. If all generalized bicharacteristic rays of $\partial_t^2 - \Delta_{\mathcal{A}}$ meet $\omega \times]0,T_c[$ for some $0 < T_c < +\infty$, then for all $\varepsilon > 0$, for all initial conditions $w_o \in H_0^1(\Omega)$, there is a control $\vartheta_\varepsilon \in L^1\left(0,\varepsilon; L^2(\Omega)\right)$ such that the solution $w \in C\left([0,\varepsilon]; L^2(\Omega)\right)$ of the Schrödinger problem*

$$\begin{cases} i\partial_t w + \Delta_{\mathcal{A}} w = \vartheta_{\varepsilon\,|\omega} & \text{in } \Omega \times]0,\varepsilon[, \\ w = 0 \quad \text{on } \partial\Omega \times]0,\varepsilon[, \\ w(\cdot,0) = w_o \quad \text{in } \Omega \end{cases} \tag{2.7}$$

satisfies $w(\cdot,\varepsilon) \equiv 0$ in Ω. Furthermore, there exists a constant $C > 0$, such that for all $\varepsilon > 0$, the estimates (2.3)–(2.4) hold.

5. The main goal of this paper is to point out that an observability result for the heat equation gives a logarithmic observability estimate for the Schrödinger equation but also that an exact control result for the wave equation gives the exact controllability for the Schrödinger equation. Moreover, we show that an exact control result for the Schrödinger equation in one space dimension implies the exact controllability

for the Schrödinger equation in bounded domain $\Omega \subset \mathbb{R}^n$, $n > 1$, when a geometrical condition is assumed. The observability results for the Schrödinger equation are established under two different kinds of geometry: either when no geometrical hypothesis is assumed (Theorem 2.1) or when the Bardos–Lebeau–Rauch geometric control condition is satisfied (Theorem 2.2 ($n = 1$) and Theorem 2.3 ($n > 1$); see (2.6) in comment 3). Also, these techniques allow us to have explicit estimates with respect to the time of controllability.

The paper is organized in the following way. The proofs of Theorems 2.1 and 2.2 (see comment 2) rest on interpolation estimates which are established in section 3. These interpolation estimates can be seen as low frequency estimates. The proof of Theorem 2.1 is then easily described. In section 4, we prove Theorem 2.2. Section 5 is devoted to the construction of the control stated in Theorem 2.3.

3. Low frequency estimates. In this section, we first state interpolation inequalities in Theorem 3.1 below, which are the key results to prove Theorems 2.1 and 2.2 (see comment 2). Next, we recall some results on the observability for parabolic problems obtained by Fernandez-Cara and Zuazua [F-CZ]. Finally, we prove Theorem 3.1.

3.1. Interpolation inequalities and the proof of Theorem 2.1. We have the following interpolation inequalities.

THEOREM 3.1. *Suppose $\Omega \subset \mathbb{R}^n$, $n \geq 1$, is of class C^∞. Let ω be a nonempty open set included in Ω. Then there exist $C > 0$, $\varepsilon_o > 0$, $\mu_o > 0$, such that for all $\varepsilon \leq \varepsilon_o$, for all $\mu \geq \mu_o$, for all initial data $u_o \in H^2(\Omega) \cap H_0^1(\Omega)$, the solution of the homogenous Schrödinger equation* (1.3) *satisfies*

$$\int_\Omega |u_o(x)|^2 \, dx \leq \exp\left(\frac{C\mu}{\varepsilon}\right) \int_\omega \int_0^\varepsilon |u(x,t)|^2 \, dtdx + \frac{\varepsilon^3}{\mu} \int_\Omega |\Delta u_o(x)|^2 \, dx. \tag{3.1}$$

Furthermore, there exist $C > 0$, $\varepsilon_o > 0$, $\mu_o > 0$, such that for all $\varepsilon \leq \varepsilon_o$, for all $\mu \geq \mu_o$, for all initial data $u_o \in H^2(\Omega) \cap H_0^1(\Omega)$, the solution of the homogenous Schrödinger equation (1.3) *satisfies*

$$\int_\Omega |u_o(x)|^2 \, dx \leq \exp\left(\frac{C\mu}{\varepsilon}\right) \int_\omega \int_0^\varepsilon |\Delta u(x,t)|^2 \, dtdx + \frac{\varepsilon^3}{\mu} \int_\Omega |\Delta u_o(x)|^2 \, dx. \tag{3.2}$$

Let us assume that the interpolation inequality (3.1) holds in order to prove Theorem 2.1. The proof of Theorem 3.1 will be given at the end of section 3.

Proof of Theorem 2.1. By taking μ such that $\mu = C_0 \frac{\int_\Omega |\Delta u_0|^2 dx}{\int_\Omega |u_0|^2 dx} \geq \mu_o$, the estimate (3.1) becomes

$$\int_\Omega |u_0|^2 \, dx \leq \exp\left(\frac{CC_0}{\varepsilon} \frac{\int_\Omega |\Delta u_0|^2 \, dx}{\int_\Omega |u_0|^2 \, dx}\right) \int_\omega \int_0^\varepsilon |u|^2 \, dtdx + \frac{\varepsilon^3}{C_0} \int_\Omega |u_0|^2 \, dx. \tag{3.3}$$

So, $\exists C > 0$, $\exists \varepsilon_o > 0$, for all $\varepsilon \leq \varepsilon_o$,

$$\int_\Omega |u_0|^2 \, dx \leq \exp\left(\frac{C}{\varepsilon} \frac{\int_\Omega |\Delta u_0|^2 \, dx}{\int_\Omega |u_0|^2 \, dx}\right) \int_\omega \int_0^\varepsilon |u|^2 \, dtdx.$$

And also, $\exists C > 0$, for all $\varepsilon > 0$,

$$\int_\Omega |u_0|^2 \, dx \leq \exp\left(C\left(1 + \frac{1}{\varepsilon}\right) \frac{\int_\Omega |\Delta u_0|^2 \, dx}{\int_\Omega |u_0|^2 \, dx}\right) \int_\omega \int_0^\varepsilon |u|^2 \, dtdx. \tag{3.4}$$

The estimate (2.1) of Theorem 2.1 is equivalent to (3.4) by using properties of the logarithmic and exponential functions. That concludes the proof of Theorem 2.1. Let us remark here that we obtain (2.5) from (3.4) by studying the case where either $\|\Delta u_o\| \leq \delta \|u_o\|$ or $\|\Delta u_o\| > \delta \|u_o\|$ (see comment 1). □

It will be useful to recall some explicit in time observability results for parabolic problems before proving Theorem 3.1.

3.2. Observability for the parabolic problem. We recall the result in [F-CZ] in the particular case of a null potential.

Theorem [F-CZ]. Let Ω be a connected bounded domain in $\mathbb{R}^n$, with smooth boundary. Let v be the solution of the following adjoint parabolic equation:

$$\begin{cases} \partial_t v + \Delta v = 0 & \text{in } \Omega\times]0,T[, \\ v = 0 & \text{on } \partial\Omega\times]0,T[. \end{cases} \tag{3.5}$$

Then there is $C > 0$, such that for all $T > 0$,

$$\int_\Omega |v(x,0)|^2 \, dx \leq \exp\left(C\left(1+\frac{1}{T}\right)\right) \int_\omega \int_0^T |v|^2 \, dtdx. \tag{3.6}$$

Remark 3.2. Theorem [F-CZ] is obtained from the works of Fursikov and Imanuvilov [FI] on Carleman estimates for adjoint parabolic equations. Another approach, based on the work of Lebeau and Robbiano [LR1] on the exact controllability of the heat equation on a Riemannian compact manifold with boundary, and Dirichlet boundary conditions, in both cases of interior or boundary controls, gives us the estimates (3.6) but not explicitly in time. Nevertheless, a logarithmic boundary observability estimate for the Schrödinger equation is presented with that approach in [P].

We deduce from Theorem [F-CZ] the following corollary.

Corollary 3.3. *Let W be the solution of the following adjoint parabolic problem:*

$$\begin{cases} \partial_t W + \Delta W = f & \text{in } \Omega\times]0,T[, \\ W = 0 & \text{on } \partial\Omega\times]0,T[, \\ W(\cdot,T) \in L^2(\Omega). \end{cases} \tag{3.7}$$

Then

$$\exists C_T > 0, \qquad \int_\Omega |W(x,0)|^2 \, dx \leq C_T \left(\int_\omega \int_0^T |W|^2 \, dtdx + \int_\Omega \int_0^T |f|^2 \, dtdx \right). \tag{3.8}$$

If, moreover, $W(\cdot,T) \in H^2 \cap H_0^1(\Omega)$, then

$$\exists C_T > 0, \qquad \int_\Omega |W(x,0)|^2 \, dx \leq C_T \left(\int_\omega \int_0^T |\Delta W|^2 \, dtdx + \int_\Omega \int_0^T |f|^2 \, dtdx \right). \tag{3.9}$$

Here, the constant C_T of the estimates (3.8), (3.9) *is of the order of*

$$C_T = \exp\left(C\left(1+\frac{1}{T}\right)\right), \tag{3.10}$$

where $C > 0$ is a constant independent of $T > 0$.

Proof of Corollary 3.3. It is easy to see that (3.8) holds from (3.6) with a classical energy method. Let us prove (3.9). We consider $z(x,t) = W(x,t) - a(x,t)$, where

$$\left\{\begin{array}{rl} \partial_t a + \Delta a = f & \text{in } \Omega\times]0,T[, \\ a = 0 & \text{on } \partial\Omega\times]0,T[, \\ a(\cdot,T) = 0 & \text{in } \Omega. \end{array}\right. \tag{3.11}$$

As $\partial_t z$ is a solution of (3.5), the regularity of $W(\cdot,T)$ and (3.6) allow us to obtain the estimate

$$\int_\Omega |\Delta z(x,0)|^2\,dx \le \exp\left(C\left(1+\frac{1}{T}\right)\right)\int_\omega\int_0^T |\partial_t z|^2\,dtdx. \tag{3.12}$$

Now we give equalities on the solution a by a classical energy method:

$$\begin{aligned} &\frac{1}{2}\int_\Omega |a(x,0)|^2\,dx + \int_0^T\int_\Omega |\nabla a|^2\,dxdt = -\int_0^T\int_\Omega fa\,dxdt, \\ &\frac{1}{2}\int_\Omega |\nabla a(x,0)|^2\,dx + \int_0^T\int_\Omega |\partial_t a|^2\,dxdt = \int_0^T\int_\Omega f\partial_t a\,dxdt. \end{aligned} \tag{3.13}$$

By Cauchy–Schwarz and Poincaré inequalities and from (3.13) we have

$$\int_\Omega |a(x,0)|^2\,dx + \int_0^T\int_\Omega |\partial_t a|^2\,dxdt \le c\int_0^T\int_\Omega |f|^2\,dxdt. \tag{3.14}$$

We obtain from (3.12) and (3.14)

(3.15)

$$\begin{aligned} \int_\Omega |W(x,0)|^2\,dx &\le 2\int_\Omega |z(x,0)|^2\,dx + 2\int_\Omega |a(x,0)|^2\,dx \\ &\le c\int_\Omega |\Delta z(x,0)|^2\,dx + 2\int_\Omega |a(x,0)|^2\,dx \\ &\le \exp\left(C\left(1+\frac{1}{T}\right)\right)\int_\omega\int_0^T |\partial_t z|^2\,dtdx + c\int_0^T\int_\Omega |f|^2\,dxdt \\ &\le \exp\left(C\left(1+\frac{1}{T}\right)\right)\left(\int_\omega\int_0^T |\partial_t W|^2\,dtdx + c\int_0^T\int_\Omega |f|^2\,dxdt\right) \\ &\quad + c\int_0^T\int_\Omega |f|^2\,dxdt \\ &\le \exp\left(C\left(1+\frac{1}{T}\right)\right)\left(\int_\omega\int_0^T |\Delta W|^2\,dtdx + c\int_0^T\int_\Omega |f|^2\,dxdt\right) \\ &\quad + c\int_0^T\int_\Omega |f|^2\,dxdt. \end{aligned}$$

That concludes the proof of (3.9) and Corollary 3.3. □

3.3. Proof of Theorem 3.1. We begin to prove (3.1) as follows.

Let $F(z) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{iz\tau}e^{-\tau^2}\,d\tau$; then $F(z) = \frac{\sqrt{\pi}}{2\pi}e^{\frac{1}{4}\left(|\operatorname{Im} z|^2 - |\operatorname{Re} z|^2\right)}e^{-\frac{i}{2}(\operatorname{Im} z \operatorname{Re} z)}$. Also, with $\lambda > 0$, let us consider

$$F_\lambda(z) = \lambda F(\lambda z) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{iz\tau}e^{-\left(\frac{\tau}{\lambda}\right)^2}\,d\tau.$$

We have

$$|F_\lambda(z)| = \frac{\sqrt{\pi}}{2\pi}\lambda e^{\frac{\lambda^2}{4}\left(|\mathrm{Im}\, z|^2 - |\mathrm{Re}\, z|^2\right)}. \tag{3.16}$$

Let $s, \ell_0 \in \mathbb{R}$, and

$$W_{\ell_0,\lambda}(s,x) = \int_{\mathbb{R}} F_\lambda(\ell_0 + is - \ell)\Phi(\ell)u(x,\ell)d\ell, \tag{3.17}$$

where $\Phi \in C_0^\infty(\mathbb{R})$. The Gaussian transformation (3.17) is inspired from the Fourier–Bros–Iagolnitzer transformation in [LR2]. We remark that $\partial_s F_\lambda(\ell_0 + is - \ell) = -i\partial_\ell F_\lambda(\ell_0 + is - \ell)$, so

$$\begin{aligned}\partial_s W_{\ell_0,\lambda}(s,x) &= \int_{\mathbb{R}} -i\partial_\ell F_\lambda(\ell_0 + is - \ell)\Phi(\ell)u(x,\ell)d\ell \\ &= \int_{\mathbb{R}} iF_\lambda(\ell_0 + is - \ell)\left\{\frac{d}{d\ell}\Phi(\ell)u(x,\ell) + \Phi(\ell)\frac{\partial}{\partial\ell}u(x,\ell)\right\}d\ell.\end{aligned}$$

As $u : (x,t) \longmapsto u(x,t)$ is the solution of (1.3), $W_{\ell_0,\lambda}$ satisfies

$$\begin{cases} \partial_s W_{\ell_0,\lambda}(s,x) + \Delta W_{\ell_0,\lambda}(s,x) = \displaystyle\int_{\mathbb{R}} iF_\lambda(\ell_0 + is - \ell)\Phi'(\ell)u(x,\ell)d\ell, \\ W_{\ell_0,\lambda}(s,x) = 0 \quad \forall x \in \partial\Omega, \\ W_{\ell_0,\lambda}(0,x) = \left(F_\lambda * \Phi u(x,\cdot)\right)(\ell_0) \quad \forall x \in \Omega. \end{cases} \tag{3.18}$$

We define $\Phi \in C_0^\infty(\mathbb{R})$ such that the following holds. Let $L > 0$, and we choose $\Phi \in C_0^\infty(]0, L[)$, $0 \le \Phi \le 1$, $\Phi \equiv 1$ on $\left[\frac{L}{4}; \frac{3L}{4}\right]$ and such that $|\Phi'| \le \frac{8}{L}$. We take $K = \left[0; \frac{L}{4}\right] \cup \left[\frac{3L}{4}; L\right]$ and $K_0 = \left[\frac{3L}{8}; \frac{5L}{8}\right]$. So, $\mathrm{mes} K_0 = \frac{L}{4}$, $\mathrm{mes} K = \frac{L}{2}$, $\mathrm{supp}(\Phi') = K$, and $\mathrm{dist}(K; K_o) = \frac{L}{8}$. We will choose $\ell_0 \in K_0$.

As an application of (3.8), $W_{\ell_0,\lambda}$ satisfies the following estimate:

(3.19)

$$\begin{aligned}\int_\Omega \left|\left(F_\lambda * \Phi u(x,\cdot)\right)(\ell_0)\right|^2 dx &\le C_T \int_\omega \int_0^T |W_{\ell_0,\lambda}(s,x)|^2 \, dsdx \\ &\quad + C_T \int_\Omega \int_0^T \left|\int_{\mathbb{R}} iF_\lambda(\ell_0 + is - \ell)\Phi'(\ell)u(x,\ell)d\ell\right|^2 dsdx.\end{aligned}$$

On the other hand, from (3.16)

$$\begin{aligned}\int_\omega \int_0^T |W_{\ell_0,\lambda}(s,x)|^2 \, dsdx &= \int_\omega \int_0^T \left|\int_{\mathbb{R}} F_\lambda(\ell_0 + is - \ell)\Phi(\ell)u(x,\ell)d\ell\right|^2 dsdx \\ &\le \int_0^T \int_\omega \left|\int_{\mathbb{R}} \frac{\sqrt{\pi}}{2\pi}\lambda e^{\frac{\lambda^2}{4}\left(s^2 - |\ell_0 - \ell|^2\right)}\Phi(\ell)\left|u(x,\ell)\right|d\ell\right|^2 dxds \\ &\le \frac{\lambda^2}{4\pi}\left(\int_0^T e^{\frac{\lambda^2}{2}s^2}ds\right)|\sup\Phi|^2 \int_\omega \left|\int_0^L |u(x,\ell)|\, d\ell\right|^2 dx \\ &\le \frac{\lambda^2}{4\pi}e^{\frac{\lambda^2}{2}T^2}T \ \ |\sup\Phi|^2 \ \ L\int_\omega \int_0^L |u(x,\ell)|^2 \, d\ell dx\end{aligned}$$

and

$$\begin{aligned}
&\int_\Omega \int_0^T \left| \int_{\mathbb{R}} iF_\lambda(\ell_0 + is - \ell)\Phi'(\ell)u(x,\ell)d\ell \right|^2 dsdx \\
&\leq \int_0^T \int_\Omega \left| \int_{\mathbb{R}} \frac{\sqrt{\pi}}{2\pi} \lambda e^{\frac{\lambda^2}{4}\left(s^2 - |\ell_0 - \ell|^2\right)} |\Phi'(\ell)| \, |u(x,\ell)| \, d\ell \right|^2 dxds \\
&\leq \frac{\lambda^2}{4\pi} e^{\frac{\lambda^2}{2}T^2} T \int_\Omega \left(\int_K e^{-\frac{\lambda^2}{2}|\ell_0 - \ell|^2} |\Phi'(\ell)|^2 \, |u(x,\ell)|^2 \, d\ell \right) \mathrm{mes}\,(K)\, dx \\
&\leq \frac{\lambda^2}{4\pi} e^{\frac{\lambda^2}{2}T^2} T \; e^{-\frac{\lambda^2}{2}\mathrm{dist}(K,K_0)^2} \; \sup |\Phi'(\ell)|^2 \, \mathrm{mes}\,(K) \int_\Omega \int_K |u(x,\ell)|^2 \, d\ell dx \\
&\leq \frac{\lambda^2}{4\pi} e^{\frac{\lambda^2}{2}T^2} T \; e^{-\frac{\lambda^2}{2}\mathrm{dist}(K,K_0)^2} \; \sup |\Phi'(\ell)|^2 \, \mathrm{mes}\,(K)^2 \int_\Omega |u_o|^2 \, dx \\
&\leq \frac{\lambda^2 T}{4\pi} \exp\left[\frac{\lambda^2}{2} \left(T^2 - \left(\frac{L}{8} \right)^2 \right) \right] \; \frac{8^2}{L^2} \frac{L^2}{4} \int_\Omega |u_o|^2 \, dx \\
&\leq \frac{4\lambda^2 T}{\pi} \exp\left[\frac{\lambda^2}{2} \left(T^2 - \left(\frac{L}{8} \right)^2 \right) \right] \int_\Omega |u_o|^2 \, dx.
\end{aligned}$$

So, the inequality (3.19) becomes

$$\begin{aligned}
\int_\Omega \left| \left(F_\lambda * \Phi u(x,\cdot) \right)(\ell_0) \right|^2 dx &\leq C_T \frac{\lambda^2 TL}{4\pi} \exp\left(\frac{\lambda^2}{2} T^2 \right) \int_\omega \int_0^L |u(x,\ell)|^2 \, d\ell dx \\
&\quad + C_T \frac{4\lambda^2 T}{\pi} \exp\left[\frac{\lambda^2}{2} \left(T^2 - \left(\frac{L}{8} \right)^2 \right) \right] \int_\Omega |u_o|^2 \, dx. \qquad (3.20)
\end{aligned}$$

With the Parseval relation, we have

$$\begin{aligned}
&\int_{\mathbb{R}} \left| \Phi(\ell_0) u\,(x,\ell_0) - \left(F_\lambda * \Phi u\,(x,\cdot) \right)(\ell_0) \right|^2 d\ell_0 \\
&= \frac{1}{2\pi} \int_{\mathbb{R}} \left| \widehat{\Phi(\ell_0) u\,(x)}, \ell_0)(\tau) \right|^2 \left(1 - e^{-\left(\frac{\tau}{\lambda}\right)^2} \right)^2 d\tau \\
&\leq \frac{1}{\pi\lambda^2} \int_{\mathbb{R}} \left| \tau \widehat{\Phi(\ell_0) u\,(x)}, \ell_0)(\tau) \right|^2 d\tau \\
&\leq \frac{2}{\lambda^2} \int_{\mathbb{R}} \left| \Phi'(\ell_0) u\,(x,\ell_0) + \Phi(\ell_0) \partial_{\ell_0} u\,(x,\ell_0) \right|^2 d\ell_0 \\
&\leq \frac{4}{\lambda^2} \left[\frac{8^2}{L^2} \int_K |u(x,\ell_0)|^2 \, d\ell_0 + \int_0^L |\partial_{\ell_0} u(x,\ell_0)|^2 \, d\ell_0 \right].
\end{aligned}$$

By integrating on Ω, we obtain

$$\begin{aligned}
&\int_\Omega \int_{\mathbb{R}} \left| \Phi(\ell_0) u\,(x,\ell_0) - \left(F_\lambda * \Phi u\,(x,\cdot) \right)(\ell_0) \right|^2 d\ell_0 dx \qquad (3.21) \\
&\leq \frac{4}{\lambda^2} \left[\frac{c^2}{L^2} \int_K \int_\Omega |u(x,\ell_0)|^2 \, d\ell_0 dx + \int_0^L \int_\Omega |\partial_{\ell_0} u(x,\ell_0)|^2 \, d\ell_0 dx \right] \\
&\leq \frac{4}{\lambda^2} \left[\frac{8^2}{L^2} \frac{L}{2} \int_\Omega |u_o|^2 \, dx + L \int_\Omega |\Delta u_o|^2 \, dx \right].
\end{aligned}$$

So, with (3.20) and (3.21),

$$\begin{aligned}\operatorname{mes}(K_0)\int_\Omega |u_o|^2\,dx &= \int_{K_0}\int_\Omega |\Phi(\ell_0)u(x,\ell_0)|^2\,dx d\ell_0\\ &\le \operatorname{mes}(K_0)\,C_T\frac{\lambda^2 TL}{4\pi}\exp\left(\frac{\lambda^2}{2}T^2\right)\int_\omega\int_0^L |u(x,\ell)|^2\,d\ell dx\\ &+ \operatorname{mes}(K_0)\,C_T\frac{4\lambda^2 T}{\pi}\exp\left[\frac{\lambda^2}{2}\left(T^2-\left(\frac{L}{8}\right)^2\right)\right]\int_\Omega |u_o|^2\,dx\\ &+ \frac{4}{\lambda^2}\left[\frac{8^2}{L^2}\frac{L}{2}\int_\Omega |u_o|^2\,dx + L\int_\Omega |\Delta u_o|^2\,dx\right].\end{aligned}$$

Finally,

$$\begin{aligned}\int_\Omega |u_o|^2\,dx &\le C_T\frac{\lambda^2 TL}{4\pi}\exp\left(\frac{\lambda^2}{2}T^2\right)\int_\omega\int_0^L |u(x,\ell)|^2\,d\ell dx\\ &+ C_T\frac{4\lambda^2 T}{\pi}\exp\left[\frac{\lambda^2}{2}\left(T^2-\left(\frac{L}{8}\right)^2\right)\right]\int_\Omega |u_o|^2\,dx\\ &+ \frac{1}{\lambda^2}\frac{4^2}{L}\left[\frac{8^2}{2L}\int_\Omega |u_o|^2\,dx + L\int_\Omega |\Delta u_o|^2\,dx\right].\end{aligned}$$

Let us consider $A>0$ real such that $\left(1-A^2\right)<0$. By choosing $L=8AT$, it becomes

$$\begin{aligned}\int_\Omega |u_o|^2\,dx \le \frac{2A}{\pi}C_T\lambda^2T^2\exp\left(\frac{\lambda^2T^2}{2}\right)\int_\omega\int_0^{8AT}|u(x,\ell)|^2\,d\ell dx + 16\frac{1}{\lambda^2}\int_\Omega |\Delta u_o|^2\,dx\\ + \left[\frac{8}{A^2}\frac{1}{\lambda^2T^2} + \frac{4}{\pi}C_T\lambda^2T\exp\left(-\frac{A^2-1}{2}\lambda^2T^2\right)\right]\int_\Omega |u_o|^2\,dx.\end{aligned}$$

With the relation (3.10), we have the following uniform in time interpolation estimate:

$$\begin{aligned}\int_\Omega |u_o|^2\,dx &\le Ae^Ce^{C/T}\lambda^2T^2\exp\left(\frac{\lambda^2T^2}{2}\right)\int_\omega\int_0^{8AT}|u(x,\ell)|^2\,d\ell dx\\ &+ \left[\frac{C}{A^2}\frac{1}{\lambda^2T^2} + e^Ce^{C/T}\lambda^2T\exp\left(-\frac{A^2-1}{2}\lambda^2T^2\right)\right]\int_\Omega |u_o|^2\,dx\\ &+ \frac{16}{\lambda^2}\int_\Omega |\Delta u_o|^2\,dx.\end{aligned}$$

We introduce $\lambda^2=\frac{\mu}{T^3}$. Let $\alpha>0$ be real such that $\left(2\alpha+1-A^2\right)<0$; hence,

$$\begin{aligned}\int_\Omega |u_o|^2\,dx &\le Ae^Ce^{C/T}\frac{\mu}{T}\exp\left(\frac{\mu}{2T}\right)\int_\omega\int_0^{8AT}|u(x,\ell)|^2\,d\ell dx\\ &+ \left[\frac{C}{A^2}\frac{T}{\mu} + e^Ce^{C/T}\frac{\mu}{T^2}\exp\left(-\frac{A^2-1}{2}\frac{\mu}{T}\right)\right]\int_\Omega |u_o|^2\,dx\\ &+ 16\frac{T^3}{\mu}\int_\Omega |\Delta u_o|^2\,dx\\ &\le Ae^Ce^{C/T}\frac{\mu}{T}\exp\left(\frac{\mu}{2T}\right)\int_\omega\int_0^{8AT}|u(x,\ell)|^2\,d\ell dx\end{aligned}$$

$$+\left[\frac{C}{A^2}\frac{T}{\mu}+e^C e^{C/T}\frac{1}{\mu}\frac{4}{\alpha^2}\exp\left(-\frac{A^2-1-2\alpha}{2}\frac{\mu}{T}\right)\right]\int_\Omega |u_o|^2\,dx$$
$$+16\frac{T^3}{\mu}\int_\Omega |\Delta u_o|^2\,dx.$$

We take $\mu > \frac{2C}{A^2-1-2\alpha}$, so

$$\int_\Omega |u_o|^2\,dx \le Ae^C\frac{\mu}{T}\exp\left(\frac{A^2\mu}{2T}\right)\int_\omega\int_0^{8AT}|u(x,\ell)|^2\,d\ell dx$$
$$+\left[\frac{C}{A^2}\frac{T}{\mu}+e^C\frac{4}{\alpha^2}\frac{1}{\mu}\right]\int_\Omega |u_o|^2\,dx+16\frac{T^3}{\mu}\int_\Omega |\Delta u_o|^2\,dx,$$

$$\int_\Omega |u_o|^2\,dx \le \frac{e^C}{A}\exp\left(\frac{A^2\mu}{2T}\right)\int_\omega\int_0^{8AT}|u(x,\ell)|^2\,d\ell dx \tag{3.22}$$
$$+\left[\frac{CT}{A^2}+\frac{4e^C}{\alpha^2}\right]\frac{1}{\mu}\int_\Omega |u_o|^2\,dx+16\frac{T^3}{\mu}\int_\Omega |\Delta u_o|^2\,dx,$$

and with $\mu_o = \max(\frac{2C}{A^2-1-2\alpha}; 2(\frac{C}{A^2}+\frac{4e^C}{\alpha^2}))$

$$\forall T\le 1,\quad \forall \mu>\mu_o,\qquad \frac{1}{2}\int_\Omega |u_o|^2\,dx \tag{3.23}$$
$$\le \frac{e^C}{A}\exp\left(\frac{A^2\mu}{2T}\right)\int_\omega\int_0^{8AT}|u(x,\ell)|^2\,d\ell dx+16\frac{T^3}{\mu}\int_\Omega |\Delta u_o|^2\,dx.$$

Therefore, (3.1) is proved by choosing $T=\frac{\varepsilon}{8A}\le 1$.

The proof of (3.2) follows the same approach by using (3.9).

As application of (3.9), $W_{\ell_0,\lambda}$ satisfies the following estimate:

(3.24)
$$\int_\Omega |(F_\lambda * \Phi u(x,\cdot))(\ell_0)|^2\,dx \le C_T\int_\omega\int_0^T |\Delta W_{\ell_0,\lambda}(s,x)|^2\,dsdx$$
$$+C_T\int_\omega\int_0^T\left|\int_{\mathbb{R}} iF_\lambda(\ell_0+is-\ell)\Phi'(\ell)u(x,\ell)d\ell\right|^2 dsdx.$$

On the other hand, from (3.16)

$$\int_\omega\int_0^T |\Delta W_{\ell_0,\lambda}(s,x)|^2\,dsdx \le \frac{\lambda^2}{4\pi}e^{\frac{\lambda^2}{2}T^2}T\ |\sup\Phi|^2\ L\int_\omega\int_0^L |\Delta u(x,\ell)|^2\,d\ell dx.$$

Consequently, the inequality (3.24) becomes

$$\int_\Omega |(F_\lambda * \Phi u(x,\cdot))(\ell_0)|^2\,dx \le C_T\frac{\lambda^2 TL}{4\pi}\exp\left(\frac{\lambda^2}{2}T^2\right)\int_\omega\int_0^L |\Delta u(x,\ell)|^2\,d\ell dx$$
$$+C_T\frac{4\lambda^2 T}{\pi}\exp\left[\frac{\lambda^2}{2}\left(T^2-\left(\frac{L}{8}\right)^2\right)\right]\int_\Omega |u_o|^2\,dx. \tag{3.25}$$

Now, due to (3.25) and (3.21), for $\left(1-A^2\right)<0$ and $T\le 1$, we have

$$\int_\Omega |u_o|^2\,dx \le Ae^C e^{C/T}\lambda^2T^2\exp\left(\frac{\lambda^2T^2}{2}\right)\int_\omega\int_0^{8AT}|\Delta u(x,\ell)|^2\,d\ell dx+\frac{16}{\lambda^2}\int_\Omega |\Delta u_o|^2\,dx$$

$$+\left[\frac{C}{A^2}\frac{1}{\lambda^2T^2}+e^Ce^{C/T}\lambda^2T\exp\left(-\frac{A^2-1}{2}\lambda^2T^2\right)\right]\int_\Omega|u_o|^2\,dx.$$

We introduce $\lambda^2=\frac{\mu}{T^3}$ and $0<\alpha<\frac{1}{2}\left(A^2-1\right)$, so that

$$\begin{aligned}\int_\Omega|u_o|^2\,dx &\le Ae^{C/T}\frac{\mu}{T}\exp\left(\frac{\mu}{2T}\right)\int_\omega\int_0^{8AT}|\Delta u(x,\ell)|^2\,d\ell dx+16\frac{T^3}{\mu}\int_\Omega|\Delta u_o|^2\,dx\\ &+\left[\frac{C}{A^2}T+e^Ce^{C/T}\left(\frac{\mu}{T}\right)^2\exp\left(-\frac{A^2-1}{2}\frac{\mu}{T}\right)\right]\frac{1}{\mu}\int_\Omega|u_o|^2\,dx\\ &\le Ae^{C/T}\frac{\mu}{T}\exp\left(\frac{\mu}{2T}\right)\int_\omega\int_0^{8AT}|\Delta u(x,\ell)|^2\,d\ell dx+16\frac{T^3}{\mu}\int_\Omega|\Delta u_o|^2\,dx\\ &+\left[\frac{C}{A^2}T+e^Ce^{C/T}\frac{4}{\alpha^2}\exp\left(-\frac{A^2-1-2\alpha}{2}\frac{\mu}{T}\right)\right]\frac{1}{\mu}\int_\Omega|u_o|^2\,dx.\end{aligned}$$

We choose μ large enough such that

$$\int_\Omega|u_o|^2\,dx\le Ae^{C/T}\frac{\mu}{T}\exp\left(\frac{A\mu}{2T}\right)\int_\omega\int_0^{8AT}|\Delta u(x,\ell)|^2\,d\ell dx+C\frac{T^3}{\mu}\int_\Omega|\Delta u_o|^2\,dx.$$

Consequently, we have the following assertion: $\exists C>0$, $\exists\mu_o>0$, for all $T\le1$, for all $\mu>\mu_o$,

$$\int_\Omega|u_o|^2\,dx\le C\exp\left(\frac{A\mu}{T}\right)\int_\omega\int_0^{8AT}|\Delta u(x,\ell)|^2\,d\ell dx+C\frac{T^3}{\mu}\int_\Omega|\Delta u_o|^2\,dx. \tag{3.26}$$

That concludes the proof of Theorem 3.1. □

4. Proof of Theorem 2.2. We now prove Theorem 2.2 by using (3.2) when $n=1$ and multiplier techniques. Let A, B, β, ε be four reals such that $A<B$, $0<2\beta<B-A$, and $\varepsilon>0$. Let $\varphi:(t,s)\in\left]0,\varepsilon\right[\times\left]A,B\right[\longmapsto\varphi(t,s)$ be the solution of the Schrödinger equation in one space dimension:

$$\left\{\begin{array}{l} i\partial_t\varphi+\partial_s^2\varphi=0\quad\text{in }\left]0,\varepsilon\right[\times\left]A,B\right[,\\ \varphi(\cdot,A)=\varphi(\cdot,B)=0\quad\text{on }\left]0,\varepsilon\right[,\\ \varphi(0,\cdot)=\varphi_o\quad\text{in }\left]A,B\right[.\end{array}\right. \tag{4.1}$$

We will prove the following stable observability estimate: $\exists C>0,$ for all $\varepsilon>0$,

$$\int_A^B\left|\partial_s^2\varphi_o\right|^2ds\le e^{C\left(1+1/\varepsilon^2\right)}\int_{B-2\beta}^{B-\beta}\int_0^\varepsilon\left|\partial_s^2\varphi\right|^2dtds. \tag{4.2}$$

Indeed, let $q\in C^2\left([A,B]\right)$ be a real function

$$\begin{aligned}iq\frac{d}{dt}\left(\partial_s\varphi\partial_s^2\overline{\varphi}\right) &=q\left(i\frac{d}{dt}\partial_s\varphi\partial_s^2\overline{\varphi}+i\partial_s\varphi\frac{d}{dt}\partial_s^2\overline{\varphi}\right)\\ &=q\left(-\partial_s^3\varphi\partial_s^2\overline{\varphi}+\partial_s\varphi\partial_s^4\overline{\varphi}\right)\\ &=q\left(-\partial_s^3\varphi\partial_s^2\overline{\varphi}+\partial_s\left(\partial_s\varphi\partial_s^3\overline{\varphi}\right)-\partial_s^2\varphi\partial_s^3\overline{\varphi}\right)\\ &=-q\partial_s\left(\partial_s^2\varphi\partial_s^2\overline{\varphi}\right)+q\partial_s\left(\partial_s\varphi\partial_s^3\overline{\varphi}\right).\end{aligned}$$

So

$$\begin{aligned}\int_0^\varepsilon\int_A^B iq\frac{d}{dt}\left(\partial_s\varphi\partial_s^2\overline{\varphi}\right)dsdt &= -\int_0^\varepsilon\int_A^B q\partial_s\left|\partial_s^2\varphi\right|^2+\int_0^\varepsilon\int_A^B q\partial_s\left(\partial_s\varphi\partial_s^3\overline{\varphi}\right)\\ &= -\int_0^\varepsilon\left[q\left|\partial_s^2\varphi\right|^2\right]_A^B+\int_0^\varepsilon\int_A^B q'\left|\partial_s^2\varphi\right|^2\\ &+\int_0^\varepsilon\left[q\partial_s\varphi\partial_s^3\overline{\varphi}\right]_A^B-\int_0^\varepsilon\int_A^B q'\left(\partial_s\varphi\partial_s^3\overline{\varphi}\right)\\ &= \int_0^\varepsilon\int_A^B q'\left|\partial_s^2\varphi\right|^2+\int_0^\varepsilon\left[q\partial_s\varphi\partial_s^3\overline{\varphi}\right]_A^B\\ &-\int_0^\varepsilon\left[q'\partial_s\varphi\partial_s^2\overline{\varphi}\right]_A^B+\int_0^\varepsilon\int_A^B\left(q''\partial_s\varphi+q'\partial_s^2\varphi\right)\partial_s^2\overline{\varphi}\\ &= \int_0^\varepsilon\left[q\partial_s\varphi\partial_s^3\overline{\varphi}\right]_A^B+2\int_0^\varepsilon\int_A^B q'\left|\partial_s^2\varphi\right|^2+\int_0^\varepsilon\int_A^B q''\partial_s\varphi\partial_s^2\overline{\varphi}.\end{aligned}$$

Finally,

$$\begin{aligned}\int_A^B\left[iq\partial_s\varphi\partial_s^2\overline{\varphi}\right]_0^\varepsilon ds &= \int_0^\varepsilon\left[q\partial_s\varphi\partial_s^3\overline{\varphi}\right]_A^B dt\\ &+2\int_0^\varepsilon\int_A^B q'\left|\partial_s^2\varphi\right|^2 dsdt+\int_0^\varepsilon\int_A^B q''\partial_s\varphi\partial_s^2\overline{\varphi}dsdt.\end{aligned}\tag{4.3}$$

Consequently, by taking the real part of (4.3), we obtain

$$\begin{aligned}&2\int_0^\varepsilon\int_A^B q'\left|\partial_s^2\varphi\right|^2 dsdt+\mathrm{Re}\int_0^\varepsilon\left[q\partial_s\varphi\partial_s^3\overline{\varphi}\right]_A^B dt\\ &=-\mathrm{Im}\int_A^B\left[q\partial_s\varphi\partial_s^2\overline{\varphi}\right]_0^\varepsilon ds-\mathrm{Re}\int_0^\varepsilon\int_A^B q''\partial_s\varphi\partial_s^2\overline{\varphi}dsdt.\end{aligned}$$

By choosing $q(s)=s-A$, we have

$$\begin{aligned}2\int_0^\varepsilon\int_A^B\left|\partial_s^2\varphi\right|^2 dsdt &= -(B-A)\,\mathrm{Re}\int_0^\varepsilon\partial_s\varphi(t,B)\,\partial_s^3\overline{\varphi}(t,B)\,dt\\ &-\mathrm{Im}\int_A^B\left[(s-A)\,\partial_s\varphi\partial_s^2\overline{\varphi}\right]_0^\varepsilon ds.\end{aligned}\tag{4.4}$$

By choosing $q(s)=\chi(s)$ with $\mathrm{supp}\chi\subset[B-2\beta,B]$ and $\chi(B)\neq 0$, we have

$$\begin{aligned}-\chi(B)\,\mathrm{Re}\int_0^\varepsilon\partial_s\varphi(t,B)\,\partial_s^3\overline{\varphi}(t,B)\,dt &= 2\int_0^\varepsilon\int_A^B\chi'(s)\left|\partial_s^2\varphi\right|^2 dsdt\\ &+\mathrm{Im}\int_A^B\left[\chi(s)\,\partial_s\varphi\partial_s^2\overline{\varphi}\right]_0^\varepsilon ds+\mathrm{Re}\int_0^\varepsilon\int_A^B\chi''(s)\,\partial_s\varphi\partial_s^2\overline{\varphi}dsdt.\end{aligned}\tag{4.5}$$

Due to (4.4) and (4.5), if, moreover, $\mathrm{supp}\chi'\subset[B-2\beta,B-\beta]$, we obtain the following assertion:

$\exists c > 0, \quad \forall \varepsilon > 0,$

$$\int_0^\varepsilon \int_A^B \left|\partial_s^2 \varphi\right|^2 ds dt \leq c \left(\int_0^\varepsilon \int_{B-2\beta}^{B-\beta} \left|\partial_s^2 \varphi\right|^2 ds dt + (1+\varepsilon) \left\|\partial_s \varphi_o\right\| \left\|\partial_s^2 \varphi_o\right\|_{L^2(]A,B[)} \right).$$

Hence

$$\int_A^B \left|\partial_s^2 \varphi_o\right|^2 ds \leq \frac{c}{\varepsilon} \int_{B-2\beta}^{B-\beta} \int_0^\varepsilon \left|\partial_s^2 \varphi\right|^2 ds dt + c \left(1 + \frac{1}{\varepsilon}\right) \left\|\partial_s \varphi_o\right\|_{L^2(]A,B[)} \left\|\partial_s^2 \varphi_o\right\|_{L^2(]A,B[)}.$$

Finally,

$$\exists c > 0, \quad \forall \varepsilon \leq 1, \qquad \int_A^B \left|\partial_s^2 \varphi_o\right|^2 ds \leq \frac{c}{\varepsilon} \int_{B-2\beta}^{B-\beta} \int_0^\varepsilon \left|\partial_s^2 \varphi\right|^2 dt ds + \frac{c}{\varepsilon^2} \left\|\partial_s \varphi_o\right\|^2_{L^2(]A,B[)}.$$

By interpolation, we have

(4.6)

$$\exists c > 0, \quad \forall \varepsilon \leq 1, \qquad \int_A^B \left|\partial_s^2 \varphi_o\right|^2 ds \leq \frac{c}{\varepsilon} \int_{B-2\beta}^{B-\beta} \int_0^\varepsilon \left|\partial_s^2 \varphi\right|^2 dt ds + \frac{c}{\varepsilon^4} \left\|\varphi_o\right\|^2_{L^2(]A,B[)}.$$

But the interpolation inequality (3.2) of Theorem 3.1 in the one-dimensional case implies that

$\exists C > 0, \exists \varepsilon_o > 0, \exists \mu_o > 0, \forall \varepsilon \leq \varepsilon_o, \forall \mu \geq \mu_o,$

$$\int_A^B \left|\varphi_o\right|^2 ds \leq \exp\left(C \frac{\mu}{\varepsilon}\right) \int_{B-2\beta}^{B-\beta} \int_0^\varepsilon \left|\partial_s^2 \varphi\right|^2 dt ds + \frac{\varepsilon^3}{\mu} \int_A^B \left|\partial_s^2 \varphi_o\right|^2 ds. \tag{4.7}$$

Let D_o be real such that $D_o = \max(2; \frac{\min(1;\varepsilon_o)\mu_o}{c})$. By choosing $\mu = D_o \frac{c}{\varepsilon} \geq \mu_o$, we conclude the proof of (4.2) from (4.7) and (4.6). And, in a standard way [Li], [F], we also have $\exists C > 0$, for all $\varepsilon > 0$,

$$\int_A^B \left|\varphi_o\right|^2 ds \leq e^{C\left(1+1/\varepsilon^2\right)} \int_{B-2\beta}^{B-\beta} \int_0^\varepsilon \left|\varphi\right|^2 dt ds. \tag{4.8}$$

That concludes the proof of Theorem 2.2. □

Remark 4.1. To prove (2.2), we used the multiplier methods and we absorbed the terms of lower order with the interpolation inequality (3.2) for $n = 1$. The same method can be used for $n > 1$ (see comment 2). Indeed, from the equality (1.21) of the work of Fabre [F, Lemma 1.9] on the exact internal controllability of the Schrödinger equation, we choose $\theta = g^2(t)$, where $g \in C_0^\infty(]0, \varepsilon[)$, $g = 1$ in $]\varepsilon/3; 2\varepsilon/3[$, and $0 \leq g \leq 1$, to obtain with standard bootstrap arguments the following assertion:

$$\exists c > 0, \quad \forall \varepsilon \leq 1, \qquad \int_\Omega \left|\Delta u_o\right|^2 dx dt \leq \frac{c}{\varepsilon} \int_0^\varepsilon \int_\omega \left|\Delta u\right|^2 dx dt + \frac{c}{\varepsilon^2} \left\|u_o\right\|^2_{H^1(\Omega)}. \tag{4.9}$$

Under the hypothesis of Lemma 1.9 of the work of Fabre [F, p. 350],we have (2.2) for $n > 1$ by applying the interpolation inequality (3.2) of Theorem 3.1.

5. Proof of Theorem 2.3. This section is devoted to proving the exact control result for Schrödinger equations with an explicit in time estimate of the control. We proceed in three steps.

5.1. Step 1. The Schrödinger equation on $\mathbb{R}$. In this section, we prove the existence of the following solution of the Schrödinger equation on $\mathbb{R}$.

PROPOSITION 5.1. *Let $T > 0$ be real, and let δ be the Dirac measure. There exists a distribution $f = f(t,s)$ defined on $]0,\varepsilon[\times \mathbb{R}_s$ such that $f_t : s \in \mathbb{R}_s \longmapsto f(t,s)$ has a support included in $J = (]-\infty,-2T[\cup]2T,+\infty[)$ and the solution $F : (t,s) \in [0,\varepsilon] \times \mathbb{R}_s \longmapsto F(t,s)$ of the Schrödinger equation*

$$\left\{ \begin{array}{l} i\partial_t F + \partial_s^2 F = f_{|J} \quad \text{in }]0,\varepsilon[\times \mathbb{R}_s, \\ F(0,\cdot) = \delta(\cdot) \quad \text{in } \mathbb{R}_s \end{array} \right. \tag{5.1}$$

satisfies $F(\varepsilon,\cdot) \equiv 0$ in $[-T,T]$. Moreover, if $H = F - E$, where E is the fundamental solution of the Schrödinger equation in one space dimension, then

$$\exists C > 0, \quad \forall \varepsilon > 0, \qquad \|H\|_{L^\infty(0,\varepsilon;L^2(]-T,T[))} \leq e^{C\left(1+1/\varepsilon^2\right)}. \tag{5.2}$$

Remark 5.2. The result of Proposition 5.1 simply says that the Schrödinger equation on the whole line can be controlled to zero with a control concentrated in the exterior of the ball. This is also true in several dimensions. The proof of this can be easily obtained from the result on a bounded domain by a cut-off argument (see also [Z]). In our case, we obtain an explicit estimate with respect to the time ε of controllability. Here T does not denote time, and let us adopt the variables $(t,s) \in [0,\varepsilon] \times \mathbb{R}_s$ when we consider the one-dimensional case.

Proof of Proposition 5.1. Using the HUM of Lions [Li] and estimate (2.2) of Theorem 2.2, we know that for all data $v_\varepsilon \in L^2(]-3T,4T[$, there exists a control $h \in L^2(]0,\varepsilon[\times]3T,4T[)$ such that the solution $v : (t,s) \longmapsto v(t,s) \in C([0,\varepsilon];L^2(]-3T,4T[))$ satisfies

$$\left\{ \begin{array}{l} i\partial_t v + \partial_s^2 v = h_{|]3T,4T[} \quad \text{in }]0,\varepsilon[\times]-3T,4T[, \\ v(\cdot,-3T) = 0,\ v(\cdot,4T) = 0 \quad \text{on }]0,\varepsilon[, \\ v(0,\cdot) = 0 \quad \text{in }]-3T,4T[, \\ v(\varepsilon,\cdot) = v_\varepsilon \quad \text{in }]-3T,4T[\end{array} \right. \tag{5.3}$$

and

$$\frac{1}{\sqrt{\varepsilon}} \|h\|_{L^1(0,\varepsilon;L^2(]3T,4T[))} \leq \|h\|_{L^2(]0,\varepsilon[\times]3T,4T[)} \leq e^{C\left(1+1/\varepsilon^2\right)} \|v_\varepsilon\|_{L^2(]-3T,4T[)}. \tag{5.4}$$

In particular, we take $v_\varepsilon(\varepsilon,s) = -\chi(s) \frac{e^{-i\frac{\pi}{4}}}{\sqrt{4\pi\varepsilon}} e^{i\frac{s^2}{4\varepsilon}}$, where $s \in]-3T,4T[$, $\chi \in C_0^\infty(]-3T,4T[)$, $0 \leq \chi \leq 1$, $\chi_{|[-T,T]} = 1$. So

$$\|v\|_{L^\infty(0,\varepsilon;L^2(]-3T,4T[))} \leq 2\|h\|_{L^1(0,\varepsilon;L^2(]3T,4T[))} \leq Ce^{C\left(1+1/\varepsilon^2\right)}. \tag{5.5}$$

Let us consider

$$H(t,s) = \left| \begin{array}{ll} v(t,s) & \text{in } [0,\varepsilon] \times [-3T,4T], \\ 0 & \text{in } [0,\varepsilon] \times (]-\infty,-3T[\cup]4T,+\infty[), \end{array} \right. \tag{5.6}$$

where v is the solution of (5.3). So

$$\left\{ \begin{array}{l} i\partial_t H + \partial_s^2 H = h_{|]3T,4T[} - \partial_s v \otimes \delta(s-4T) + \partial_s v \otimes \delta(s+3T) \quad \text{in }]0,\varepsilon[\times \mathbb{R}_s, \\ H(0,\cdot) = 0 \quad \text{in } \mathbb{R}_s, \\ H(\varepsilon,s) = -\chi(s) \frac{e^{-i\frac{\pi}{4}}}{\sqrt{4\pi\varepsilon}} e^{i\frac{s^2}{4\varepsilon}} \end{array} \right. \tag{5.7}$$

and

$$\exists C>0,\quad \forall \varepsilon>0,\qquad \|H\|_{L^\infty(0,\varepsilon;L^2(]-3T,4T[))}\leq e^{C\left(1+1/\varepsilon^2\right)}. \tag{5.8}$$

Let E be the fundamental solution of the Schrödinger equation on the whole line

$$E(t,s)=\frac{e^{-i\frac{\pi}{4}}}{\sqrt{4\pi t}}e^{i\frac{s^2}{4t}}. \tag{5.9}$$

The solution $E\in C^\infty(\{t>0\}\times\mathbb{R}_s)\cap C\left([0,+\infty[\,;H^{-1/2-\epsilon}(\mathbb{R}_s)\right)$ satisfies

$$\begin{cases} i\partial_t E+\partial_s^2 E=0 \quad \text{in } \{t>0\}\times\mathbb{R}_s,\\ E(0,\cdot)=\delta(\cdot)\in H^{-1/2-\epsilon}(\mathbb{R}_s). \end{cases} \tag{5.10}$$

We finally consider $f=h_{|]3T,4T[}-\partial_s v\otimes\delta(s-4T)+\partial_s v\otimes\delta(s+3T)$ such that the solution $F=E+H$ satisfies

$$\begin{cases} i\partial_t F+\partial_s^2 F=f_{|J} \quad \text{in }]0,\varepsilon[\times\mathbb{R}_s,\\ i\partial_t F+\partial_s^2 F=0 \quad \text{in }]0,\varepsilon[\times[-T,T],\\ F(0,\cdot)=\delta(\cdot) \quad \text{in } \mathbb{R}_s,\\ F(\varepsilon,\cdot)\equiv 0 \quad \text{in } [-T,T]. \end{cases} \tag{5.11}$$

That concludes the proof of Proposition 5.1. □

5.2. Step 2. Controllability for the hyperbolic problem. The following exact controllability result holds.

LEMMA 5.3. *Suppose $\Omega\subset\mathbb{R}^n$, $n\geq 1$, is of class C^∞, and there is no infinite order of contact between the boundary $\partial\Omega$ and the bicharacteristics of $\partial_t^2-\Delta$. If all generalized bicharacteristic rays meet $\omega\times]0,T_c[$ for some $0<T_c<+\infty$, then for all $T>T_c$, for all initial condition $w_o\in H_0^1(\Omega)$, there exists a control $g\in L^2(\Omega\times]-T,T[)$ such that the solution $y\in C\left(\mathbb{R}_t;H_0^1(\Omega)\right)\cap C^1\left(\mathbb{R}_t;L^2(\Omega)\right)$ satisfies*

$$\begin{cases} \partial_t^2 y-\Delta y=g_{|\omega\times]-T,T[} \quad \text{in } \Omega\times\mathbb{R}_t,\\ y=0 \quad \text{on } \partial\Omega\times\mathbb{R}_t,\\ y(\cdot,0)=w_o,\ \partial_t y(\cdot,0)=0 \quad \text{in } \Omega,\\ y\equiv 0 \quad \text{in } \Omega\times(]-\infty,-T]\cup[T,+\infty[) \end{cases} \tag{5.12}$$

and

$$\|g\|_{L^2(\omega\times]-T,T[)}\leq C_{\omega,T}\|\nabla w_o\|_{L^2(\Omega)}. \tag{5.13}$$

Remark 5.4. The result of Lemma 5.3 holds by a simple reflection argument as a consequence of the theorem of Bardos, Lebeau, and Rauch [BLR] on the exact controllability for hyperbolic equations which are obtained with microlocal techniques and propagation of singularities of the solution of hyperbolic systems. We recall their result to be complete.

Theorem [BLR]. Suppose $\Omega\subset\mathbb{R}^n$, $n\geq 1$, is of class C^∞, and there is no infinite order of contact between the boundary $\partial\Omega$ and the bicharacteristics of $\partial_t^2-\Delta$. If all generalized bicharacteristic rays meet $\omega\times]0,T_c[$ for some $0<T_c<+\infty$, then for all $T>T_c$, for all $\theta\in C_0^\infty(]0,T[)$, for all initial conditions $(w_o,w_1)\in H_0^1(\Omega)\times L^2(\Omega)$, there exists a control $\varrho\in L^2(\Omega\times\mathbb{R}_t)$ such that the solution $\Psi\in C\left(\mathbb{R}_t;H_0^1(\Omega)\right)\cap C^1\left(\mathbb{R}_t;L^2(\Omega)\right)$ satisfies

$$\begin{cases} \partial_t^2\Psi-\Delta\Psi=\theta\varrho_{|\omega} \quad \text{in } \Omega\times\mathbb{R}_t,\\ \Psi=0 \quad \text{on } \partial\Omega\times\mathbb{R}_t,\\ \Psi(\cdot,0)=w_o,\ \partial_t\Psi(\cdot,0)=w_1 \quad \text{in } \Omega,\\ \Psi(\cdot,T)=\partial_t\Psi(\cdot,T)=0 \quad \text{in } \Omega \end{cases}$$

and

$$\|\theta\varrho\|_{L^2(\omega\times]0,T[)} \leq C_{\omega,T}\,\|(\nabla w_o, w_1)\|_{L^2(\Omega)}\,.$$

Proof of Lemma 5.3. We choose $w_1 = 0$ and extend Ψ in a symmetric way by taking

$$y\,(x,t) = \left|\begin{array}{ll} \Psi\,(x,t) & \text{in } \Omega\times[0,T]\,,\\ \Psi\,(x,-t) & \text{in } \Omega\times[-T,0[\,. \end{array}\right.$$

The control g will be given by

(5.14)
$$g\,(x,t)_{|\omega\times]-T,T[} = \theta\,(t)\,\varrho\,(x,t)_{|\omega\times]0,T[} + \theta\,(-t)\,\varrho\,(x,-t)_{|\omega\times]-T,0[} \quad \text{in } \Omega\times\,]-T,T[\,,$$

where $\theta \in C_0^\infty\,(]0,T[)$ so that $g\,(\cdot,-T) = g\,(\cdot,0) = g\,(\cdot,T) = 0$. □

5.3. Step 3. Construction of the control. Now we are able to construct and estimate the control of Theorem 2.3 as follows.

We define $w : (x,t) \in \Omega\times[0,\varepsilon] \longmapsto w\,(x,t)$ such that

$$w\,(x,t) = \int_{\mathbb{R}} F\,(t,\ell)\,y\,(x,\ell)\,d\ell, \tag{5.15}$$

where $F : (t,\ell) \in [0,\varepsilon]\times\mathbb{R}_\ell \longmapsto F\,(t,\ell)$ is obtained from Proposition 5.1:

$$\left\{\begin{array}{l} i\partial_t F + \partial_\ell^2 F = f_{|J} \quad \text{in } \,]0,\varepsilon[\,\times\mathbb{R}_\ell,\\ i\partial_t F + \partial_\ell^2 F = 0 \quad \text{in } \,]0,\varepsilon[\,\times[-T,T]\,,\\ \qquad F\,(0,\cdot) = \delta\,(\cdot) \quad \text{in } \mathbb{R}_\ell,\\ \qquad F\,(\varepsilon,\cdot) \equiv 0 \quad \text{in } \,[-T,T]\,, \end{array}\right. \tag{5.16}$$

and $y : (x,\ell) \in \Omega\times\mathbb{R}_\ell \longmapsto y\,(x,\ell)$ given by Lemma 5.3 satisfies

$$\left\{\begin{array}{l} \partial_\ell^2 y - \Delta y = g_{|\omega\times]-T,T[} \quad \text{in } \Omega\times\mathbb{R}_\ell,\\ \qquad y = 0 \quad \text{on } \partial\Omega\times\mathbb{R}_\ell,\\ \quad y(\cdot,0) = w_o \in H_0^1\,(\Omega)\,,\ \partial_\ell y(\cdot,0) = 0 \quad \text{in } \Omega,\\ \qquad y \equiv 0 \quad \text{in } \Omega\times(]-\infty,-T]\cup[T,+\infty[)\,. \end{array}\right. \tag{5.17}$$

Let us calculate $i\partial_t w\,(x,t)$:

$$\begin{aligned} i\partial_t w\,(x,t) &= \int_{\mathbb{R}} i\partial_t F\,(t,\ell)\,y\,(x,\ell)\,d\ell\\ &= \int_{\mathbb{R}} \left[-\partial_\ell^2 F\,(t,\ell) + f_{|J}\right] y\,(x,\ell)\,d\ell\\ &= \int_{\mathbb{R}} -F\,(t,\ell)\,\partial_\ell^2 y\,(x,\ell)\,d\ell + \int_{-T}^{T} \left[f_{|J}\right] y\,(x,\ell)\,d\ell\\ &= \int_{\mathbb{R}} F\,(t,\ell)\left[-\Delta y\,(x,\ell) - g\,(x,\ell)_{|\omega\times]-T,T[}\right] d\ell. \end{aligned}$$

Remark 5.5. The key point is that the solution $y : (x,\ell) \longmapsto y\,(x,\ell)$, where $\ell \in \mathbb{R}$, is identically null for ℓ out of the domain $]-T,T[$. Next, we need that the solution $F : (t,\ell) \longmapsto F\,(t,\ell)$ is defined for $\ell \in \mathbb{R}$. Moreover, F must satisfy $F\,(0,\cdot) = \delta\,(\cdot)$

in $\mathbb{R}$, $F(\varepsilon,\cdot)\equiv 0$ in $[-T,T]$, and also the homogenous Schrödinger equation in the domain $]0,\varepsilon[\times[-T,T]$. Out of the domain $[-T,T]$, F is solution of the Schrödinger equation with a second member f, but we do not see it in the integrations by parts because y is null on the support of f.

Our conclusion is

$$\left\{\begin{array}{r} i\partial_t w+\Delta w=\vartheta_{\varepsilon|\omega} \quad \text{in } \Omega\times]0,\varepsilon[\,,\\ w=0 \quad \text{on } \partial\Omega\times]0,\varepsilon[\,,\\ w(\cdot,0)=w_o \quad \text{in } \Omega,\\ w(\cdot,\varepsilon)=0 \quad \text{in } \Omega \end{array}\right. \tag{5.18}$$

with an estimate of the control ϑ_ε in $\omega\times]0,\varepsilon[$, given by

$$\begin{aligned} \vartheta_\varepsilon(x,t) &= \int_{-T}^{T} -F(t,\ell)\,g(x,\ell)\,d\ell\\ &= \int_{-T}^{T} -(E+H)(t,\ell)\,g(x,\ell)\,d\ell\\ &= \vartheta_{\varepsilon,1}(x,t)+\vartheta_{\varepsilon,2}(x,t), \end{aligned} \tag{5.19}$$

where, from Proposition 5.1, (5.2), and (5.13), we have

$$\begin{aligned} \|\vartheta_{\varepsilon,1}(\cdot,t)\|_{L^2(\omega)} &= \left(\int_\omega \left|\int_{-T}^{T} E(t,\ell)\,g(x,\ell)\,d\ell\right|^2 dx\right)^{1/2}\\ &\le c\left(\int_\omega \left|\frac{1}{\sqrt{t}}\,\|g\|_{L^2(]-T,T[)}\right|^2 dx\right)^{1/2}\\ &\le \frac{1}{\sqrt{t}} C_{\omega,T}\,\|\nabla w_o\|_{L^2(\Omega),} \end{aligned} \tag{5.20}$$

$$\begin{aligned} \|\vartheta_{\varepsilon,2}(\cdot,t)\|_{L^2(\omega)} &= \left(\int_\omega \left|\int_{-T}^{T} H(t,\ell)\,g(x,\ell)\,d\ell\right|^2 dx\right)^{1/2}\\ &\le \left(\|H(t,\cdot)\|^2_{L^2(]-T,T[)}\int_\omega \|g(x,\cdot)\|^2_{L^2(]-T,T[)}\,dx\right)^{1/2}\\ &\le e^{C\left(1+1/\varepsilon^2\right)}\ C_{\omega,T}\,\|\nabla w_o\|_{L^2(\Omega).} \end{aligned} \tag{5.21}$$

We conclude with an estimate of $(\int_\Omega |\nabla w(x,t)|^2dx)^{1/2} = (\int_\Omega |\int_{\mathbb{R}}(E+H)(t,\ell)\nabla y(x,\ell)d\ell|^2dx)^{1/2}$:

$$\begin{aligned} \left(\int_\Omega \left|\int_{\mathbb{R}} E(t,\ell)\,\nabla y(x,\ell)\,d\ell\right|^2 dx\right)^{1/2} &= \left(\int_\Omega \left|\int_{-T}^{T} E(t,\ell)\,\nabla y(x,\ell)\,d\ell\right|^2 dx\right)^{1/2}\\ &\le c\left(\int_\Omega \left|\frac{1}{\sqrt{t}}\,\|\nabla y(x,\cdot)\|_{L^2(]-T,T[)}\right|^2 dx\right)^{1/2}\\ &\le \frac{1}{\sqrt{t}} C_{\omega,T}\,\|\nabla w_o\|_{L^2(\Omega)}, \end{aligned} \tag{5.22}$$

(5.23)

$$\left(\int_\Omega \left|\int_{\mathbb{R}} H(t,\ell)\,\nabla y(x,\ell)\,d\ell\right|^2 dx\right)^{1/2} = \left(\int_\Omega \left|\int_{-T}^{T} H(t,\ell)\,\nabla y(x,\ell)\,d\ell\right|^2 dx\right)^{1/2}$$
$$\leq \left(\|H(t,\cdot)\|^2_{L^2(]-T,T[)} \int_\Omega \|\nabla y(x,\cdot)\|^2_{L^2(]-T,T[)}\,dx\right)^{1/2}$$
$$\leq e^{C\left(1+1/\varepsilon^2\right)}\, C_{\omega,T}\,\|\nabla w_o\|_{L^2(\Omega)}\,.$$

Remark 5.6. If we choose $w(x,t) = \int_{\mathbb{R}} F(t,\ell)\,y(x,\ell)\,d\ell$ with F given by (5.16) and y given by (5.17) where the operator Δ is replaced by $\Delta_{\mathcal{A}}$, then w is solution of (2.7).

Acknowledgments. The author would like to thank Professor C. Bardos and Professor E. Zuazua for their encouragement and for very useful criticism which allowed for the significant improvement of the presentation of the paper.

REFERENCES

[BdM] L. Boutet de Monvel, *Propagation des singularités des solutions d'équations analogues à l'équation de Schrödinger*, Lecture Notes in Math. 459, Springer, New York, 1975.

[B] N. Burq, *Contrôle de l'équation des plaques en présence d'obstacles strictement convexes*, Mém. Soc. Math. France (N.S.) 55, Marseille, 1993.

[BLR] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[F] C. Fabre, *Résultats de contrôlabilité exacte interne pour l'équation de Schrödinger et leurs limites asymptotiques: Application à certaines équations de plaques vibrantes*, Asymptot. Anal., 5 (1992), pp. 343–379.

[F-CZ] E. Fernandez-Cara and E. Zuazua, *The cost of approximate controllability for heat equations: The linear case*, Adv. Differential Equations, 5 (2000), pp. 465–514.

[FI] A.V. Fursikov and O. Yu Imanuvilov, *Controllability of Evolution Equations*, Lecture Notes Ser. 34, Seoul National University, Seoul, Korea, 1996.

[HL] M.A. Horn and W. Littman, *Boundary control of a Schrödinger equation with non-constant principal part*, in Control of Partial Differential Equations and Applications (Laredo, 1994), Lecture Notes in Pure and Appl. Math. 174, Dekker, New York, 1996, pp. 101–106.

[I] A.E. Ingham, *Some trigonometrical inequalities with applications to the theory of series*, Math. Z., 41 (1936), pp. 367–369.

[KS] L. Kapitanski and Y. Safarov, *Dispersive Smoothing for Schrödinger Equations*, preprint, 1995.

[Le] G. Lebeau, *Contrôle de l'équation de Schrödinger*, J. Math. Pures Appl. (9), 71 (1992), pp. 267–291.

[Li] J.-L. Lions, *Contrôlabilité exacte, stabilisation et perturbation des systèmes distribués*, Rech. Math. Appl. 1, Masson, Paris, 1998.

[LR1] G. Lebeau and L. Robbiano, *Contrôle exacte de l'équation de la chaleur*, Comm. Partial Differential Equations, 20 (1995), pp. 335–356.

[LR2] G. Lebeau and L. Robbiano, *Stabilisation de l'équation des ondes par le bord*, Duke Math. J., 86 (1997), pp. 465–491.

[LT] I. Lasiecka and R. Triggiani, *Optimal regularity, exact controllability and uniform stabilisation of Schrödinger equations with Dirichlet control*, Differential Integral Equations, 5 (1992), pp. 521–535.

[M] E. Machtyngier, *Exact controllability for the Schrödinger equation*, SIAM J. Control Optim., 32 (1994), pp. 24–34.

[MZ] S. Micu and E. Zuazua, *Boundary controllability of a linear hybrid system arising in the control of noise*, SIAM J. Control Optim., 35 (1997), pp. 1614–1637.

[P] K.-D. Phung, *Observabilité de l'équation de Schrödinger*, Prépublication LAPT/UMR 99-11, l'Université de Provence, Provence, France, 1999.

[Ro] L. Robbiano, *Fonction de coût et contrôle des solutions des équations hyperboliques*, Asymptot. Anal., 10 (1995), pp. 95–115.

[Ru] D. Russell, *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Stud. Appl. Math., 52 (1973), pp. 189–212.

[T] D. Tataru, *Carleman estimates and unique continuation for solutions to boundary value problems*, J. Math. Pures Appl. (9), 75 (1996), pp. 367–408.

[Z] E. Zuazua, *Exponential decay for the semilinear wave equation with localized damping in unbounded domains*, J. Math. Pures Appl. (9), 70 (1991), pp. 513–529.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 1, pp. 231–252

MOBILE POINT CONTROLS VERSUS LOCALLY DISTRIBUTED ONES FOR THE CONTROLLABILITY OF THE SEMILINEAR PARABOLIC EQUATION*

ALEXANDER KHAPALOV†

Abstract. It is well known now that a rather general semilinear parabolic equation with globally Lipschitz nonlinear term is both approximately and exactly null-controllable in $L^2(\Omega)$, when governed in a bounded domain by the locally distributed controls. In this paper we intend to show that, in fact, in one space dimension ($\Omega = (0,1)$) the very same results can be achieved by employing *at most* two mobile point controls with support on the curves properly selected within an arbitrary subdomain of $Q_T = (0,1) \times (0,T)$. We will show that such curves can be described by a certain differential inequality and the explicit examples are provided. We also discuss some extensions of our main results to the superlinear terms and to the case of several dimensions.

Key words. linear and semilinear parabolic equations, controllability, observability, point controls

AMS subject classifications. 93, 35

PII. S0363012999358038

1. Introduction.

1.1. Problem description. In modeling physical processes in bounded domains by controlled PDEs two types of controls—boundary and internal—are typically used. The boundary controls act upon the system from outside, while the internal controls act in the interior of the system's space domain. Each of these controls can be both distributed (i.e., depending both on x and t) and lumped (depending on t only). In general, it seems obvious that the former ones are more powerful. However, the latter ones seem more preferred in terms of applications. In this paper we are interested in the following question arising in this context: *Do the locally distributed controls really (always) work "better" than the internal point ones?*

We consider the Dirichlet initial-boundary value problem for the following one-dimensional parabolic equation:

$$u_t = u_{xx} + b(x,t)u_x + a(x,t)u + f(u) + (Bv)(x,t) \ \text{ in } Q_T = \Omega \times (0,T) = (0,1) \times (0,T),$$

$$\text{(S)} \qquad u(0,t) = u(1,t) = 0 \quad \text{in } \ (0,T), \quad u \ |_{t=0} = u_0 \in L^2(0,T),$$

$$a \in C(\bar{Q}_T), \ \ b \in C^{0,1}(\bar{Q}_T).$$

Here the term $(Bv)(x,t)$ models an internal control: B denotes the control operator (it describes how the "control mechanism" acts and is "fixed"), and v is the value of control.

Accordingly, in the case of locally distributed controls we have

$$(Bv)(x,t) = \chi_\omega(x)v(x,t), \tag{1.1}$$

*Received by the editors June 21, 1999; accepted for publication (in revised form) January 10, 2001; published electronically May 31, 2001. This work was supported in part by NATO grant CRG.CRG.972964.

http://www.siam.org/journals/sicon/40-1/35803.html

†Department of Pure and Applied Mathematics, Washington State University, Pullman, WA 99164-3113 (khapala@wsu.edu).

where $\omega = (l_1, l_2)$ is the given subdomain of $\Omega = (0,1)$ on which control v is supported (so $\chi_\omega(x)$ is the characteristic function of ω), and $v = v(x,t)$ is the function of *both* the time and space variables.

To the contrary, the *lumped* controls, as it follows from their name, are the functions of *time only.* Typically two kinds of internal lumped controls are employed:

(a) *the point controls*

$$(Bv)(x,t) = v(t)\delta(x - s(t)), \tag{1.2}$$

where $s(t)$ is the preassigned point support of control $v = v(t)$ at time t (so $\delta(x-s(t))$ denotes Dirac's mass concentrated at $s(t)$); and

(b) *the averaged (or zone) controls*

$$(Bv)(x,t) = v(t)\chi_{\omega(t)}(x), \tag{1.3}$$

where $\omega(t) \subset \Omega$ is the given support of control $v = v(t)$ at time t. Note that in the latter case (1.3) the *same* value $v(t)$ of the control function applies at every point of the set $\omega(t)$.

If $s(t) \equiv x_0$ in (1.2) and $\omega(t) \equiv \omega$ in (1.3) for all $t \in (0,T)$, the lumped controls are "static"; otherwise, they are called "mobile" or "scanning."

The lumped controls are strongly motivated by numerous applications. They can be regarded as a degenerate class of locally distributed ones. (The latter in turn can be viewed as a collection of infinitely many former ones.) To analyze these two types of control, one usually needs quite different methods. For example, the controllability property by means of the locally distributed controls is essentially based on the unique continuation property of solutions to the linear parabolic equation from an *open* set. This property cannot, obviously, be associated with the case of lumped controls.

Generally, one cannot expect equally strong results for these two types of controls. Nonetheless, *we intend to show below that for a rather general system like* (S)*, whenever* (S) *is globally controllable by the locally distributed controls* (1.1) *supported in* $\omega \times (0,T)$*, it is also globally controllable by means of at most two mobile point controls* (1.2) *(see* (1.6) *below), which are supported on the curves* $s_1(\cdot)$ *and* $s_2(\cdot)$ *suitably selected within the very same set* $\omega \times (0,T)$. Our special concern is their explicit description.

Let us recall in this respect that the geometry of control support is critical for controllability by means of lumped controls. Namely, unlike the locally distributed ones, the outcome in terms of controllability for the lumped controls is generally *unstable* with respect to their support. For example, it is well known (see [4], [7], [26]) that the standard heat equation in $\Omega = (0,1)$ with the static point control $v(t)\delta(x - x_0), x_0 \in (0,1)$, is *not* controllable for any rational location x_0. However, it becomes approximately controllable in $L^2(0,1)$ at any positive time T for x_0 being any irrational number and exactly null-controllable for almost all irrational x_0. (These results are based on the explicit Fourier series approach and do not apply to a *linear* system like (S) with $f = 0$, convection term, and time-varying coefficients.) In this respect, nonetheless, our main results below deal with quite feasible geometric Assumptions 1.1–1.3 that are "stable" with respect to either the C-, or, at most, $C^{2,1}$-topology for the choice of control curves.

Before we proceed any further, let us recall the classical definitions of controllability.

Assume that system (S) has a unique solution in the space $C([0,T];H)$ for any $u_0 \in H$ and $v \in V$, where H and V are some Banach spaces.

DEFINITION 1.1. *Given $T > 0$, system* (S) *is said to be approximately controllable in H at time T if for an arbitrary $\varepsilon > 0$ and $u_0, u_T \in H$, there is a suitable $v \in V$ such that for the corresponding solution to* (S) *we have*

$$\| u(\cdot, T) - u_T \|_H \leq \varepsilon. \tag{1.4}$$

DEFINITION 1.2. *Assume $f(0) = 0$. Given $T > 0$, system* (S) *is said to be exactly null-controllable in H at time T if for an arbitrary $u_0 \in H$ there is a suitable $v \in V$ such that the corresponding solution to* (S) *reaches the zero state (the equilibrium of* (S)*) at time T; that is,*

$$u(\cdot, T) = 0. \tag{1.5}$$

Everywhere below we deal with $H = L^2(0,1)$ and controls either in $L^2(\omega \times (0,T))$ in the case of locally distributed controls or in $L^2(0,T)$ in the case of point controls.

1.2. Main results. We further assume that the control term Bv is represented by no more than two point controls as follows:

$$(Bv)(x,t) = v_1(t)\delta(x - s_1(t)) + v_2(t)\delta(x - s_2(t)). \tag{1.6}$$

To ensure both the mathematical well-posedness of the system at hand (see Theorem A1, below) and to preserve the physical meaning of $s_1(\cdot)$ and $s_2(\cdot)$ as of the trajectories for the point controls, the following conditions are assumed throughout the remainder of this paper.

Assumption 1.1. (i) The functions $s_1 = s_1(t)$ and $s_2 = s_2(t)$ are defined on some segments lying in $[0,T]$ and are continuous functions with values in $[0,1]$. (This means that the actual controls act only where s_i's are defined and are inactive otherwise.)

(ii) Any "horizontal" line $\{(x,t) \mid t = t_*\}$ can cross any of the trajectories $s_i(\cdot)$ at most at *one* point. (Indeed, if $s_i(\cdot), i = 1,2$ represent the paths of point controls, then at every moment of time these controls can be supported at the single points only; see Figures 1.1 and 1.2.)

Our first two results—Theorems 1.1 and 1.2—deal with the linear version of system (S) and are, respectively, about the approximate and exact null-controllability properties. Theorem 1.1 employs the following additional geometric condition.

Assumption 1.2. (meeting condition). There is an interval $[t_1, t_2] \subset [0,T)$ such that the functions $x = s_1(t)$ and $x = s_2(t)$ are continuous and one-to-one on it with values in $[0,1]$ and

$$s_1(t_2) = s_2(t_2), \quad s_1(t) < s_2(t) \quad \forall t \in [t_1, t_2).$$

This condition means that the two control point controls at hand arrive ("meet") at time t_2 at the same point, as shown in Figure 1.2, for example.

THEOREM 1.1 (approximate controllability). *Let $T > 0$ be given, and let Assumptions* 1.1 *and* 1.2 *hold. Then the linear version of system* (S)*, namely, with $f = 0$ and Bv as in* (1.6) *(or, system* (2.2) *below, which is the same) is approximately controllable in $L^2(0,1)$ at time T. In turn its dual system* (2.5) *with two point sensors is observable (in the sense described in* (2.7) *below).*

Our next exact null-controllability result makes use of the following conditions.

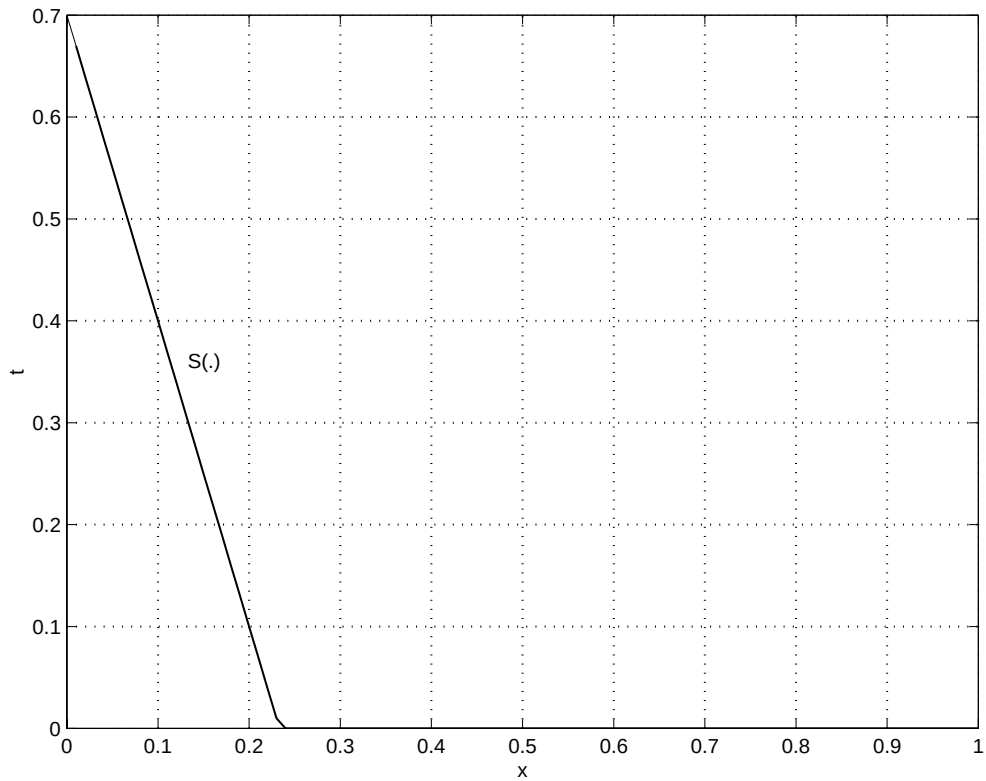


FIG. 1.1. *One point control.*

Assumption 1.3. (i) In addition to Assumption 1.2, assume that on the interval $[t_1, t_2] \subset [0, T]$ the functions $s_1(\cdot)$ and $s_2(\cdot)$ are, respectively, strictly monotone increasing and decreasing.

(ii) Assume that the connected geometric curve $s(\cdot)$ in $\bar{Q}_T \subset R^2$, composed from $s_1(\cdot)$ and $s_2(\cdot)$ on the interval $[t_1, t_2]$ (recall that they "meet" at t_2 by Assumption 1.2), is smooth and admits the following representation:

$$s(\cdot) = \{(x,t) \mid (x,t) \in \mathcal{A},\ F(x,t) = 0\},$$

where $\mathcal{A} = \{(x,t) \mid x \in [0,1], t \in [t_1, t_2]\}$ and F is an element of $C^{2,1}(\mathcal{A})$ and

$$-F_t(x,t) + 2b(x,t)F_x(x,t) - \{b(x,t)F(x,t)\}_x + F_{xx}(x,t) \le 0 \quad \forall (x,t) \in \mathcal{A}^*, \tag{1.7}$$

$$F(x,t) > 0 \quad \forall (x,t) \in \operatorname{int}\{\mathcal{A}^*\}, \quad F(x,t) \le 0 \quad \forall (x,t) \in \mathcal{A} \backslash \mathcal{A}^*, \tag{1.8}$$

where $\mathcal{A}^* = \{(x,t) \mid s_1(t) \le x \le s_2(t),\ t \in [t_1, t_2]\}$.

Assumption 1.3(i) means that any "vertical" line ($x =$ constant) within the layer $\mathcal{A}$ crosses at most one of the curves $s_1(\cdot), s_2(\cdot)$ at no more than one point. This assumption (as well as (ii)) can be relaxed somewhat: we assume it to simplify our further integration by parts. Several explicit examples illustrating Assumption 1.3 are given in section 6 below.

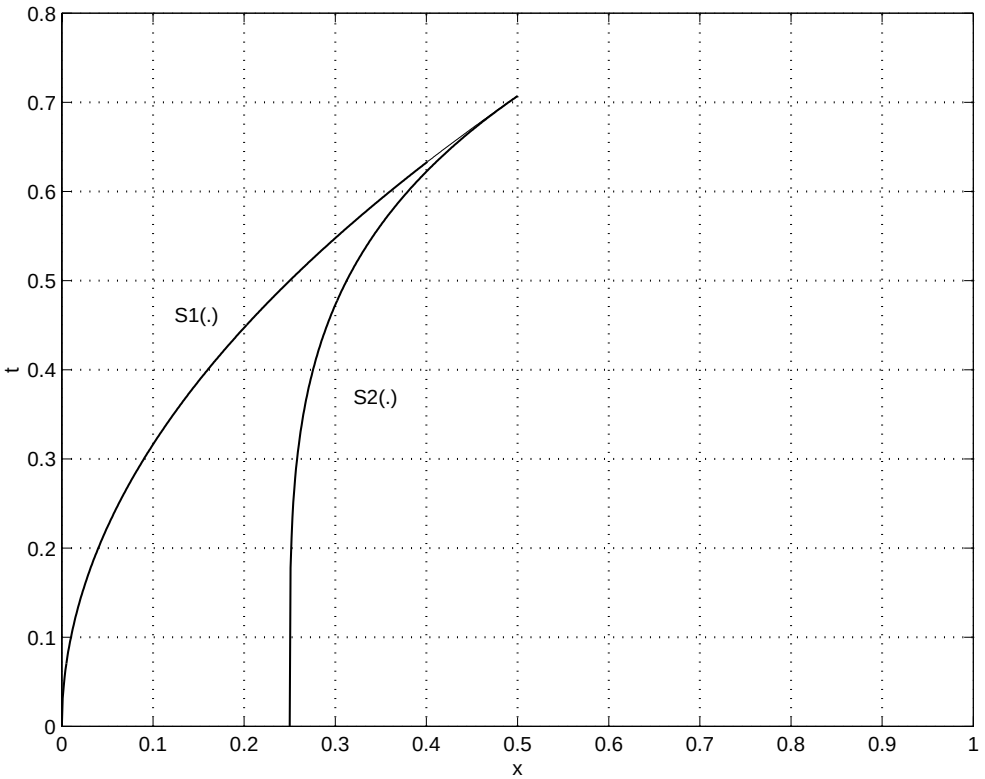


FIG. 1.2. *Two point controls.*

THEOREM 1.2 (exact null-controllability). *Let $T > 0$ be given, and let Assumptions* 1.1–1.3 *hold. Then the linear version of system* (S)*, namely, with $f = 0$ and Bv as in* (1.6) *(see system* (2.2) *below), is exactly null-controllable in $L^2(\Omega)$ at time T.*

The approximate and exact null-controllability properties of a system like (S) with globally Lipschitz f and the locally distributed controls were established (in several space dimensions) in [14], [13], based on the method of Carleman estimates (see also [24] for the case of the standard heat equation). For a different variational approach relevant solely to the issue of approximate controllability we refer to [6]. In this respect we have the following "lumped" result.

THEOREM 1.3 (the semilinear case). *Suppose that f is globally Lipschitz, differentiable at zero, and vanishes at zero: $f(0) = 0$. Let $T > 0$ be given, and let Assumptions* 1.1–1.3 *hold. Then system* (S) *with Bv as in* (1.6) *(or,* (5.1) *below, which is the same) is exactly null-controllable in $L^2(\Omega)$ at time T.*

Remark 1.1.

- In Corollaries 2.1 and 5.1 and in Remark 2.1 we distinguish the cases when all the above results hold with a single point control active only.
- In section 7.4 we also discuss possible extensions of Theorem 1.3 to some superlinear growth rates for f, which were the subject of the recent works [16], [8], [20], [21], [9], [2], [1], [11] dealing with the locally distributed controls.
- For the one-dimensional semilinear heat equation with the *uniformly bounded* globally Lipschitz term, governed by the static point controls like in (1.2)

with $s(t) \equiv x_0$, the approximate controllability was shown in [28].

- In [19] the approximate controllability was established for the *static* zone lumped controls like in (1.3), assuming that either (a) f is a sublinear $\log^r \mid u \mid$-like function or (b) it is superlinear like $\mid f(x,t,u,u_x) \mid \leq \beta(t)(\mid u \mid^{r_1} + \mid u_x \mid^{r_2})$, where r_1, r_2 can exceed 1 and $\beta(t) \to 0$ faster than any of $e^{-\nu/t}$, $\nu > 0$ as $t \to 0$, *and* the dissipativity condition holds for such f. In [22] the last restriction on the growth of f in t was avoided, assuming that $b = 0$, $a = a(t)$ in (S), and, in addition to the locally distributed control (1.3), one can use the coefficient $a = a(t)$ as an extra bilinear control.

In the appendix we prove the following supporting existence and regularity result which insures the well-posedness of the boundary problem (S), (1.6). (To separate it from our main controllability results and to indicate that its proof is delegated to the appendix, we mark it as "A.1.")

THEOREM A.1. *Let Assumption* 1.1 *hold, let f be globally Lipshitz, and let $f(0) = 0$. Then system* (S), (1.6) *admits a unique generalized solution in $C([0,T]; L^2(0,1)) \bigcap L^2(0,T; H_0^1(0,1))$ for which the following two estimates hold:*

$$\| u \|_{L^6(Q_T)} \leq c(T) \left(\| u_0 \|_{L^2(\Omega)} + \| v_1 \|_{L^2(0,T)} + \| v_2 \|_{L^2(0,T)} \right), \tag{1.9a}$$

$$\| u \|_{C([0,T];L^2(0,1))} + \left(\int_0^T \int_0^1 u_x^2 \, dxdt \right)^{1/2} \leq c(T) \left(\| u_0 \|_{L^2(\Omega)} + \| v_1 \|_{L^2(0,T)} + \| v_2 \|_{L^2(0,T)} \right), \tag{1.9b}$$

where $c(T)$ is nondecreasing and depends also on the $C(\bar{Q}_T)$-norms of $a(x,t)$ and $b(x,t)$ and the Lipschitz constant of f.

The remainder of the paper is organized as follows. In sections 2–4 we deal with the case when (S) is linear. Sections 4 and 5 consider the semilinear case with globally Lipschitz f. Several explicit examples are given in section 6. In section 7 we analyze some possible extensions of our main results.

2. The linear case: The dual observed system. In the linear case, to separate two types of controls, we represent (S) with the locally distributed controls as

$$p_t \ = \ p_{xx} \ + \ b(x,t)p_x \ + \ a(x,t)p \ + \ v(x,t)\chi_\omega(x) \quad \text{in} \quad Q_T, \tag{2.1}$$

$$p(0,t) = p(1,t) = 0 \quad \text{in} \ (0,T), \quad p \mid_{t=0} = p_0 \in L^2(0,1), \ \ v \in L^2(\omega \times (0,T)),$$

while with *two* point controls (1.6) system (S) will look as follows:

$$\begin{aligned} u_t \ &= u_{xx} + b(x,t)u_x \ + \ a(x,t)u \ + \ v_1(t)\delta(x - s_1(t)) \\ &\quad + \ v_2(t)\delta(x - s_2(t)) \quad \text{in} \quad Q_T, \end{aligned} \tag{2.2}$$

$$u(0,t) = u(1,t) = 0 \quad \text{in} \ (0,T), \quad u \mid_{t=0} = u_0 \in L^2(0,1), \ \ v_1, v_2 \in L^2(0,T).$$

It is known for (2.1) (and (2.5); see, e.g., [23]) and it is shown in the appendix for (2.2) that these systems possess unique solutions in the space $\Xi = C([0,T]; L^2(0,1)) \bigcap L^2(0,T; H_0^1(0,1))$, which we further endow with the norm

$$\| z \|_\Xi = \| z \|_{C([0,T];L^2(0,1))} \ + \ \left(\int_0^T \int_0^1 z_x^2 \, dxdt \right)^{1/2}, \tag{2.3}$$

and the estimates like (1.9a) and (1.9b) hold.

2.1. Approximate controllability: The linear case. Clearly, to analyze this property in this case, it is sufficient to do it when the initial data is the zero-state:

$$p_0 = 0 \ \text{ for } \ (2.1) \ \text{ and } \ u_0 = 0 \ \text{ for } \ (2.2). \tag{2.4}$$

Then, as is well known (see, e.g., [5]), the issue of global approximate controllability of systems (2.1) and (2.2) is tantamount to the observability property of the corresponding dual boundary problem:

$$y_t \ = \ y_{xx} \ - \ (b(x,T-t)y)_x \ + \ a(x,T-t)y \quad \text{ in } \ Q_T, \tag{2.5}$$

$$y(0,t) = y(1,t) = 0, \quad y \ |_{t=0} = y_0 \in L^2(0,1).$$

Namely, for (2.1) (or, for (2.1), (2.4), which is the same) system (2.5) must be *observable* with respect to the locally distributed observation over $\omega \times (0,T)$; that is,

$$y \equiv 0 \ \text{ in } \ \omega \times (0,T) \ \ \Longrightarrow \ \ y \equiv 0 \ \text{ in } \ Q_T. \tag{2.6}$$

In turn, for system (2.2) the *dual* observability property means that

$$y \equiv 0 \ \text{ along } \ s_1(T-\cdot), s_2(T-\cdot) \ \ \Longrightarrow \ \ y \equiv 0 \ \text{ in } \ Q_T. \tag{2.7}$$

This classical conclusion follows from the duality relations

$$\int_0^1 p(x,T)y_0(x)dx \ = \ \int_0^T \int_\omega v(x,t)y(x,T-t)dxdt,$$

$$\int_0^1 u(x,T)y_0(x)\,dx \ = \ \int_0^T v_1(t)y(s_1(t),T-t)\,dt \ + \ \int_0^T v_2(t)y(s_2(t),T-t)\,dt,$$

which can be derived by multiplying accordingly (2.1), (2.4) and (2.2), (2.4) by $y(x,T-t)$ and further integration by parts over Q_T.

Note now that the statement (2.6) is equivalent to the unique continuation property of solutions to (2.5) (also possessing the backward uniqueness property, e.g., [3]) from $\omega \times (0,T)$ to Q_T, which holds for any nondegenerate interval $\omega = (l_1, l_2)$.

We now intend to show that for any nondegenerate $\omega = (l_1, l_2) \subseteq (0,1)$ one can select two curves $s_1(\cdot)$ and $s_2(\cdot)$, lying in $\omega \times (0,T)$, which ensure (2.7), and hence the approximate controllability of (2.2). Our results here are linked to the geometric Assumptions 1.2 and 1.3 on these curves. In this subsection we employ the former.

We start with the discussion of the well-posedness of point observations.

Note that we have $(b(\cdot,T-\cdot)y)_x, a(\cdot,T-\cdot)y \in L^2(Q_T)$, while by the smoothing effect $u(\cdot,t_*) \in H_0^1(0,1)$ for any $t_* \in (0,T]$ (see, e.g., [23, pp. 178–180], [25]). Hence solutions to (2.5) on (t_*,T) can be viewed as the ones of the standard heat equation with the source term in $L^2((0,1)\times(t_*,T))$ and the initial data in $H_0^1(0,1)$. Hence they are continuous on any $[0,1]\times[t_*,T]$, i.e., on $[0,1]\times(0,T]$ for all $y_0 \in L^2(0,1)$. Moreover, the following estimate holds for any $t_* \in (0,T)$ (e.g., [25]):

$$\| \, y \, \|_{C([0,1]\times[t_*,T])} \le \ c(t_*) \left(\| \, y(\cdot,t_*) \, \|_{H_0^1(0,1)} \ + \ \| \, (b(\cdot,T-\cdot)y)_x + a(\cdot,T-\cdot)y \, \|_{L^2(Q_T)} \right), \tag{2.8a}$$

where the symbol $c(s)$ denotes any generic (i.e., they can be different) finite positive function of $s > 0$. (The symbol C is reserved for a generic positive constant.)

On the other hand, from the classical regularity results (based on the Fourier series approach) it is known (again viewing (2.5) as the standard heat equation with the source term in $L^2(Q_T)$) that

$$\| y(\cdot,t_*) \|_{H_0^1(0,1)} \le c(t_*) \big(\| y_0 \|_{L^2(0,1)} + \| (b(\cdot,T-\cdot)y)_x + a(\cdot,T-\cdot)y \|_{L^2(Q_T)} \big) \le c(t_*) \| y_0 \|_{L^2(0,1)}, \tag{2.8b}$$

(recall $c(\cdot)$ is generic), where we also used the following classical estimate for solutions of (2.5) (see (1.9a), (1.9b), and (2.3)):

$$\| y \|_{\Xi} \le c(T) \| y_0 \|_{L^2(\Omega)} .$$

Combining all of the above, we obtain

$$\| y \|_{C([0,1]\times[t_*,T])} \le c(t_*) \| y_0 \|_{L^2(0,1)} \quad \forall t_* \in (0,T), \tag{2.9}$$

which ensures the well-posedness of point observation.

Note also that for any continuous curve $s(t) \in (0,1), t \in (0,T)$ satisfying Assumption 1.1, by the continuity of embedding $H_0^1(0,1) \subset C[0,1]$, and by Poincaré–Friedrichs's inequality, we have

$$\| y(s(T-\cdot),\cdot) \|_{L^2(0,T)} \le C \| y \|_{L^2(0,T;H_0^1(0,1))} .$$

We are now ready to prove Theorem 1.1.

Proof of Theorem 1.1. *The case of the standard heat equation.* Note that by the smoothing effect, the solutions to (2.5) are classical in the cylinder

$$\mathcal{B} = \{(x,t) \mid x \in [0,1], t \in [T-t_2, T-t_1]\} = \{(x,t) \mid (x,T-t) \in \mathcal{A}\}$$

for any $y_0 \in L^2(0,1)$. This permits us to apply the classical maximum principle in it, which states that the maximum and minimum of y in (the closed set) $\mathcal{B}$ are attained on the boundary

$$\Gamma_{T-t_2,T-t_1} = \{(x,t) \mid x = 0,1; t \in [T-t_2, T-t_1]\} \bigcup \{(x,t) \mid t = T-t_2, x \in [0,1]\}$$

of this cylinder. Now we would like to remind the reader of the classical proof of this statement in order to show that it remains true for the set

$$\mathcal{B}^* = \{(x,t) \mid (x,T-t) \in \mathcal{A}^*\}$$

in place of $\mathcal{B}$ as well.

Assume it is false, e.g., that y reaches its maximum in $\mathcal{B}$, say, M, at the point $(x_0,t_0) \in \mathcal{B} \backslash \Gamma_{T-t_2,T-t_1}$, where $M = m + \varepsilon$, $m = \max\{y(x,t) \mid (x,t) \in \Gamma_{T-t_1,T-t_2}\}$, and $\varepsilon > 0$. Introduce an auxiliary function $v(x,t) = y(x,t) + k(t_0 - t)$.

Now select a positive parameter k sufficiently small to ensure that

$$\| y - v \|_{C(\mathcal{B})} \le \varepsilon/2.$$

Then v also reaches its maximum in $\mathcal{B}$, say, at the point $(x_1,t_1) \in \mathcal{B} \backslash \Gamma_{T-t_2,T-t_1}$, because

$$v(x_1,t_1) \ge v(x_0,t_0) = m + \varepsilon,$$

while v cannot exceed $m+\varepsilon/2$ on $\Gamma_{T-t_2,T-t_1}$. At the maximum point (x_1,t_1) we have

$$y_{xx}(x_1,t_1)=v_{xx}(x_1,t_1)\ \le\ 0\ \le\ v_t(x_1,t_1)=y_t(x_1,t_1)-k.$$

Therefore, the heat equation does not hold at (x_1,t_1), which is a contradiction.

Analyzing this proof, the reader can see that, under Assumptions 1.1 and 1.2, it holds true without any changes for the set $\mathcal{B}^*=\{(x,t)\mid s_1(T-t)\le x\le s_2(T-t),\ t\in[T-t_2,T-t_1]\}$ in place of $\mathcal{B}$ and for the combined *connected* (by Assumption 1.2) geometric curve $s(\cdot)=s_1(T-\cdot)\bigcup s_2(T-\cdot)$ in place of $\Gamma_{T-t_2,T-t_1}$ as well. Thus we establish the following:

$$\begin{aligned}\|\ y\ \|_{C(\mathcal{B}^*)}\ &\le\ \max_{i=1,2}\{\|\ y(s_i(T-\cdot),\cdot)\ \|_{C[T-t_2,T-t_1]}\}\\ &=\ \max_{i=1,2}\{\|\ y(s_i(\cdot),T-\cdot)\ \|_{C[t_1,t_2]}\}.\end{aligned}\tag{2.10}$$

In other words, if the solution y to (2.5) vanishes on the curves $s_1(T-t)$ and $s_2(T-t),t\in[T-t_2,T-t_1]$, "emitted" from the point $(s_1(t_2)=s_1(t_2),T-t_2)$, then y vanishes everywhere in $\mathcal{B}^*$.

Furthermore, by the unique continuation property (e.g., [27]) y vanishes in the horizontal layer $\mathcal{B}$. By backward (and forward) uniqueness this solution vanishes in Q_T. Thus we have (2.7) and hence the approximate controllability of dual (2.2).

The general case. Step 1. Represent the equation in (2.5) as follows:

$$y_t\ =\ y_{xx}\ -\ b(x,T-t)y_x\ +\ d(x,T-t)y,$$

where $d(x,T-t)=a(x,T-t)-b_x(x,T-t)$.

We know (see (2.8a–b)–(2.9)) that $y(x,T-t_2)\in H_0^1(0,1)$. Let us assume for a while that $y(x,T-t_2)$ and the coefficients $b(x,T-t)$ and $d(x,T-t)$ are infinitely many times continuously differentiable in $\mathcal{B}$. Then y will be the classical solution to (2.5) in the latter set; see, e.g., [12, p. 65].

Moreover, without loss of generality, we can assume that $d(x,T-t)\le 0$ in $\mathcal{B}$. (Indeed, this can be achieved by a simple change of variable $\hat{y}\to y:\ \hat{y}=e^{\lambda t}y$ with properly selected parameter λ.) If so, y satisfies the maximum principle in any set $\mathcal{P}_{(t)}=\mathcal{B}^*\bigcap\{(x,t)\mid t\ge t_*\}$, where $t_*\in(T-t_2,T-t_1]$ (see [12], pp. 34–35); that is, $|\ y(x,t)\ |$ reaches its maximum on the lower boundary of the set $\mathcal{P}_{(t)}$:

$$|\ y(x,t)\ |\le\ \max\{\|\ y(s_i(T-\cdot),\cdot)\ \|_{C[T-t_*,T-t_1]},\ i=1,2;\ \max_{r\in[s_1(T-t_*,s_2(T-t_*)]}|\ y(r,t_*)\ |\}$$

$$\forall(x,t)\in\mathcal{P}_{(t)};$$

see Figure 2.1. Since $s_1(t_2)=s_2(t_2)$ with $t_*\to T-t_2$, this estimate implies (2.10) whenever y is the classical solution.

Step 2. Consider now any solution to (2.5). Similar to the argument leading to (2.9), we can show that it can be approximated in the $C(\mathcal{B})$-norm by a sequence of the classical solutions to the boundary problem

$$y_{jt}\ =\ y_{jxx}\ -\ b_j(x,T-t)y_{jx}\ +\ d_j(x,T-t)y_j\quad\text{in}\ \ \mathcal{B},$$

$$y_j(0,t)=y_j(1,t)=0,\quad y_j\ |_{t=T-t_2}=y_{j0},$$

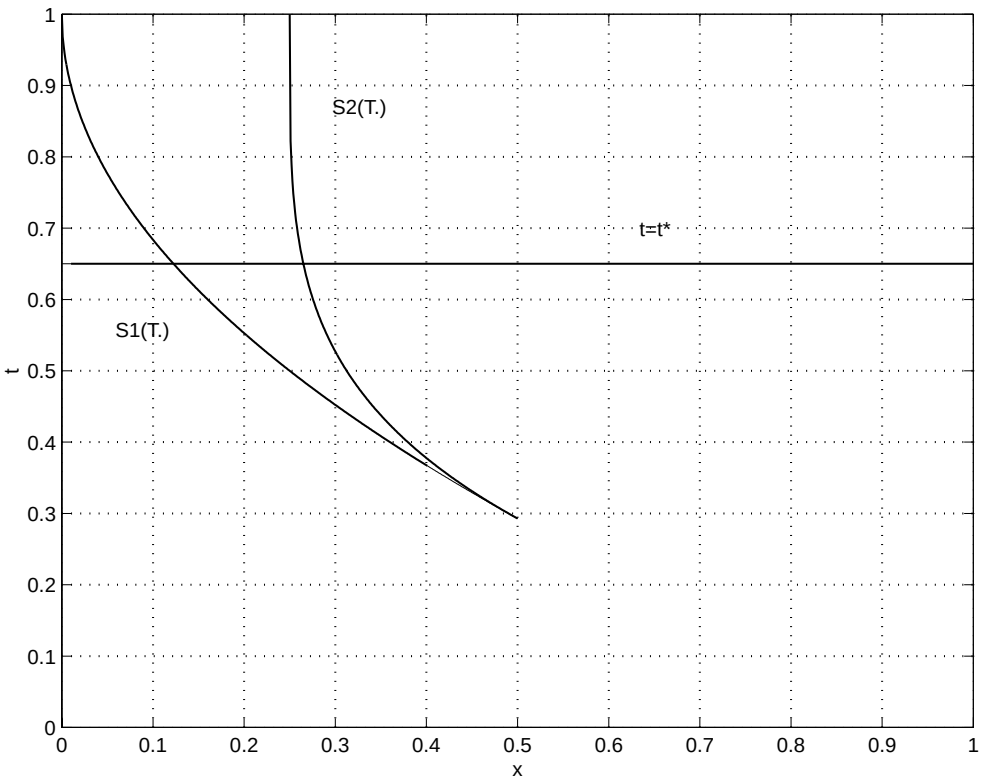


FIG. 2.1.

where $b_j(x, T-t), d_j(x, T-t)$, and $y_{j0}(x)$ are infinitely many times continuously differentiable functions converging, respectively, in the $C(\mathcal{B})$-norm to $b(x, T-t)$ and $d(x, T-t)$, and in the $H_0^1(0,1)$-norm to $y_0^*(x) = y(x, T-t_2)$. Moreover, d_j's can be selected to preserve the inequality $d_j(x, T-t) \leq 0$ in $\mathcal{B}$ for all $j = 1, \ldots$.

Indeed, to approximate the coefficients, one may first continuously extend their domain to a larger set and then use a suitable averaging procedure. To ensure the above-mentioned inequality for d_j's, if necessary, $d(x, T-t)$ should be approximated first by a sequence of continuous functions for which this inequality is strict in $\mathcal{B}$. To approximate y_0^*, it is sufficient to recall that the infinitely many times continuously differentiable functions with compact support in $(0,1)$ are dense in $H_0^1(0,1)$.

Then, for the difference $z_j = y - y_j$, we have the following boundary problem:

$$z_{jt} \;=\; z_{jxx} \;-\; b_j(x, T-t) z_{jx} \;+\; d_j(x, T-t) z_j \;+ f_j \quad \text{in} \quad \mathcal{B}, \tag{2.11}$$

$$z_j(0,t) = z_j(1,t) = 0, \quad z_j \;|_{t=T-t_2} = y_0^* - y_{j0},$$

where $f_j(x,t) = -(b(x, T-t) - b_j(x, T-t)) y_x + (d(x, T-t) - d_j(x, T-t)) y$. An estimate like (1.9a)–(1.9b) applied to (2.5) implies that

$$\| f_j \|_{L^2(\mathcal{B})} \leq \; C \left(\| b - b_j \|_{C(\mathcal{A})} + \| d - d_j \|_{C(\mathcal{A})} \right) \| y_0 \|_{L^2(0,1)}$$

for some $C > 0$.

Similar to (2.8a) and (2.8b), that is, viewing (2.11) as the standard heat equation with the source term $-b_j(x,T-t)z_{jx}+d_j(x,T-t)z_j+f_j(x,t)$ in $L^2(\mathcal{B})$, we obtain that

$$\| z_j \|_{C(\mathcal{B})}$$

$$\leq C_1 \left(\| z_j(\cdot,T-t_2) \|_{H_0^1(0,1)} + \| -b_j(\cdot,T-t)z_{jx}+d_j(\cdot,T-t)z_j+f_j \|_{L^2(\mathcal{B})} \right)$$

$$\leq C_2(\| z_j(\cdot,T-t_2) \|_{H_0^1(0,1)} + (\| b_j \|_{C(\mathcal{A})} + \| d_j \|_{C(\mathcal{A})}) \| z_j \|_{\Xi(\mathcal{B})}$$

$$+ \left(\| b-b_j \|_{C(\mathcal{A})} + \| d-d_j \|_{C(\mathcal{A})}\right) \| y_0 \|_{L^2(0,1)}])$$

for some positive constants $C_i, i=1,2$, where $\Xi(\mathcal{B})$ is the restriction of the space Ξ defined on Q_T to $\mathcal{B}$). Now, applying the estimates like (1.9a) and (1.9b) to z_j as the solution of (2.11), we derive that

$$\| z_j \|_{\Xi(\mathcal{B})} \leq C \left(\| z_j(\cdot,T-t_2) \|_{L^2(0,1)} + \| f_j \|_{L^2(\mathcal{B})} \right)$$

for some $C>0$, with $\| f_j \|_{L^2(\mathcal{B})}$ already evaluated in the above. Hence

$$\lim_{j\to\infty} \| z_j \|_{C(\mathcal{B})} = \lim_{j\to\infty} \| y-y_j \|_{C(\mathcal{B})} = 0. \tag{2.12}$$

Step 3. Now, if y vanishes on the curves $s_1(T-t)$ and $s_2(T-t)$, $t\in[T-t_2,T-t_1]$, lying in $\mathcal{B}$, then, by (2.12),

$$\lim_{j\to\infty} \max_{i=1,2} \{ \| y_k(s_i(T-\cdot),\cdot) \|_{C[T-t_2,T-t_1]} \} = 0. \tag{2.13}$$

Combining (2.13) with (2.10) applied to the classical solutions $y_j, j=1,\dots$ (see Step 1), yields that

$$\lim_{j\to\infty} \| y_j \|_{C(\mathcal{B}^*)} = 0,$$

which in view of (2.12) means that y vanishes in $\mathcal{B}^*$.

The end of the proof in the general case of (2.5) is identical to the case of the standard heat equation in the above with one correction: in the general case we don't have enough regularity to make use of the unique continuation result of [27]. Instead, we will use the estimate (4.4) below with $D = \operatorname{int}\mathcal{B}^*$, which gives $y|_{t=T}=0$ and, by duality discussed in the beginning of this subsection, the approximate controllability of (2.2). To obtain (2.7), one needs also to use the backward uniqueness property of solutions to (2.5) (e.g., [3]). This completes the proof of Theorem 1.1. □

Corollary 2.1. *Note that, due to the zero boundary condition in* (2.5) *(i.e., y vanishes on the lines $x=0$ and $x=1$), the conclusion of Theorem* 1.1 *also holds when we have just one point control, whose trajectory "hits" the boundary of* $(0,1)$ *at some time t_2, while approaching it from the interior of $\Omega=(0,1)$. (The actual proof of Theorem* 1.1 *is given below for this case.)*

2.2. Exact null-controllability: The linear case. For a system like (2.1) with varying coefficients and the locally distributed controls this property was shown in [14], [13] (see [24] for the standard heat equation). The methods of these works employ Carleman's estimates to enhance the unique continuation property, namely, by establishing the following estimate for the solutions to (2.5):

$$\| y(\cdot,T) \|_{L^2(0,1)} \leq C \| y \|_{L^2(\omega\times(0,T))} . \tag{2.14}$$

Note that (2.14) also provides the unique continuation property of y from $\omega\times(0,T)$ to Q_T, and hence the approximate controllability of (2.1). However, it means more, namely, that the operator which maps the trace y on $\omega \times (0,T)$ to $y(\cdot,T)$ on $(0,1)$ is well defined and continuous with respect to the spaces in (2.14). By duality, this classically yields the exact null-controllability of (2.1) in $L^2(0,1)$ at time T.

In this respect our next goal is to derive an estimate analogous to (2.14) for two point sensors "dual" of the two point controls in (2.2), which, by duality, is tantamount to Theorem 1.2. Our main result here is as follows.

THEOREM 2.2 (observability estimate). *Let $T > 0$ be given, and let Assumptions* 1.1–1.3 *hold. Then for any solution to the system* (2.5) *we have the following observability estimate (also implying* (2.7)*)*:

$$\| y(\cdot,T) \|_{L^2(0,1)} \leq C \left(\int_{T-t_2}^{T-t_1} (y^2(s_1(T-t),t) + y^2(s_2(T-t),t))dt \right)^{1/2} . \tag{2.15}$$

The proof of Theorem 2.2 is given in the next two sections.

Remark 2.1. Both Theorems 2.2 and 1.2 can be extended as in Corollary 2.1.

3. An auxiliary observability estimate (3.6). In this section our goal is to derive the estimate (3.6) under the assumptions of Theorem 2.2. For simplicity we will further assume that $t_1 = 0, t_2 = T$ and formally set $s_1(t) = 0$ for $t \in [0,T]$ (as in Corollary 2.1; otherwise, see Remark 3.1).

Let

$$\| a(x,t) \|_{C(\bar{Q}_T)} = L. \tag{3.1}$$

Let F be as in Assumption 1.3. Given $T > 0$, put

$$\varphi(x,t) = \begin{cases} F(x,T-t) & \text{for } (x,t) \in \mathcal{B}^*, \\ 0 & \text{for } (x,t) \in \mathcal{B}\backslash\mathcal{B}^*. \end{cases} \tag{3.2a}$$

Under Assumption 1.3, $\varphi \in C^{2,1}(\mathcal{B}^*)$ is nonnegative, vanishes in $\mathcal{B}\backslash\mathcal{B}^*$, and

$$\varphi = 0 \ \text{ on } \ s_2(T-\cdot) \ \text{ and } \ \varphi > 0 \ \text{ in } \ \operatorname{int}\{\mathcal{B}^*\}. \tag{3.2b}$$

Multiplication of (2.5) by φy and further integration by parts over Q_t (for all $t \in [0,T]$) or over the set $\mathcal{B}_t^* = \mathcal{B}^* \bigcap \{(x,\tau) \mid 0 < \tau < t\}$, which is the same, yield

$$\frac{1}{2}\int_0^1 \varphi(x,t)y^2(x,t)dx$$

$$= \int_0^t \int_0^1 \left(a(x,T-\tau)\varphi y^2 + \frac{1}{2}\varphi_\tau y^2 + y_{xx}\varphi y \ - \ (b(x,T-\tau)y)_x y\varphi \right) dxd\tau$$

$$= \int\int_{\mathcal{B}_t^*} a(x,T-\tau)\varphi y^2 dxd\tau \; - \; \frac{1}{2}\int\int_{\mathcal{B}_t^*} \varphi_x (y^2)_x dxd\tau$$

$$- \int\int_{\mathcal{B}_t^*} \varphi y_x^2 dxd\tau \; + \; \frac{1}{2}\int\int_{\mathcal{B}_t^*} (\varphi_\tau y^2 + 2b(x,T-\tau)\varphi_x y^2 + b(x,T-\tau)\varphi(y^2)_x) dxd\tau$$

$$\leq \; L\int\int_{\mathcal{B}_t^*} \varphi y^2 dxd\tau \; - \; \frac{1}{2}\int\int_{\mathcal{B}_t^*} \varphi_x (y^2)_x dxd\tau$$

$$+ \frac{1}{2}\int\int_{\mathcal{B}_t^*} \{\varphi_\tau + 2b(x,T-\tau)\varphi_x - (b(x,T-\tau)\varphi)_x\}\, y^2 dxd\tau. \tag{3.3}$$

From (3.2a) we derive, using Green's formula, that

$$-\int\int_{\mathcal{B}_t^*} \varphi_x (y^2)_x dxd\tau \; \leq \; \int\int_{\mathcal{B}_t^*} \varphi_{xx} y^2 dxd\tau + \int_0^t \mid \varphi_x(s_2(T-\tau),\tau) \mid y^2(s_2(T-\tau),\tau)d\tau. \tag{3.4}$$

Combining further (3.4) with (3.3), we obtain

$$\int_0^1 \varphi(x,t)y^2(x,t)dx \; \leq \; 2L\int_0^t\int_0^1 \varphi y^2 dxd\tau + \max_{(x,t)\in\bar{s}(T-\cdot)} \mid \varphi_x \mid \int_0^T y^2(s_2(T-\tau),\tau)d\tau$$

$$+ \int\int_{\mathcal{B}_t^*} (\varphi_\tau + 2b(x,T-\tau)\varphi_x - (b(x,T-\tau)\varphi)_x + \varphi_{xx}) y^2 dxd\tau.$$

In turn, recalling (3.2a) and (1.7) yields

$$\int_0^1 \varphi(x,t)y^2(x,t)dx$$

$$\leq \; 2L\int_0^t\int_0^1 \varphi y^2 dxd\tau \; + \; \max_{(x,t)\in\bar{\omega}(\cdot)} \mid F_x \mid \int_0^T y^2(s_2(T-\tau),\tau)d\tau. \tag{3.5}$$

Making use of Bellman–Gronwall's lemma, we can obtain from (3.5) that

$$\int_0^1 F(x,T-t)y^2(x,t)dx \; \leq \; e^{2LT} \max_{(x,t)\in\bar{s}(\cdot)} \mid F_x \mid \int_0^T y^2(s_2(T-\tau),\tau)d\tau \quad \forall t \in (0,T).$$

Hence, by (3.2b), there is an open subset D_* in Q_T (say, near the line $\{(x,t) \mid t = T\}$) for which

$$\int_{D_*} y^2(x,t)dxdt \; \leq \; Me^{2\|a\|_{C(\bar{Q}_T)}T} \int_0^T y^2(s_2(T-t),t)dt, \tag{3.6}$$

where M is a positive constant, which does not depend on $a(x,t)$. (This circumstance is critical for the fixed point argument in section 5.)

Remark 3.1. In the general case (i.e., when $s_1(\cdot)$ is present) in the above proof we will have just one more term similar to that in the above containing $s_2(T-\cdot)$ for the other branch of the curve $s(\cdot)$ defined in Assumption 1.3.

4. Auxiliary observability estimate for locally distributed support. It was shown in [14], [13, p. 24] that system (2.1) is exactly null-controllable from any $p_0 \in H_0^1(0,1)$ at any positive time T by using a locally distributed control whose magnitude is bounded as follows:

$$\| v \|_{L^2(\omega\times(0,T))} \le C \| p_0 \|_{H_0^1(0,1)}, \tag{4.1}$$

where C depends on $\omega \times (0,T)$.

The same result is also true for $p_0 \in L^2(0,1)$ and the locally distributed controls supported on any open subset of Q_T. Indeed, it is sufficient to show this for a control support like $\Upsilon = (l_1,l_2)\times(t_1,t_2)$, where $(l_1,l_2) \subset (0,1)$, $0 < t_1 < t_2 < T$ (i.e., with control $v = v(x,t)$ vanishing on $Q_T \backslash \Upsilon$).

Indeed, by the regularity of solutions to (2.1) with $p_0 \in L^2(0,1)$ and $v = 0$ on $(0,t_1)$, we have (similar to (2.8a) and (2.8b)) that

$$\| p(\cdot,t_1) \|_{H_0^1(0,1)} \le C \| p_0 \|_{L^2(0,1)} . \tag{4.2}$$

Applying the above exact null-controllability result on the time-interval (t_1,t_2) yields that there is a control $v = v(x,t)$ with support on Υ such that $p(\cdot,t_2) = 0$, while (4.1) holds with (t_1,t_2) in place of $(0,T)$ and $p(\cdot,t_1)$ in place of p_0. Hence, with $v = 0$ on (t_2,T), $p(\cdot,T) = 0$ also. Combining this with (4.2), we derive the estimate

$$\| v \|_{L^2(\Upsilon)} \le C(\Upsilon) \| p_0 \|_{L^2(0,1)}, \tag{4.3}$$

in which Υ can thus be any open subset of Q_T.

By the classical duality argument (4.3) implies the following dual observability estimate for (2.5):

$$\| y(\cdot,T) \|_{L^2(0,1)} \le C(D) \| y \|_{L^2(D)}, \tag{4.4}$$

where, again, $D = \{(x,t) \mid (x,T-t) \in \Upsilon\}$ can be any open subset of Q_T.

If one selects $D = D_*$ as in (3.6), then combining (4.4) and (3.6) provides the conclusion of Theorem 2.2.

Remark 4.1. Estimates (4.3) and (4.4) are immediate "adjustments" of the corresponding results in the above-cited works [14] and [13, p. 24]. The reader can find much more refined estimates of this type in [10]. (They were applied to the semilinear case in [11].)

5. The semilinear case.

Proof of Theorem 1.3. Let us give first the explicit form of system (S), (1.6) in Theorem 1.3, which is as follows:

$$u_t = u_{xx} + b(x,t)u_x + a(x,t)u + f(u) + v_1(t)\delta(x-s_1(t)) + v_2(t)\delta(x-s_2(t)) \quad \text{in } Q_T, \tag{5.1}$$

$$u(0,t) = u(1,t) = 0 \quad \text{in } (0,T), \quad u\,|_{t=0} = u_0 \in L^2(0,1), \;\; v_1, v_2 \in L^2(0,T).$$

Now note that the proof of Theorem 1.3 follows, in fact, from Theorems 2.2 and 1.2 by the fixed point argument (see, e.g., [6], [14], [13]). Its idea is to seek a suitable solution u to (5.1) satisfying (1.5) as a special ("fixed point") solution to the linear system like (2.2) with the potential $a(x,t) + f(z(x,t))/z(x,t)$ in place of $a(x,t)$ when z runs over $L^2(Q_T)$, namely, for which $z = u$. In its abstract operator form, this argument, based

on the estimate (1.9a)–(1.9b) and Theorems 2.2 and 1.2, is principally the same as for the locally distributed controls [18]. Therefore, we omit it here. □

Remark 5.1. Theorem 1.3 can easily be extended to the case of $f = f(x,t,u)$ for which the properties described in it hold uniformly in x and t.

COROLLARY 5.1. *The result of Theorem* 1.3 *holds with respect to approximate controllability, provided* $t_2 = T$ *in Assumption* 1.3.

This is an immediate traditional "structural" consequence of the exact null-controllability; see, e.g., [13, pp. 35–38]. Indeed, fix any $\varepsilon > 0$, and select any two initial and target states u_0 and u_1 in $L^2(0,1)$ for (5.1). Then, by the continuity of solutions to (5.1) in time (see, e.g., Theorem A.1), we can find T_* close enough to T and u_* close enough to u_1 such that the solution to (5.1) on (T_*, T), which we denote by $\bar{u}$ with the "initial" state $\bar{u}(\cdot, T_*) = u_*$ and $v_1 = v_2 = 0$, satisfies

$$\| \bar{u}(\cdot, T) - u_1 \|_{L^2(0,1)} \leq \varepsilon. \tag{5.2}$$

We build the control required to obtain (1.4) as follows.

On $(0, T_*)$ we apply the zero controls, while on (T_*, T) we employ controls v_1 and v_2, which solve the following auxiliary exact null-controllability problem:

$$\hat{u}_t = \hat{u}_{xx} + b(x,t)\hat{u}_x + a(x,t)\hat{u} + f(\hat{u} + \bar{u}) - f(\bar{u}) + v_1(t)\delta(x - s_1(t))$$

$$+ v_2(t)\delta(x - s_2(t)) \quad \text{in} \quad (0,1) \times (T_*, T),$$

$$\hat{u}(0,t) = \hat{u}(1,t) = 0 \quad \text{in} \ (T_*, T), \quad \hat{u} \mid_{t=T_*} = u(\cdot, T_*) - u_*,$$

$$\hat{u}(\cdot, T) = 0 \quad \text{in} \ (0,1). \tag{5.3}$$

This is possible due to Theorem 1.3 applied on (T_*, T) (recall we assumed that $s_1(\cdot)$ and $s_2(\cdot)$ "meet" at $t_2 = T$) for the function $f(x,t,s) = f(s + \bar{u}(x,t)) - f(\bar{u}(x,t))$ satisfying the assumptions of Theorem 1.3 along with Remark 5.1. Then on (T_*, T) we have

$$u = \hat{u} + \bar{u},$$

and hence, since $u(\cdot, T) = \bar{u}(\cdot, T)$, (5.3) and (5.2) imply (1.4).

6. Examples.

Example 6.1. Let $\Omega = (0,1)$ and the equation in (2.2) has the form

$$u_t = u_{xx} + bu_x + v(t)\delta(x - \alpha(T - t)) \quad \text{in} \quad Q_T, \quad v \in L^2(0,T). \tag{6.1}$$

We assume that $\alpha > 0$ and the velocity of convection b is positive and constant. Thus (6.1) is the case of the point control, which moves across Ω from the point $x = \min\{1, \alpha T\}$ to the left end-point $x = 0$ of the space domain Ω at the constant speed α; see Figure 1.1.

Put

$$F(x,t) = -x + \alpha(T - t),$$

$$\mathcal{A}^* = \{(x,t) \in \bar{Q}_T \mid x \leq \alpha(T - t)\}.$$

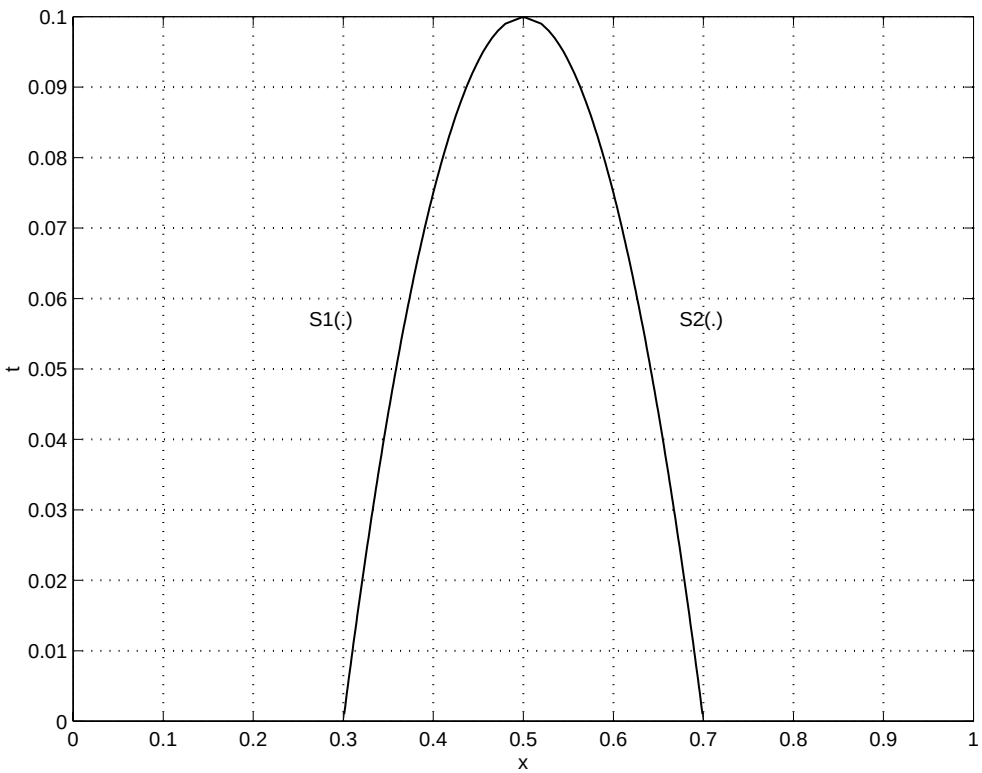


FIG. 6.1.

Take any $0 < T < 1/\alpha$. In turn, condition (1.7) gives

$$\alpha \ \leq \ b.$$

Hence, to make Theorem 1.2 work along Remark 2.1 with a *single point control*, we need the latter to move in the direction of convection with a speed which does not exceed that of convection. Note that our point control moves within the interval $[0, \alpha T]$.

Example 6.2. In Example 6.1, one can consider a parabolic trajectory $F(x,t) = -(x-0.5)^2 + \alpha(T-t) \ = \ 0$, whose two branches on the left and on the right of the line $x = 0.5$ form the control curves for two point controls:

$$s_1(t) = 0.5 - \sqrt{\alpha(T-t)}, \ \ s_2(t) = 0.5 + \sqrt{\alpha(T-t)}, \ \ t \in (0,T);$$

see Figure 6.1.

To satisfy Theorem 1.2, one may select, e.g., $\alpha \in (0,2]$, $T < 0.25/\alpha$, while setting $b = 0$, i.e., omitting convection. Note that our point controls move within the interval $[0.5 - \sqrt{\alpha T}, 0.5 + \sqrt{\alpha T}]$. (For instance, for $\alpha = 0.4, T = 0.1$ this interval will be $[0.3, 0.7]$.)

Example 6.3. The parabolic trajectories as in Example 6.2 will satisfy any a and b described in (S). Indeed, in this case (1.8) holds as before, and (1.7) is as follows:

$$\alpha T \ + \ b(x,t)(-2(x-0.5)) \ - \ b_x(x,t)(-(x-0.5)^2 + \alpha(T-t)) \ - 2 \ \leq \ 0.$$

The estimate will hold if, e.g., $t_2 = T$, while αT and $t_2 - t_1$ are sufficiently small positive numbers.

Examples 6.1–6.3 admit immediate extension to the semilinear case like in Theorem 1.3.

7. Concluding remarks. The above results can be extended in a number of ways.

7.1. More general coefficients. In (S) we may introduce a varying coefficient for the leading term; that is, $q(x,t)u_{xx}, q(x,t) > 0$ instead of u_{xx}. In particular, along the scheme of section 3 dealing with the integration by parts, this will change the condition (1.7) to the following:

$$-F_t(x,t) \ + \ 2b(x,t)F_x(x,t) - \{b(x,t)F(x,t)\}_x \ + \ (q(x,t)F)_{xx} \ \le 0 \quad \forall (x,t) \in \mathcal{A}^*.$$

7.2. Alternative $L^\infty(\varepsilon, T)$-estimates. Note that if $[t_1 = 0, t_2] \subset [0, T)$, then by (2.9) the trace of y on the curve $s(\cdot)$ in Assumption 1.3 is a continuous function. Hence (2.15), derived for two point sensors, implies

$$\| \, y(\cdot, T) \, \|_{L^2(0,1)} \le \ C(\varepsilon) \, \| \, y(s(T - \cdot), \cdot) \, \|_{L^\infty(\varepsilon, T)}, \tag{7.1}$$

where $\varepsilon = T - t_2$.

The same type of estimates was the subject of the works [15] and [17], where the path $s(\cdot)$ for a *single* point sensor was selected within Q_T following a certain algorithmic optimization procedure with infinitely many steps. Unlike the results of this paper, the method of [15] and [17] did not provide the explicit description of $s(\cdot)$.

The idea of [15] and [17] was to select (a) a countable set of solutions $\{y_1, \ldots\}$ to (2.5) which is dense (in suitable sense) in the set of all possible solutions to (2.5) and (b) an arbitrary sequence of moments $0 < t_1 < t_2 \ldots$ in (ε, T), and then (c) to associate them with a sequence of points $x_k \in (0,1)$ such that

$$\| \, y_k(\cdot, T) \, \|_{L^2(0,1)} \le \ C \mid y_k(x_k, t_k) \mid$$

for the same $C > 0$ for all $k = 1, \ldots$. The latter was achieved making use of the maximum principle in [15] and its generalized version in [17]. Then to have (7.1), $s(\cdot)$ is to be selected as any continuous curve passing through all the points $(t_k, x_k), k = 1, \ldots$.

The estimate (7.1) yields the observability in the sense of (2.7) of system (2.5) for the corresponding $s(\cdot)$ and hence, as it was described in section 2, the approximate controllability of (2.2) by means of $L^2(0,T)$-controls. Moreover, in the dual fashion leading to Theorem 1.2, the estimate (7.1) also provides a similar exact null-controllability result *but* with controls from the space dual of $L^\infty(\varepsilon, T)$.

It seems plausible that the results of [15] and [17] can be extended to narrow the (implicit) selection of the trajectories $s(\cdot)$ to those lying within any open set $\omega \times (\varepsilon, T)$, based on the technique of section 4 in the above, making use of the estimate (4.4) instead of the maximum principle as in these works. However, one should expect a highly irregular ("impractical") behavior of such curves when following this strategy.

7.3. The case of several dimensions. The just-mentioned "linear" results of [15] and [17] for single point sensors actually hold in the several space dimensions as well, in which case solutions to the dual linear system like (2.2) with $(L^\infty(0,T))'$-controls can be defined by the duality technique. It is *not* so in the semilinear case like (5.1).

It should also be noted that in several space dimensions the well-posedness of the point observations requires more regularity than $L^2(\Omega)$ for the initial data. Hence the dual controllability of the corresponding multidimensional version of (2.2) can be achieved only in the spaces that are weaker than $L^2(\Omega)$. This explains, in particular, why in the above we focused on the one-dimensional case.

On the other hand, the results of this paper for the controllability in $L^2(\Omega)$ can easily be extended to the case of several dimensions in the sense that in n space dimensions we should use controls that are supported on the surfaces of dimension $[n-1]$. In this way in [15] we established the exact null-controllability of the semilinear reaction-diffusion-convection equation with the superlinear reaction term like $f(u) = -\mid u \mid u^r, r > 0$ and the additive linear convection term, assuming that the control supporting surface (a) satisfies a differential inequality like (1.7) *and* (b) separates the top of the cylinder Q_T from its bottom.

7.4. Superlinear growth. In the series of recent works [16], [8], [20], [21], [9], [2], [1], [11], [22] the global exact null- and/or approximate controllability of a multidimensional system like (S) were shown by means of locally distributed controls for various types of superlinear growth of nonlinear terms.

- *Superlinear logarithmic growth.* In [8] the exact null-controllability property was shown in $L^2(\Omega)$ (or appropriate Sobolev space) assuming that f can grow superlinearly at the rate $\lim_{|p|\to\infty} f(p)/(p \log \mid p \mid) = 0$. This result was improved in [2] to the rate $\lim_{|p|\to\infty} f(p)/(p(\log \mid p \mid)^{3/2}) = 0$ under the additional dissipativity condition that $-f(u)u \geq -cu^2$, where $c > 0$. In [9], [11] it was shown that the latter condition can be avoided, both in terms of exact null- and approximate controllability. Interestingly, [9], [11] deal with the system which admits blow-up.
- *Polynomial growth.* It is well known that if f is dissipative and admits the polynomial growth at infinity, then, in general, it is *not* globally controllable in any of $L^p(\Omega)$-type spaces; see [13], [6], and the references therein. Nonetheless, it was shown in [20] that this property indeed holds in some spaces that are weaker than any of the above-mentioned L^p-spaces. Alternatively, the global approximate controllability in $L^2(\Omega)$ can be achieved for such systems by using an additional bilinear lumped control [22] (see also Remark 1.1). The result of [21] extends that of [19] for the superlinear time-dependent nonlinearities with "fast" convergence to zero as $t \to 0$ to the n-dimensional case with the locally distributed controls in place of the lumped ones as in (1.3) in the one-dimensional case; see Remark 1.1.
- *Polynomial growth: Finite dimensional controllability.* The local aspects at an equilibrium of this property for dissipative nonlinearities were analyzed in [29]. For the global aspects we refer to [20].

The extension of our Theorem 1.3 to the above-listed superlinear terms is an open question. On the one hand, since in this article we reduce the issue of controllability with the point controls to that with the locally distributed ones (namely, (2.15) follows from (3.6) and (4.4)), this extension does not seem impossible, at least in some situations. On the other hand, there are many serious difficulties, both technical and conceptual, in this direction. Let us mention just two of them here.

- The existence of a solution to (S) with the point control (1.2) seems to be a highly technical issue and can intrinsically be ill-posed (involving potential blow-up) when the dissipativity condition is not assumed (see, [8], [9], [11]).
- In the superlinear case the uniqueness of solutions to the boundary problem

at hand typically is not guaranteed. This means that Definitions 1.1 and 1.2 are ill-posed, which requires their certain generalization. In the works [8], [9], [11], and [2] these properties were established in the sense that there is at least one solution to the PDE at hand for which either (1.4) or (1.5) holds. In [16], [20], [21], and [22] controllability was achieved by selecting a control that can uniformly steer all possible multiple solutions.

Appendix A. In this section we prove Theorem A.1 formulated at the end of subsection 1.2. For simplicity of notations our proof deals with the system (S) or (5.1), which is the same, and one point control as in (1.2).

Proof of Theorem A.1. The uniqueness follows by the standard technique.

Our proof of existence is based on that of the corresponding existence result in [23, pp. 467–474], established there for (5.1) in the absence of the control term $v(t)\delta(x-s(t))$. The degenerate nature of the latter is the crux here.

For simplicity we further omit the coefficient a, assuming that it is incorporated into the globally Lipschitz f.

First of all, let us recall that we deal here with generalized solutions, understood in the sense of the following identity, obtained by formal integration by parts of (5.1) multiplied by an arbitrary smooth test function ψ, vanishing at $x = 0, 1$:

$$\int_0^t \int_\Omega (-u\psi_t - u_x\psi_x dxdt \ - \ bu_x\psi \ + \ f(u)\psi)dxdt$$

$$= \int_\Omega (\psi(x,0)u_0 - \psi(x,t)u(x,t))dx \ + \ \int_0^T v\psi(s(\tau))d\tau \quad \forall t \in [0,T]. \tag{A.1}$$

Here, by the Lipschitz property of f and the embedding theorems (e.g., [23]) $f(u) \in L^6(Q_T)$ for all $u \in \Xi$ and

$$\| \psi(s(\cdot),\cdot) \|_{L^2(0,T)} \leq \ C \| \psi \|_{L^2(0,T;H_0^1(0,1))} \tag{A.2}$$

for some positive constant C.

Following [23], we apply Galerkin's method. Namely, we look for an approximate solution u^N in the form

$$u^N(x,t) = \sum_{k=1}^N c_k^N(t)\psi_k(x),$$

where $\{\psi_k\}$ is a fundamental system in $H_0^1(Q_T)$, i.e., ψ_k's are linear independent and span the entire $H_0^1(Q_T)$, and

$$\int_\Omega \psi_k\psi_l dx \ = \ \delta_k^l \quad \text{and} \quad \| \psi_k, \ \psi_{kx} \|_{L^\infty(Q_T)} \leq \ c_k \quad k = 1, \dots .$$

The substitution of u^N into (5.1) provides the following system of ordinary differential equations in c_k^N's:

$$\frac{\partial c_k}{\partial t} = \int_\Omega \left(u_{x_i}^N\psi_{kx} \ + \ b(x,t)u_x^N\psi_k \ - \ f(u^N)\psi_k\right) dx \ + \ v(t)\psi_k(s(t),t), \quad k = 1, \dots, N, \tag{A.3}$$

with the initial conditions

$$c_k^N(0) = \int_\Omega u_0(x)\psi_k(x,0)dx, \quad k = 1, \dots, N.$$

Multiplication of the kth equation in (A.3) by $c_k^N(t)$ and summation of all the equations and integration over $(0,T)$ yield that

$$\frac{1}{2} \parallel u^N(x,t) \parallel^2_{L^2(\Omega)} \le \frac{1}{2} \parallel u^N(\cdot,0) \parallel^2_{L^2(\Omega)} - \int_0^t \int_\Omega (u_x^N)^2 dxd\tau$$

$$+ \max_{(x,t)\in\bar{Q}_T} \{\mid b(x,t) \mid\} \int_0^t \int_\Omega u_x^N u^N dxd\tau$$

$$+ L \int_0^t \int_\Omega (u^N)^2 dxdt + \parallel v \parallel_{L^2(0,T)} \parallel u^N(s(\cdot),\cdot) \parallel_{L^2(0,t)}, \tag{A.4}$$

where L is the Lipschitz constant for f.

Young's and Poincaré–Friedrichs's inequalities and (A.2) provide us with the following estimates, valid for any $s > 0$:

$$\parallel v \parallel_{L^2(0,T)} \parallel u^N(s(\cdot),\cdot) \parallel_{L^2(0,t)} \le \frac{1}{s} \parallel v \parallel^2_{L^2(0,T)} + s \parallel u^N(s(\cdot),\cdot) \parallel^2_{L^2(0,t)}$$

$$\le \frac{1}{s} \parallel v \parallel^2_{L^2(0,T)} + sC_* \int_0^t \int_\Omega (u_x^N)^2 dxd\tau \tag{A.5}$$

for some $C_* > 0$, and

$$\int_0^t \int_\Omega u_x^N u^N dxd\tau \le \frac{1}{s} \parallel u^N \parallel^2_{L^2(Q_t)} + s \int_0^t \int_\Omega (u_x^N)^2 dxd\tau. \tag{A.6}$$

Select $s > 0$ in (A.5) and (A.6) so that

$$1 - sC_* - s \max_{(x,t)\in\bar{Q}_T} \mid b(x,t) \mid > \frac{1}{2}.$$

Then we derive from (A.4)–(A.6) that

$$\max_{\tau\in[0,t]} \parallel u^N(\cdot,\tau) \parallel^2_{L^2(\Omega)} + \int_0^t \int_\Omega (u_x^N)^2 dxd\tau$$

$$\le 2 \Bigg[\parallel u^N(\cdot,0) \parallel^2_{L^2(\Omega)} + 2L \int_0^t \int_\Omega (u^N)^2 dxd\tau$$

$$+ 2 \max_{(x,t)\in\bar{Q}_T} \{b(x,t)\} \frac{1}{s} \int_0^t \int_\Omega (u^N)^2 dxd\tau + \frac{2}{s} \parallel v \parallel^2_{L^2(0,T)} \Bigg].$$

Since

$$\int_0^t \int_\Omega (u^N(x,\tau))^2 dx d\tau \;\le\; \int_0^t \left(\max_{s\in[0,\tau]} \| u^N(\cdot,s) \|^2_{L^2(\Omega)} \;+\; \int_0^\tau \int_\Omega (u^N_x)^2 dx ds \right) d\tau,$$

we further have

$$\max_{\tau\in[0,t]} \| u^N(\cdot,\tau) \|^2_{L^2(\Omega)} \;+\; \int_0^t \int_\Omega (u^N_x)^2 dx d\tau$$

$$\le\; 2 \| u^N(\cdot,0) \|^2_{L^2(\Omega)} \;+ \frac{4}{s} \| v \|^2_{L^2(0,T)}$$

$$+ c_0 \int_0^t \left(\max_{s\in[0,\tau]} \| u^N(\cdot,s) \|^2_{L^2(\Omega)} \;+\; \int_0^\tau \int_\Omega (u^N_x)^2 dx ds \right) d\tau$$

for $c_0 = 4(c_3 + \max_{(x,t)\in\bar{Q}_T}\{b(x,t)\}\frac{1}{s})$. Employing Bellman–Gronwall's lemma, we obtain

$$\| u^N \|_\Xi = \left(\max_{t\in[0,T]} \| u^N(x,t) \|^2_{L^2(\Omega)} \;+\; \int_0^t \int_\Omega (u^N_x)^2 dx dt \right)^{\frac{1}{2}}$$

$$\le\; c(T) \left(\| u^N(x,0) \|_{L^2(\Omega)} \;+\; \| v \|_{L^2(0,T)} \right) \tag{A.7}$$

for some constant $c(T)$, which does not depend on N and $a(x,t)$. This is exactly the estimate (6.47) in [23, p. 468] from which the rest of the proof of Theorem A.1 follows the lines of that of Theorem 6.7 in [23, Ch. V, pp. 466–474] (by appropriate limit passage). Note that (A.7) provides the second estimate in (1.9a) and (1.9b), while the first one follows by continuous embedding $\Xi \subset L^6(Q_T)$; see, e.g., [23, p. 466]. □

Acknowledgment. The author wishes to acknowledge the referee's very helpful suggestions on the improvement of the paper.

REFERENCES

[1] S. Anita and V. Barbu, *Null controllability of nonlinear convective heat equations*, ESAIM Control Optim. Calc. Var., 5 (2000), pp. 157–173.

[2] V. Barbu, *Exact controllability of the superlinear heat equation*, Appl. Math. Optim., 42 (2000), pp. 73–89.

[3] C. Bardos and L. Tartar, *Sur l'unicité rétrograde des équations paraboliques et quelques questions voisines*, Arch. Rational Mech. Anal., 50 (1973), pp. 10–25.

[4] Sz. Dolecki, *Observation for the one-dimensional heat equation*, Stadia Math., 48 (1973), pp. 291–305.

[5] Sz. Dolecki and D. L. Russell, *A general theory of observation and control*, SIAM J. Control Optim., 15 (1977), pp. 185–220.

[6] C. Fabre, J.-P. Puel, and E. Zuazua, *Approximate controllability for the semilinear heat equations*, Proc. Roy. Soc. Edinburgh Sect. A, 125 (1995), pp. 31–61.

[7] H. O. Fattorini and D. L. Russell, *Uniform bounds on biorthogonal functions for real exponentials with an application to the control theory of parabolic equations*, Quart. Appl. Math., 32 (1974/75), pp. 45–69.

[8] E. Fernández-Cara, *Null controllability of the semilinear heat equation*, ESAIM Control Optim. Calc. Var., 2 (1997), pp. 87–103.

[9] E. Fernández-Cara and E. Zuazua, *Controllability for blowing up semilinear parabolic equations*, C. R. Acad. Sci. Paris Sér. I Math., 330 (2000), pp. 199–204.

[10] E. Fernández-Cara and E. Zuazua, *The cost of approximate controllability for heat equations: The linear case*, Adv. Differential Equations, 5 (2000), pp. 465–514.
[11] E. Fernández-Cara and E. Zuazua, *Null and approximate controllability for weakly blowing-up semilinear heat equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, to appear.
[12] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.
[13] A. Fursikov and O. Imanuvilov, *Controllability of Evolution Equations*, Lect. Notes Ser. 34, Seoul National University, Seoul, Korea, 1996.
[14] O. Imanuvilov, *Controllability for parabolic equations*, Mat. Sb., 186 (1995), pp. 109–132.
[15] A. Khapalov, *L^∞-exact observability of the heat equation with scanning pointwise sensor*, SIAM J. Control Optim., 32 (1994), pp. 1037–1051.
[16] A. Y. Khapalov, *Some aspects of the asymptotic behavior of the solutions of the semilinear heat equation and approximate controllability*, J. Math. Anal. Appl., 194 (1995), pp. 858–882.
[17] A. Khapalov, *On unique continuation of the solutions of the parabolic equation from a curve*, Control Cybernet., 25 (1996), pp. 451–463.
[18] A. Y. Khapalov, *Exact Null-Controllability for the Semilinear Heat Equation with Mobile Controls of Degenerate Support*, Tech. report 98-2, Math. Dept., Washington State University, Pullman, WA, 1998.
[19] A. Khapalov, *Approximate controllability and its well-posedness for the semilinear reaction-diffusion equation with internal lumped controls*, ESAIM Control Optim. Calc. Var., 4 (1999), pp. 83–98.
[20] A. Y. Khapalov, *Global approximate controllability properties for the semilinear heat equation with superlinear term*, Rev. Mat. Complut., 12 (1999), pp. 511–535.
[21] A. Y. Khapalov, *A class of globally controllable semilinear heat equations with superlinear terms*, J. Math. Anal. Appl., 242 (2000), pp. 271–283.
[22] A. Y. Khapalov, *Bilinear control for global controllability of the semilinear parabolic equations with superlinear terms*, in Control of Nonlinear Distributed Parameter Systems, Dedicated to David Russell, Marcel Dekker, New York, 2001, pp. 139–155.
[23] O. H. Ladyzhenskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasi-Linear Equations of Parabolic Type*, AMS, Providence, RI, 1968.
[24] G. Lebeau and L. Robbiano, *Contrôle exact de l'equation de la chaleur*, Comm. Partial Differential Equations, 20 (1995), pp. 335–356.
[25] V. P. Mikhailov, *Partial Differential Equations*, Mir, Moscow, 1978.
[26] Y. Sakawa, *Observability and related problems for partial differential equations of parabolic type*, SIAM J. Control, 13 (1975), pp. 14–27.
[27] J.-C. Saut and B. Scheurer, *Unique continuation for some evolution equations*, J. Differential Equations, 66 (1987), pp. 118–139.
[28] H. X. Zhou, *A note on approximate controllability for semilinear one-dimensional heat equation*, Appl. Math. Optim., 8 (1982), pp. 275–285.
[29] E. Zuazua, *Finite dimensional null controllability for the semilinear heat equation*, J. Math. Pures Appl. (9), 76 (1997), pp. 237–264.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 1, pp. 253–269

LIE-ALGEBRAIC STABILITY CRITERIA FOR SWITCHED SYSTEMS*

ANDREI A. AGRACHEV† AND DANIEL LIBERZON‡

Abstract. It was recently shown that a family of exponentially stable linear systems whose matrices generate a solvable Lie algebra possesses a quadratic common Lyapunov function, which implies that the corresponding switched linear system is exponentially stable for arbitrary switching. In this paper we prove that the same properties hold under the weaker condition that the Lie algebra generated by given matrices can be decomposed into a sum of a solvable ideal and a subalgebra with a compact Lie group. The corresponding local stability result for nonlinear switched systems is also established. Moreover, we demonstrate that if a Lie algebra fails to satisfy the above condition, then it can be generated by a family of stable matrices such that the corresponding switched linear system is not stable. Relevant facts from the theory of Lie algebras are collected at the end of the paper for easy reference.

Key words. switched system, asymptotic stability, Lie algebra

AMS subject classifications. 93D20, 93B25, 93B12, 17B30

PII. S0363012999365704

1. Introduction.

A switched system can be described by a family of continuous-time subsystems and a rule that orchestrates the switching between them. Such systems arise, for example, when different controllers are being placed in the feedback loop with a given process, or when a given process exhibits a switching behavior caused by abrupt changes of the environment. For a discussion of various issues related to switched systems, see the recent survey article [13].

To define more precisely what we mean by a switched system, consider a family $\{f_p : p \in \mathcal{P}\}$ of sufficiently regular functions from $\mathbb{R}^n$ to $\mathbb{R}^n$, parameterized by some index set $\mathcal{P}$. Let $\sigma : [0, \infty) \to \mathcal{P}$ be a piecewise constant function of time, called a *switching signal*. A *switched system* is then given by the following system of differential equations in $\mathbb{R}^n$:

$$\dot{x} = f_\sigma(x). \tag{1}$$

We assume that the state of (1) does not jump at the switching instants, i.e., the solution $x(\cdot)$ is everywhere continuous. Note that infinitely fast switching (chattering), which calls for a concept of generalized solution, is not considered in this paper. In the particular case when all the individual subsystems are linear (i.e., $f_p(x) = A_p x$, where $A_p \in \mathbb{R}^{n \times n}$ for each $p \in \mathcal{P}$), we obtain a *switched linear system*

$$\dot{x} = A_\sigma x. \tag{2}$$

This paper is concerned with the following problem: find conditions on the individual subsystems which guarantee that the switched system is asymptotically stable

*Received by the editors December 14, 1999; accepted for publication (in revised form) January 22, 2001; published electronically June 26, 2001.
http://www.siam.org/journals/sicon/40-1/36570.html
†Steklov Math. Inst., Moscow, Russia, and S.I.S.S.A.–I.S.A.S., via Beirutz-4 Trieste, 34014 Italy (agrachev@sissa.it).
‡Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, 1308 West Main Street, Urbana, IL 61801 (liberzon@uiuc.edu). The research of this author was supported in part by ARO grant DAAH04-95-1-0114, by NSF grant ECS 9634146, and by AFOSR grant F49620-97-1-0108.

for an arbitrary switching signal σ. In fact, a somewhat stronger property is desirable, namely, asymptotic or even exponential stability that is uniform over the set of all switching signals. Clearly, all the individual subsystems must be asymptotically stable, and we will assume this to be the case throughout the paper. Note that it is not hard to construct examples where instability can be achieved by switching between asymptotically stable systems (section 4 contains one such example), so one needs to determine what additional requirements must be imposed. This question has recently generated considerable interest, as can be seen from the work reported in [9, 12, 16, 17, 18, 19, 21, 22].

Commutation relations among the individual subsystems play an important role in the context of the problem posed above. This can be illustrated with the help of the following example. Consider the switched linear system (2), take $\mathcal{P}$ to be a finite set, and suppose that the matrices A_p commute pairwise: $A_pA_q = A_qA_p$ for all $p, q \in \mathcal{P}$. Then it is easy to show directly that the switched linear system is exponentially stable, uniformly over all switching signals. Alternatively, one can construct a quadratic common Lyapunov function for the family of linear systems

$$\dot{x} = A_p x, \qquad p \in \mathcal{P}, \tag{3}$$

as shown in [18], which is well known to lead to the same conclusion.

In this paper we undertake a systematic study of the connection between the behavior of the switched system and the commutation relations among the individual subsystems. In the case of the switched linear system (2), a useful object that reveals the nature of these commutation relations is the Lie algebra $\mathfrak{g} := \{A_p : p \in \mathcal{P}\}_{LA}$ generated by the matrices A_p, $p \in \mathcal{P}$ (with respect to the standard Lie bracket $[A_p, A_q] := A_pA_q - A_qA_p$). The observation that the structure of this Lie algebra is relevant to stability of (2) goes back to the paper by Gurvits [9]. That paper studied the discrete-time counterpart of (2) taking the form

$$x(k+1) = A_{\sigma(k)}x(k), \tag{4}$$

where σ is a function from nonnegative integers to a finite index set $\mathcal{P}$ and $A_p = e^{L_p}$, $p \in \mathcal{P}$, for some matrices L_p. Gurvits conjectured that if the Lie algebra $\{L_p : p \in \mathcal{P}\}_{LA}$ is nilpotent (which means that Lie brackets of sufficiently high order equal zero), then the system (4) is asymptotically stable for any switching signal σ. He was able to prove this conjecture for the particular case when $\mathcal{P} = \{1, 2\}$ and the third-order Lie brackets vanish: $[L_1, [L_1, L_2]] = [L_2, [L_1, L_2]] = 0$.

It was recently shown in [12] that the switched linear system (2) is exponentially stable for arbitrary switching if the Lie algebra $\mathfrak{g}$ is solvable (see section A.3 for the definition). The proof relied on the facts that matrices in a solvable Lie algebra can be simultaneously put in the upper-triangular form (Lie's theorem) and that a family of linear systems with stable upper-triangular matrices has a quadratic common Lyapunov function. For the result to hold, the index set $\mathcal{P}$ does not need to be finite (although a suitable compactness assumption is required). One can derive the corresponding result for discrete-time systems in similar fashion, thereby confirming and directly generalizing the statement conjectured by Gurvits (because every nilpotent Lie algebra is solvable).

In the present paper we continue the line of work initiated in the above references. Our main theorem is a direct extension of the one proved in [12]. The new result states that one still has exponential stability for arbitrary switching if the Lie algebra $\mathfrak{g}$ is a semidirect sum of a solvable ideal and a subalgebra with a compact Lie group

(which amounts to saying that all the matrices in this second subalgebra have purely imaginary eigenvalues). The corresponding local stability result for the nonlinear switched system (1) is also established. Being formulated in terms of the original data, such Lie-algebraic stability criteria have an important advantage over results that depend on a particular choice of coordinates, such as the one reported in [16]. Moreover, we demonstrate that the above condition is in some sense the strongest one that can be given on the Lie algebra level. Loosely speaking, we show that if a Lie algebra does not satisfy this condition, then it could be generated by a switched linear system that is not stable.

More precisely, the main contributions of the paper can be summarized as follows. (See the appendix for an overview of relevant definitions and facts from the theory of Lie algebras.) Given a matrix Lie algebra $\hat{\mathfrak{g}}$ which contains the identity matrix, we are interested in the following question. Is it true that any set of stable generators for $\hat{\mathfrak{g}}$ gives rise to a switched system that is exponentially stable, uniformly over all switching signals? We discover that this property depends only on the structure of $\hat{\mathfrak{g}}$ as a Lie algebra and not on the choice of a particular matrix representation of $\hat{\mathfrak{g}}$. The following equivalent characterizations of the above property can be given.

1. The factor algebra $\hat{\mathfrak{g}}$ mod $\mathfrak{r}$, where $\mathfrak{r}$ denotes the radical, is a compact Lie algebra.
2. The Killing form is negative semidefinite on $[\hat{\mathfrak{g}}, \hat{\mathfrak{g}}]$.
3. The Lie algebra $\hat{\mathfrak{g}}$ does not contain any subalgebras isomorphic to $sl(2, \mathbb{R})$.

We will also show how the investigation of stability (in the above sense) of a switched linear system in $\mathbb{R}^n$, $n > 2$, whose associated Lie algebra is low-dimensional, can be reduced to the investigation of stability of a switched linear system in $\mathbb{R}^2$. For example, take $\mathcal{P} = \{1, 2\}$, and define $\tilde{A}_i := A_i - \frac{1}{n}\text{trace}(A_i)I$, $i = 1, 2$. Assume that all iterated Lie brackets of the matrices $\tilde{A}_1$ and $\tilde{A}_2$ are linear combinations of $\tilde{A}_1$, $\tilde{A}_2$, and $[\tilde{A}_1, \tilde{A}_2]$. This means that if we consider the Lie algebra $\mathfrak{g} = \{A_1, A_2\}_{LA}$ and add to it the identity matrix (if it is not already there), the resulting Lie algebra $\hat{\mathfrak{g}}$ has dimension at most 4. In this case, the following algorithm can be used to verify that the switched linear system generated by A_1 and A_2 is uniformly exponentially stable or, if this is not possible, to construct a second-order switched linear system whose uniform exponential stability is equivalent to that of the original one.

Step 1. If $[\tilde{A}_1, \tilde{A}_2]$ is a linear combination of $\tilde{A}_1$ and $\tilde{A}_2$, stop: the system is stable. Otherwise, write down the matrix of the Killing form for the Lie algebra $\tilde{\mathfrak{g}} := \{\tilde{A}_1, \tilde{A}_2\}_{LA}$ relative to the basis given by $\tilde{A}_1$, $\tilde{A}_2$, and $[\tilde{A}_1, \tilde{A}_2]$. (This is a symmetric 3×3 matrix; see section A.4 for the definition of the Killing form.)

Step 2. If this matrix is degenerate or negative definite, stop: the system is stable. Otherwise, continue.

Step 3. Find three matrices h, e, and f in $\tilde{\mathfrak{g}}$ with commutation relations $[h, e] = 2e$, $[h, f] = -2f$, and $[e, f] = h$ (this is always possible in the present case). We can then write $\tilde{A}_i = \beta_i e + \gamma_i f + \delta_i h$, where $\alpha_i, \beta_i, \gamma_i$ are constants, $i = 1, 2$.

Step 4. Compute the largest eigenvalue of h. It will be an integer; denote it by k. Then the given system is stable if and only if the switched linear system generated by the 2×2 matrices

$$\hat{A}_1 := \begin{pmatrix} \frac{\text{trace}(A_1)}{nk} - \delta_1 & -\beta_1 \\ -\gamma_1 & \frac{\text{trace}(A_1)}{nk} + \delta_1 \end{pmatrix}, \quad \hat{A}_2 := \begin{pmatrix} \frac{\text{trace}(A_2)}{nk} - \delta_2 & -\beta_2 \\ -\gamma_2 & \frac{\text{trace}(A_2)}{nk} + \delta_2 \end{pmatrix}$$

is stable.

All the steps in the above reduction procedure involve only elementary matrix operations (addition, multiplication, and computation of eigenvalues and eigenvectors). Details and justification are given in section 4.

Before closing the introduction, we make one more remark to further motivate the work reported here and point out its relationship to a more classical branch of control theory. Assume that $\mathcal{P}$ is a finite set, say, $\mathcal{P} = \{1, \dots, m\}$. The switched system (1) can then be recast as

$$\dot{x} = \sum_{i=1}^{m} f_i(x) u_i, \tag{5}$$

where the admissible controls are of the form $u_k = 1$, $u_i = 0 \; \forall i \neq k$. (This corresponds to $\sigma = k$.) In particular, the switched linear system (2) gives rise to the bilinear system

$$\dot{x} = \sum_{i=1}^{m} A_i x u_i.$$

It is intuitively clear that asymptotic stability of (1) for arbitrary switching corresponds to lack of controllability for (5). Indeed, it means that for any admissible control function the resulting solution trajectory must approach the origin. Lie-algebraic techniques have received a lot of attention in the context of the controllability problem for systems of the form (5). As for the literature on stability analysis of switched systems, despite the fact that it is vast and growing, Lie-algebraic methods do not yet seem to have penetrated it. The present work can be considered as a step towards filling this gap.

The rest of the paper is organized as follows. In section 2 we establish a sufficient condition for stability (Theorem 2) and discuss its various implications. In section 3 we prove a converse result (Theorem 4). Section 4 contains a detailed analysis of switched systems whose associated Lie algebras are isomorphic to the Lie algebra $gl(2, \mathbb{R})$ of real 2×2 matrices. This leads to, among other things, the reduction algorithm sketched above and to a different (and arguably more illuminating) proof of Theorem 4. To make the paper self-contained, in the appendix we provide an overview of relevant facts from the theory of Lie algebras.

2. Sufficient conditions for stability. The switched system (1) is called (locally) *uniformly exponentially stable* if there exist positive constants M, c, and μ such that for any switching signal σ the solution of (1) with $\|x(0)\| \leq M$ satisfies

$$\|x(t)\| \leq c e^{-\mu t} \|x(0)\| \quad \forall t \geq 0. \tag{6}$$

The term "uniform" is used here to describe uniformity with respect to switching signals. If there exist positive constants c and μ such that the estimate (6) holds for any switching signal σ and any initial condition $x(0)$, then the switched system is called *globally uniformly exponentially stable*. Similarly, one can also define the property of uniform *asymptotic* stability, local or global. For switched linear systems all the above concepts are equivalent (see [15]). In fact, as shown in [1], in the linear case global uniform exponential stability is equivalent to the seemingly weaker property of asymptotic stability for any switching signal.

In the context of the switched linear system (2), we will always assume that $\{A_p : p \in \mathcal{P}\}$ is a *compact* (with respect to the usual topology in $\mathbb{R}^{n \times n}$) set of real $n \times n$ matrices with eigenvalues in the open left half-plane. Let $\mathfrak{g}$ be the Lie algebra

defined by $\mathfrak{g} = \{A_p : p \in \mathcal{P}\}_{LA}$ as before. The following stability criterion was established in [12]. It will be crucial in proving Theorem 2 below.

THEOREM 1 (see [12]). *If $\mathfrak{g}$ is a solvable Lie algebra, then the switched linear system* (2) *is globally uniformly exponentially stable.*

Remark 1. The proof of this result given in [12] relies on a construction of a quadratic common Lyapunov function for the family of linear systems (3). The existence of such a function actually implies global uniform exponential stability of the time-varying system $\dot{x} = A_\sigma x$ with σ not necessarily piecewise constant. This observation will be used in the proof of Theorem 2.

The above condition can always be checked directly in a finite number of steps if $\mathcal{P}$ is a finite set. Alternatively, one can use the standard criterion for solvability in terms of the Killing form. Similar criteria exist for checking the other conditions to be presented in this paper—see sections A.3 and A.4 for details.

We now consider a Levi decomposition of $\mathfrak{g}$, i.e., we write $\mathfrak{g} = \mathfrak{r} \oplus \mathfrak{s}$, where $\mathfrak{r}$ is the radical and $\mathfrak{s}$ is a semisimple subalgebra (see section A.4). Our first result is the following generalization of Theorem 1.

THEOREM 2. *If $\mathfrak{s}$ is a compact Lie algebra, then the switched linear system* (2) *is globally uniformly exponentially stable.*

Proof. For an arbitrary $p \in \mathcal{P}$, write $A_p = r_p + s_p$ with $r_p \in \mathfrak{r}$ and $s_p \in \mathfrak{s}$. Let us show that r_p is a stable matrix. Writing

$$e^{(r_p+s_p)t} = e^{s_p t} B_p(t), \tag{7}$$

we have the following equation for $B_p(t)$:

$$\dot{B}_p(t) = e^{-s_p t} r_p e^{s_p t} B_p(t), \qquad B_p(0) = I. \tag{8}$$

To verify (8), differentiate the equality (7) with respect to t, which gives

$$(r_p + s_p) e^{(r_p+s_p)t} = s_p e^{s_p t} B_p + e^{s_p t} \dot{B}_p.$$

Using (7) again, we have

$$r_p e^{s_p t} B_p + s_p e^{s_p t} B_p = s_p e^{s_p t} B_p + e^{s_p t} \dot{B}_p;$$

hence (8) holds. Define $c_p(t) := e^{-s_p t} r_p e^{s_p t}$. Clearly, $\operatorname{spec}(c_p(t)) = \operatorname{spec}(r_p)$ for all t. It is well known that for any two matrices A and B one has

$$e^{-A} B e^{A} = e^{\operatorname{ad} A}(B) = B + [A, B] + \frac{1}{2}[A, [A, B]] + \cdots; \tag{9}$$

hence we obtain the expansion

$$c_p(t) = r_p + [s_p t, r_p] + \frac{1}{2}[s_p t, [s_p t, r_p]] + \cdots.$$

Since $[\mathfrak{s}, \mathfrak{r}] \subseteq \mathfrak{r}$, we see that $c_p(t) \in \mathfrak{r}$. According to Lie's theorem, there exists a basis in which all matrices from $\mathfrak{r}$ are upper-triangular. Combining the above facts, it is not hard to check that $\operatorname{spec}(B_p(t)) = e^{t \operatorname{spec}(r_p)}$. Now it follows from (8) that $\operatorname{spec}(r_p)$ lies in the open left half of the complex plane. Indeed, as $t \to \infty$, we have $e^{(r_p+s_p)t} \to 0$ because the matrix A_p is stable. Since $\mathfrak{s}$ is compact, there exists a constant $C > 0$ such that we have $|e^s x| \geq C|x|$ for all $s \in \mathfrak{s}$ and $x \in \mathbb{R}^n$; thus we cannot have $e^{s_p t} x \to 0$ for $x \neq 0$. Therefore, $B_p(t) \to 0$, and so r_p is stable.

Since $p \in \mathcal{P}$ was arbitrary, we see that all the matrices r_p, $p \in \mathcal{P}$, are stable. Theorem 1 implies that the switched linear system generated by these matrices is globally uniformly exponentially stable. Moreover, the same property holds for matrices in the extended set $\bar{\mathfrak{r}} := \{\bar{A} : \exists\, p \in \mathcal{P} \text{ and } s \in \mathfrak{s} \text{ such that } \bar{A} = e^{-s} r_p e^s\}$. This is true because the matrices in this set are stable and because they belong to $\mathfrak{r}$. (The last statement follows from the expansion (9) again since $[\mathfrak{s}, \mathfrak{r}] \subseteq \mathfrak{r}$.) Now the transition matrix of the original switched linear system (2) at time t takes the form

$$\Phi(t,0) = e^{(r_{p_k}+s_{p_k})t_k} \cdots e^{(r_{p_1}+s_{p_1})t_1} = e^{s_{p_k}t_k} B_{p_k}(t_k) \cdots e^{s_{p_1}t_1} B_{p_1}(t_1),$$

where t_1, t_1+t_2, $\ldots$, $t_1+t_2+\cdots+t_{k-1} < t$ are switching instants, $t_1+\cdots+t_k = t$, and, as before, $\dot{B}_{p_i}(t) = e^{-s_{p_i}t} r_{p_i} e^{s_{p_i}t} B_{p_i}(t)$, $i = 1, \ldots, k$. To simplify the notation, let $k = 2$. (In the general case one can adopt the same line of reasoning or use induction on k.) We can then write

$$\Phi(t,0) = e^{s_{p_2}t_2} e^{s_{p_1}t_1} e^{-s_{p_1}t_1} B_{p_2}(t_2) e^{s_{p_1}t_1} B_{p_1}(t_1) = e^{s_{p_2}t_2} e^{s_{p_1}t_1} \tilde{B}_{p_2}(t_2) B_{p_1}(t_1),$$

where $\tilde{B}_{p_2}(t) := e^{-s_{p_1}t_1} B_{p_2}(t) e^{s_{p_1}t_1}$. We have

$$\begin{aligned}
\frac{d}{dt}\tilde{B}_{p_2}(t) &= e^{-s_{p_1}t_1} e^{-s_{p_2}t} r_{p_2} e^{s_{p_2}t} B_{p_2}(t) e^{s_{p_1}t_1} \\
&= e^{-s_{p_1}t_1} e^{-s_{p_2}t} r_{p_2} e^{s_{p_2}t} e^{s_{p_1}t_1} e^{-s_{p_1}t_1} B_{p_2}(t) e^{s_{p_1}t_1} \\
&= e^{-s_{p_1}t_1} e^{-s_{p_2}t} r_{p_2} e^{s_{p_2}t} e^{s_{p_1}t_1} \tilde{B}_{p_2}(t).
\end{aligned}$$

Thus we see that

$$\Phi(t,0) = e^{s_{p_2}t_2} e^{s_{p_1}t_1} \cdot \bar{B}(t), \tag{10}$$

where $\bar{B}(t)$ is the transition matrix of a switched/time-varying system generated by matrices in $\bar{\mathfrak{r}}$, i.e., $\frac{d}{dt}\bar{B}(t) = \bar{A}(t)\bar{B}(t)$ with $\bar{A}(t) \in \bar{\mathfrak{r}}\ \forall t \geq 0$. The norm of the first term in the above product is bounded by compactness, while the norm of the second goes to zero exponentially by Theorem 1 (see also Remark 1), and the statement of the theorem follows. □

Remark 2. The fact that $\mathfrak{r}$ is the radical, implying that $\mathfrak{s}$ is semisimple, was not used in the proof. The statement of Theorem 2 remains valid for any decomposition of $\mathfrak{g}$ into the sum of a solvable ideal $\mathfrak{r}$ and a subalgebra $\mathfrak{s}$. Among all possible decompositions of this kind, the one considered above gives the strongest result. If $\mathfrak{g}$ is solvable, then $\mathfrak{s} = 0$ is of course compact, and we recover Theorem 1 as a special case.

Example 1. Suppose that the matrices A_p, $p \in \mathcal{P}$, take the form $A_p = -\lambda_p I + S_p$, where $\lambda_p > 0$ and $S_p^T = -S_p$ for all $p \in \mathcal{P}$. These are automatically stable matrices. Suppose also that $\operatorname{span}\{A_p, p \in \mathcal{P}\} \ni I$. Then the condition of Theorem 2 is satisfied. Indeed, take $\mathfrak{r} = \{\lambda I : \lambda \in \mathbb{R}\}$ (scalar multiples of the identity matrix) and observe that the Lie algebra $\{S_p : p \in \mathcal{P}\}_{LA}$ is compact because skew-symmetric matrices have purely imaginary eigenvalues.

In [12] the global uniform exponential stability property was deduced from the existence of a quadratic common Lyapunov function. In the present case we found it more convenient to obtain the desired result directly. However, under the hypothesis of Theorem 2, a quadratic common Lyapunov function for the family of linear systems (3) can also be constructed, as we now show. Let $\bar{V}(x) = x^T Q x$ be a quadratic

common Lyapunov function for the family of linear systems generated by matrices in $\bar{\mathfrak{r}}$ (which exists according to [12]). Define the function

$$V(x) := \int_{\mathcal{S}} \bar{V}(Sx)dS = x^T \cdot \int_{\mathcal{S}} S^T Q S dS \cdot x,$$

where $\mathcal{S}$ is the Lie group corresponding to $\mathfrak{s}$ and the integral is taken with respect to the Haar measure invariant under the right translation on $\mathcal{S}$ (see section A.4). Using (10), it is straightforward to show that the derivative of V along solutions of the switched linear system (2) satisfies

$$\begin{aligned}\frac{d}{dt}V(x(t)) &= \frac{d}{dt}\int_{\mathcal{S}} \bar{V}(S\bar{B}(t)x(0))dS \\ &= \int_{\mathcal{S}} x^T(0)\bar{B}^T(t)S^T((S\bar{A}(t)S^{-1})^T Q + QS\bar{A}(t)S^{-1})S\bar{B}(t)x(0)dS < 0.\end{aligned}$$

The first equality in the above formula follows from the invariance of the measure, and the last inequality holds because $S\bar{A}(t)S^{-1} \in \bar{\mathfrak{r}}$ for all $t \geq 0$ and all $S \in \mathcal{S}$.

Remark 3. It is now clear that the above results remain valid if piecewise constant switching signals are replaced by arbitrary measurable functions (cf. Remark 1).

The existence of a quadratic common Lyapunov function will be used to prove Corollary 3 below. It is also an interesting fact in its own right because, although the converse Lyapunov theorem proved in [15] implies that global uniform exponential stability always leads to the existence of a common Lyapunov function, in some cases it is not possible to find a quadratic one [4]. Incidentally, this clearly shows that the condition of Theorem 2 is not necessary for uniform exponential stability of the switched linear system (2). Another way to see this is to note that the property of uniform exponential stability is robust with respect to small perturbations of the parameters of the system, whereas the condition of Theorem 2 is not. In fact, no Lie-algebraic condition of the type considered here can possess the indicated robustness property. This follows from the fact, proved in section A.6, that in an arbitrarily small neighborhood of any pair of $n \times n$ matrices there exists a pair of matrices that generate the entire Lie algebra $gl(n, \mathbb{R})$.

We conclude this section with a local stability result for the nonlinear switched system (1). Let $f_p : D \to \mathbb{R}^n$ be continuously differentiable with $f_p(0) = 0$ for each $p \in \mathcal{P}$, where D is a neighborhood of the origin in $\mathbb{R}^n$. Consider the linearization matrices

$$F_p := \frac{\partial f_p}{\partial x}(0), \qquad p \in \mathcal{P}.$$

Assume that the matrices F_p are stable, that $\mathcal{P}$ is a compact subset of some topological space, and that $\frac{\partial f_p}{\partial x}(x)$ depends continuously on p for each $x \in D$. Consider the Lie algebra $\tilde{\mathfrak{g}} := \{F_p : p \in \mathcal{P}\}_{LA}$ and its Levi decomposition $\tilde{\mathfrak{g}} = \tilde{\mathfrak{r}} \oplus \tilde{\mathfrak{s}}$. The following statement is a generalization of [12, Corollary 5].

COROLLARY 3. *If $\tilde{\mathfrak{s}}$ is a compact Lie algebra, then the switched system* (1) *is uniformly exponentially stable.*

Proof. This is a relatively straightforward application of Lyapunov's first method (see, e.g., [11]). For each $p \in \mathcal{P}$ we can write $f_p(x) = F_p x + g_p(x)x$. Here $g_p(x) = \frac{\partial f_p}{\partial x}(z) - \frac{\partial f_p}{\partial x}(0)$, where z is a point on the line segment connecting x to the origin. We have $g_p(x) \to 0$ as $x \to 0$. Under the present assumptions, the family of linear systems $\dot{x} = F_p x$, $p \in \mathcal{P}$, has a quadratic common Lyapunov function. Because of

compactness of $\mathcal{P}$ and continuity of $\frac{\partial f_p}{\partial x}$ with respect to p, it is not difficult to verify that this function is a common Lyapunov function for the family of systems $\dot{x} = f_p(x)$, $p \in \mathcal{P}$, on a certain neighborhood $\bar{D}$ of the origin. Thus the switched system (1) is uniformly exponentially stable on $\bar{D}$. □

An important problem for future research is to investigate how the structure of the Lie algebra generated by the original nonlinear vector fields f_p, $p \in \mathcal{P}$, is related to stability properties of the switched system (1). Taking higher-order terms into account, one may hope to obtain conditions that guarantee stability of nonlinear switched systems when the above linearization test fails. A first step in this direction is the observation made in [21] that a finite family of commuting nonlinear vector fields giving rise to exponentially stable systems has a local common Lyapunov function. Imposing certain additional assumptions, it is possible to obtain analogues of Lie's theorem which yield triangular structure for families of nonlinear systems generating nilpotent or solvable Lie algebras (see [3, 10, 14]). However, the methods described in these papers require that the Lie algebra have full rank, and so typically they do not apply to families of systems with common equilibria of the type treated here.

3. A converse result. We already remarked that the condition of Theorem 2 is not necessary for uniform exponential stability of the switched linear system (2). It is natural to ask whether this condition can be improved. A more general question that arises is to what extent the structure of the Lie algebra can be used to distinguish between stable and unstable switched systems. The findings of this section will shed some light on these issues.

We find it useful to introduce a possibly larger Lie algebra $\hat{\mathfrak{g}}$ by adding to $\mathfrak{g}$ the scalar multiples of the identity matrix if necessary. In other words, define $\hat{\mathfrak{g}} := \{I, A_p : p \in \mathcal{P}\}_{LA}$. The Levi decomposition of $\hat{\mathfrak{g}}$ is given by $\hat{\mathfrak{g}} = \hat{\mathfrak{r}} \oplus \mathfrak{s}$ with $\hat{\mathfrak{r}} \supseteq \mathfrak{r}$ (because the subspace $\mathbb{R}I$ belongs to the radical of $\hat{\mathfrak{g}}$). Thus $\hat{\mathfrak{g}}$ satisfies the hypothesis of Theorem 2 if and only if $\mathfrak{g}$ does.

Our goal in this section is to show that if this hypothesis is not satisfied, then $\hat{\mathfrak{g}}$ can be generated by a family of stable matrices (which might in principle be different from $\{A_p : p \in \mathcal{P}\}$) with the property that the corresponding switched linear system is not stable. Such a statement could in some sense be interpreted as a converse of Theorem 2. It would imply that by working just with $\hat{\mathfrak{g}}$ it is not possible to obtain a stronger result than the one given in the previous section.

We will also see that there exists another set of stable generators for $\hat{\mathfrak{g}}$ which does give rise to a uniformly exponentially stable switched linear system. In fact, we will show that both generator sets can always be chosen in such a way that they contain the same number of elements as the original set that was used to generate $\hat{\mathfrak{g}}$. Thus, if the Lie algebra does not satisfy the hypothesis of Theorem 2, this Lie algebra alone (even together with the knowledge of how many stable matrices were used to generate it) does not provide enough information to determine whether or not the original switched linear system is stable.

Let $\{A_1, A_2, \ldots, A_m\}$ be any finite set of stable generators for $\hat{\mathfrak{g}}$. (If the index set $\mathcal{P}$ is infinite, a suitable finite subset can always be extracted from it.) Then the following holds.

THEOREM 4. *Suppose that $\mathfrak{s}$ is not a compact Lie algebra. Then there exists a set of m stable generators for $\hat{\mathfrak{g}}$ such that the corresponding switched linear system is not uniformly exponentially stable. There also exists another set of m stable generators for $\hat{\mathfrak{g}}$ such that the corresponding switched linear system is globally uniformly exponentially stable.*

Proof. To prove the second statement of the theorem, we simply subtract λI from each of the generators $A_1, A_2, \ldots, A_m$, where $\lambda > 0$ is large enough. Namely, take λ to be any number larger than the largest eigenvalue of $(A_i + A_i^T)/2$ for all $i = 1, \ldots, m$. Then it is easy to check that the linear systems defined by the matrices $A_1 - \lambda I, A_2 - \lambda I, \ldots, A_m - \lambda I$ all share the common Lyapunov function $V(x) = x^T x$. To prove that these matrices indeed generate $\hat{\mathfrak{g}}$, it is enough to show that the span of these matrices and their iterated Lie brackets contains the identity matrix I. We know that I can be written as a linear combination of the matrices $A_1, A_2, \ldots, A_m$, and their suitable Lie brackets. Replacing each A_i in this linear combination by $A_i - \lambda I$, we obtain a scalar multiple of I. If it is nonzero, we are done; otherwise, we just have to increase λ by an arbitrary amount.

We now turn to the first statement of the theorem. Since $\mathfrak{s}$ is not compact, it contains a subalgebra that is isomorphic to $sl(2, \mathbb{R})$. Such a subalgebra can be constructed as shown in section A.5. The existence of this subalgebra is the key property that we will explore.

It follows from basic properties of solutions to differential inclusions that if a family of matrices gives rise to a uniformly exponentially stable switched linear system, then all convex linear combinations of these matrices are stable. (This fact is easily seen to be true from the converse Lyapunov theorems of [15, 4], although in [15] it was actually used to prove the result; see also Remark 5 below.) To prove the theorem, we will first find a pair of stable matrices B_1, B_2 that lie in the subalgebra isomorphic to $sl(2, \mathbb{R})$ and have an unstable convex combination, and then we will use them to construct a desired set of generators for $\hat{\mathfrak{g}}$. (An alternative method of proof will be presented in the next section.)

Since every matrix representation of $sl(2, \mathbb{R})$ is a direct sum of irreducible ones, there is no loss of generality in considering only irreducible representations. Their complete classification in all dimensions (up to equivalence induced by linear coordinate transformations) is available. In particular, it is known that any irreducible representation of $sl(2, \mathbb{R})$ contains two matrices of the following form:

$$\tilde{B}_1 = \begin{pmatrix} 0 & \mu_1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \mu_r \\ 0 & \cdots & \cdots & 0 \end{pmatrix} \quad \text{and} \quad \tilde{B}_2 = \begin{pmatrix} 0 & \cdots & \cdots & 0 \\ 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 \end{pmatrix}$$

(cf. section A.2). The matrix $\tilde{B}_1$ has positive entries $\mu_1, \ldots, \mu_r$ immediately above the main diagonal and zeros elsewhere, and the matrix $\tilde{B}_2$ has ones immediately below the main diagonal and zeros elsewhere.

It is not hard to check that the nonnegative matrix $\tilde{B} := (\tilde{B}_1 + \tilde{B}_2)/2$ is irreducible[1] and as such satisfies the assumptions of the Perron–Frobenius theorem (see, e.g., [6, Chapter XIII]). According to that theorem, $\tilde{B}$ has a positive eigenvalue. Then for a small enough $\epsilon > 0$ the matrix $B := \tilde{B} - \epsilon I$ also has a positive eigenvalue. We have $B = (\tilde{B}_1 - \epsilon I + \tilde{B}_2 - \epsilon I)/2$. This implies that a desired pair of matrices in the given irreducible matrix representation of $sl(2, \mathbb{R})$ can be defined by $B_1 := \tilde{B}_1 - \epsilon I$ and $B_2 := \tilde{B}_2 - \epsilon I$. Indeed, these matrices are stable, but their average B is not.

For $\alpha \geq 0$, define $A_1(\alpha) := B_1 + \alpha A_1$ and $A_2(\alpha) := B_2 + \alpha A_2$. If α is small enough, then $A_1(\alpha)$ and $A_2(\alpha)$ are stable matrices, while $(A_1(\alpha) + A_2(\alpha))/2$ is unstable. Thus

[1] A matrix is called *irreducible* if it has no proper invariant subspaces spanned by coordinate vectors.

the matrices $A_1(\alpha), A_2(\alpha), A_3, \ldots, A_m$ yield a switched system that is not uniformly exponentially stable. Moreover, it is not hard to show that for α small enough these matrices generate $\hat{\mathfrak{g}}$. Indeed, consider a basis for $\hat{\mathfrak{g}}$ formed by $A_1, \ldots, A_m$, and their suitable Lie brackets. Replacing A_1 and A_2 in these expressions by $A_1(\alpha)$ and $A_2(\alpha)$ and writing the coordinates of the resulting elements relative to this basis, we obtain a square matrix $\Delta(\alpha)$. Its determinant is a polynomial in α whose value tends to ∞ as $\alpha \to \infty$, and therefore it is not identically zero. Thus $\Delta(\alpha)$ is nondegenerate for all but finitely many values of α; in particular, we will have a basis for $\hat{\mathfrak{g}}$ if we take α sufficiently small. This completes the proof. □

Remark 4. Given the matrices B_1 and B_2 as in the above proof, it is of course quite easy to construct a set of stable generators for $\hat{\mathfrak{g}}$ giving rise to a switched linear system that is not uniformly exponentially stable: just take any set of generators for $\hat{\mathfrak{g}}$ containing $-I$, B_1, and B_2, and make them into stable ones by means of subtracting positive multiples of the identity if necessary. The above more careful construction has the advantage of producing a set of generators with the same number of elements as in the original generating set for $\hat{\mathfrak{g}}$.

Remark 5. The existence of an unstable convex combination actually leads to more specific conclusions than simply a lack of uniform exponential stability. Namely, one can find a sequence of solutions of the switched system that converges in a suitable sense to a trajectory of the unstable linear system associated with such a convex combination. This is a consequence of the so-called *relaxation theorem* which in our case says that the set of solutions to the differential inclusion $\dot{x} \in \{A_p x : p \in \mathcal{P}\}$ is dense in the set of solutions to the differential inclusion $\dot{x} \in \mathrm{co}\{A_p x : p \in \mathcal{P}\}$, where $\mathrm{co}(K)$ denotes the convex hull of a set $K \subset \mathbb{R}^n$. For details, see [2, 5].

The results that we have obtained so far reveal the following important fact: the property of $\hat{\mathfrak{g}}$ which is being investigated here, namely, global uniform exponential stability of any switched system whose associated Lie algebra is $\hat{\mathfrak{g}}$, depends only on the structure of $\hat{\mathfrak{g}}$ (i.e., on the commutation relations between its matrices) and is independent of the choice of a particular representation.

4. Switched linear systems with low-dimensional Lie algebras. In the proof of Theorem 4 in the previous section, we needed to construct a pair of stable matrices in a representation of $sl(2, \mathbb{R})$ which give rise to an unstable switched system. To achieve this, we relied on the fact that a switched system defined by two matrices is not stable if these matrices have an unstable convex combination. However, even if all convex combinations are stable, stability of the switched system is not guaranteed. As a simple example that illustrates this, consider the switched system in $\mathbb{R}^2$ defined by the matrices $A_1 := \tilde{A}_1 - \epsilon I$ and $A_2 := \tilde{A}_2 - \epsilon I$, where

$$\tilde{A}_1 := \begin{pmatrix} 0 & k \\ -1 & 0 \end{pmatrix}, \qquad \tilde{A}_2 := \begin{pmatrix} 0 & 1 \\ -k & 0 \end{pmatrix}$$

with $\epsilon > 0$ and $k > 1$. It is easy to check that all convex combinations of A_1 and A_2 are stable. When $\epsilon = 0$, the trajectories of the corresponding individual systems look as shown in Figure 1 (left) and Figure 1 (center), respectively. It is not hard to find a switching signal $\sigma : [0, \infty) \to \{1, 2\}$ that makes the switched system $\dot{x} = \tilde{A}_\sigma x$ unstable: simply let $\sigma = 1$ when $xy > 0$ and $\sigma = 2$ otherwise. For an arbitrary initial state, this results in the switched system $\dot{x} = \tilde{A}_{\sigma(t)} x$ whose solutions grow exponentially. Therefore, the original switched system $\dot{x} = A_\sigma x$ will also be destabilized by the same switching signal, provided that ϵ is sufficiently small.

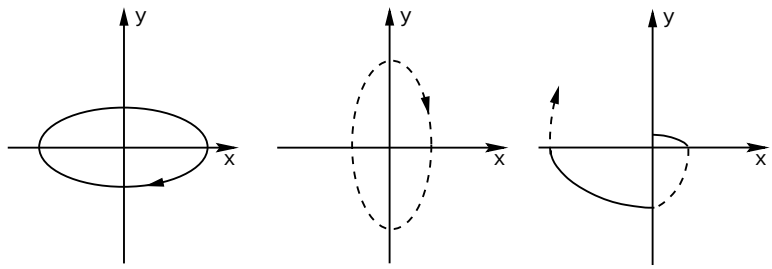


FIG. 1. *Unstable switched system in the plane.*

As a step toward understanding the behavior of switched systems in higher dimensions, in view of the findings of this paper it is natural to investigate the case when given matrices generate a Lie algebra that is isomorphic to one generated by 2×2 matrices. This is the goal of the present section.

Consider the Lie algebra $\mathfrak{g} := \{A_p : p \in \mathcal{P}\}_{LA}$, and assume that $\mathfrak{g} = \mathbb{R}I_{n\times n} \oplus sl(2,\mathbb{R})$. Here $sl(2,\mathbb{R})$ means an n-dimensional matrix representation, which we take to be irreducible. (As before, this will not introduce a loss of generality because every matrix representation of $sl(2,\mathbb{R})$ is a direct sum of irreducible ones.) Then for each $p \in \mathcal{P}$ we can write

$$A_p = (n-1)\alpha_p I_{n\times n} + \beta_p \phi(e) + \gamma_p \phi(f) + \delta_p \phi(h), \tag{11}$$

where $\beta_p, \gamma_p, \delta_p$ are constants, ϕ is the standard representation of $sl(2,\mathbb{R})$ constructed in section A.2 (n here corresponds to $k+1$ there), $\{e, h, f\}$ is the canonical basis for $sl(2,\mathbb{R})$, and $\alpha_p = \frac{1}{n(n-1)}\mathrm{trace}(A_p)$. For each $p \in \mathcal{P}$, define the following 2×2 matrix:

$$\hat{A}_p := \alpha_p I_{2\times 2} - \beta_p e - \gamma_p f - \delta_p h. \tag{12}$$

We now demonstrate that the task of investigating stability of the switched system generated by the matrices A_p, $p \in \mathcal{P}$, reduces to that of investigating stability of the two-dimensional switched system generated by the matrices $\hat{A}_p$, $p \in \mathcal{P}$.

PROPOSITION 5. *The switched linear system* (2) *with* A_p *given by* (11) *is globally uniformly exponentially stable if and only if the switched linear system* $\dot{x} = \hat{A}_\sigma x$ *with* $\hat{A}_p$ *given by* (12) *is globally uniformly exponentially stable.*

Proof. The transition matrix of the switched system (2) for any particular switching signal takes the form

$$\Phi(t,0) = e^{(n-1)(\alpha_{p_k}t_k + \cdots + \alpha_{p_1}t_1)I} e^{(\beta_{p_k}\phi(e) + \gamma_{p_k}\phi(f) + \delta_{p_k}\phi(h))t_k} \cdots e^{(\beta_{p_1}\phi(e) + \gamma_{p_1}\phi(f) + \delta_{p_1}\phi(h))t_1}.$$

Consider the (n-dimensional) linear space $P^{n-1}[x,y]$ of polynomials in x and y, homogeneous of degree $n-1$, with the basis chosen as in section A.2. Denote the elements of this basis by $p_1, \ldots, p_n$. (These are monomials in x and y.) Fix an arbitrary polynomial $p \in P^{n-1}[x,y]$, and let $a_1, \ldots, a_n$ be its coordinates relative to the above basis. As an immediate consequence of the calculations given in section A.2, for any values of x and y we have

$$\begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix} \Phi(t,0) \begin{pmatrix} p_1 \\ \vdots \\ p_n \end{pmatrix} = p\left(\hat{\Phi}(t,0)\begin{pmatrix} x \\ y \end{pmatrix}\right),$$

where

$$\hat{\Phi}(t,0) = e^{(\alpha_{p_k} t_k + \cdots + \alpha_{p_1} t_1) I} e^{(-\beta_{p_1} e - \gamma_{p_1} f - \delta_{p_1} h) t_1} \cdots e^{(-\beta_{p_k} e - \gamma_{p_k} f - \delta_{p_k} h) t_k}.$$

Since the polynomial p was arbitrary, it is clear that $\Phi(t,0)$ approaches the zero matrix as $t \to \infty$, uniformly over the set of all switching signals, if and only if $\hat{\Phi}(t,0)$ does so. But $\hat{\Phi}(t,0)$ is the transition matrix of the switched system $\dot{x} = \hat{A}_\sigma x$, corresponding to the "reversed" switching signal on $[0,t]$. We conclude that this switched system is globally asymptotically stable, uniformly over σ, if and only if the same property holds for the original system (2). The statement of the proposition now follows from the fact that for switched linear systems, uniform asymptotic stability is equivalent to uniform exponential stability. □

We are now in position to justify the reduction procedure outlined in the introduction. Assume that $\hat{\mathfrak{g}}$ has dimension at most 4. We know from section A.5 that any noncompact semisimple Lie algebra contains a subalgebra isomorphic to $sl(2,\mathbb{R})$. Thus $\hat{\mathfrak{g}}$ contains a noncompact semisimple subalgebra if only if its dimension exactly equals 4 and the Killing form is nondegenerate and sign-indefinite on $\tilde{\mathfrak{g}} = \{\tilde{A}_1, \tilde{A}_2\}_{LA} = \hat{\mathfrak{g}} \bmod \mathbb{R}I$ (see section A.4). In this case $\tilde{g}$ is isomorphic to $sl(2,\mathbb{R})$. An $sl(2)$-triple $\{h,e,f\}$ can be constructed as explained in section A.5. (The procedure given there for a general noncompact semisimple Lie algebra certainly applies to $sl(2,\mathbb{R})$ itself.) Specifically, as h we can take any element of the subspace on which the Killing form is positive definite, normalized in such a way that the eigenvalues of $\mathrm{ad}h$ equal 2 and -2. The corresponding eigenvectors yield e and f. The resulting representation of $sl(2,\mathbb{R})$ is not necessarily irreducible; the dimension of the largest invariant subspace is equal to $k+1$, where k is the largest eigenvalue of h. If the switched linear system restricted to this invariant subspace is globally uniformly exponentially stable, then the same property holds for the switched linear system restricted to any other invariant subspace. This is true because, in view of the role of the scalar $k = n-1$ in the context of Proposition 5, the matrices of the reduced (second-order) system associated with the system evolving on the largest invariant subspace are obtained from those of the reduced system associated with the system evolving on another invariant subspace by subtracting positive multiples of the identity matrix, and this cannot introduce instability (to see why this last statement is true, one can appeal to the existence of a convex common Lyapunov function [15]). Note that we do not need to identify the invariant subspaces; we need to know only the dimension of the largest one. Thus the outcome of the algorithm depends on the matrix representation of $\hat{\mathfrak{g}}$ and not just on the structure of $\hat{\mathfrak{g}}$ as a Lie algebra, but it does so in a rather weak way.

As another application of Proposition 5, we can obtain an alternative proof of Theorem 4. Indeed, let the matrices $\tilde{B}_1$ and $\tilde{B}_2$ be as in the proof of Theorem 4 given in the previous section. (The existence of a subalgebra isomorphic to $sl(2,\mathbb{R})$ remains crucial.) Define the matrices $B_1 := -k\tilde{B}_1 + \tilde{B}_2 - \epsilon I$ and $B_2 := -\tilde{B}_1 + k\tilde{B}_2 - \epsilon I$, where $\epsilon > 0$ and $k > 1$. Then the switched system

$$\dot{x} = B_\sigma x, \qquad \sigma : [0,\infty) \to \{1,2\} \tag{13}$$

is not stable for ϵ small enough (even though all convex combinations of B_1 and B_2 are stable). This follows from Proposition 5 and from the example presented at the beginning of this section; in fact, a specific (periodic) destabilizing switching signal for the system (13) can be constructed with the help of that example. Interestingly,

it appears to be difficult to establish the same result by a direct analysis of (13). The rest of the proof of Theorem 4 can now proceed exactly as before.

It was shown by Shorten and Narendra in [22] that two stable two-dimensional linear systems $\dot{x} = A_1 x$ and $\dot{x} = A_2 x$ possess a quadratic common Lyapunov function if and only if all pairwise convex combinations of matrices from the set $\{A_1, A_2, A_1^{-1}, A_2^{-1}\}$ are stable. Combined with Proposition 5, this yields the following result.

COROLLARY 6. *Let $\mathcal{P} = \{1, 2\}$. Suppose that all pairwise convex combinations of matrices from the set $\{\hat{A}_1, \hat{A}_2, \hat{A}_1^{-1}, \hat{A}_2^{-1}\}$, with A_1 and A_2 given by* (12)*, are stable. Then the switched linear system* (2)*, with A_p given by* (11)*, is globally uniformly exponentially stable.*

The above corollary provides only sufficient and not necessary conditions for global uniform exponential stability of (2). This is due to the fact that, as we already mentioned earlier, it may happen that a switched linear system is globally uniformly exponentially stable while there is no quadratic common Lyapunov function for the individual subsystems (see the example in [4]).

Appendix. Basic facts about Lie algebras. In this appendix we give an informal overview of basic properties of Lie algebras. Only those facts that directly play a role in the developments of the previous sections are discussed. Most of the material is adopted from [8, 20], and the reader is referred to these and other standard references for more details.

A.1. Lie algebras and their representations. A *Lie algebra* $\mathfrak{g}$ is a finite-dimensional vector space equipped with a *Lie bracket*, i.e., a bilinear, skew-symmetric map $[\cdot, \cdot] : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$ satisfying the Jacobi identity $[a, [b, c]] + [b, [c, a]] + [c, [a, b]] = 0$. Any Lie algebra $\mathfrak{g}$ can be identified with a tangent space at the identity of a Lie group $\mathcal{G}$ (an analytic manifold with a group structure). If $\mathfrak{g}$ is a matrix Lie algebra, then the elements of $\mathcal{G}$ are given by products of the exponentials of the matrices from $\mathfrak{g}$. In particular, each element $A \in \mathfrak{g}$ generates the *one-parameter subgroup* $\{e^{At}, t \in \mathbb{R}\}$ in $\mathcal{G}$. For example, if $\mathfrak{g}$ is the Lie algebra $gl(n, \mathbb{R})$ of all real $n \times n$ matrices with the standard Lie bracket $[A, B] = AB - BA$, then the corresponding Lie group is given by the invertible matrices.

Given an abstract Lie algebra $\mathfrak{g}$, one can consider its (matrix) representations. A *representation* of $\mathfrak{g}$ on an n-dimensional vector space V is a homomorphism (i.e., a linear map that preserves the Lie bracket) $\phi : \mathfrak{g} \to gl(V)$. It assigns to each element $g \in \mathfrak{g}$ a linear operator $\phi(g)$ on V, which can be described by an $n \times n$ matrix. A representation ϕ is called *irreducible* if V contains no nontrivial subspaces invariant under the action of all $\phi(g)$, $g \in \mathfrak{g}$. A particularly useful representation is the *adjoint* one, denoted by "ad." The vector space V in this case is $\mathfrak{g}$ itself, and for $g \in \mathfrak{g}$ the operator $\mathrm{ad}g$ is defined by $\mathrm{ad}g(a) := [g, a]$, $a \in \mathfrak{g}$. There is also *Ado's theorem*, which says that every Lie algebra is isomorphic to a subalgebra of $gl(V)$ for some finite-dimensional vector space V. (Compare this with the adjoint representation which is in general not injective.)

A.2. Example: $sl(2, \mathbb{R})$ and $gl(2, \mathbb{R})$. The *special linear Lie algebra* $sl(2, \mathbb{R})$ consists of all real 2×2 matrices of trace 0. A canonical basis for this Lie algebra is given by the matrices

$$h := \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad e := \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad f := \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}. \tag{14}$$

They satisfy the relations $[h,e]=2e$, $[h,f]=-2f$, $[e,f]=h$ and form what is sometimes called an $sl(2)$-*triple*. One can also consider other representations of $sl(2,\mathbb{R})$. Although all irreducible representations of $sl(2,\mathbb{R})$ can be classified by working with the Lie algebra directly (see [20, pp. 27–30]), for our purposes it is more useful to exploit the corresponding Lie group $SL(2,\mathbb{R})=\{S\in\mathbb{R}^{n\times n}:\det S=1\}$. Let $P^k[x,y]$ denote the space of polynomials in two indeterminates x and y that are homogeneous of degree k (where k is a positive integer). A homomorphism ϕ that makes $SL(2,\mathbb{R})$ act on $P^k[x,y]$ can be defined as

$$\phi(S)p\left(\begin{pmatrix}x\\y\end{pmatrix}\right)=p\left(S^{-1}\begin{pmatrix}x\\y\end{pmatrix}\right),$$

where $S\in SL(2,\mathbb{R})$ and $p\in P^k[x,y]$. The corresponding representation of the Lie algebra $sl(2,\mathbb{R})$, which we denote also by ϕ with slight abuse of notation, is obtained by considering the one-parameter subgroups of $SL(2,\mathbb{R})$ and differentiating the action defined above at $t=0$. For example, for e as in (14) we have

$$\phi(e)p\left(\begin{pmatrix}x\\y\end{pmatrix}\right)=\frac{d}{dt}\Big|_{t=0}p\left(e^{-et}\begin{pmatrix}x\\y\end{pmatrix}\right)=\frac{d}{dt}\Big|_{t=0}p\left(\begin{pmatrix}1&-t\\0&1\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}\right)=-y\frac{\partial}{\partial x}p\left(\begin{pmatrix}x\\y\end{pmatrix}\right).$$

Similarly, $\phi(f)p=-x\frac{\partial}{\partial y}p$ and $\phi(h)p=(-x\frac{\partial}{\partial x}+y\frac{\partial}{\partial y})p$. With respect to the basis in $P^k[x,y]$ given by the monomials $y^k,-ky^{k-1}x,k(k-1)y^{k-2}x^2,\dots,(-1)^kk!x^k$, the corresponding differential operators are realized by the matrices

$$h\mapsto\begin{pmatrix}k&\cdots&\cdots&0\\\vdots&k-2&&\vdots\\\vdots&&\ddots&\vdots\\0&\cdots&\cdots&-k\end{pmatrix},\ e\mapsto\begin{pmatrix}0&\mu_1&\cdots&0\\\vdots&\ddots&\ddots&\vdots\\\vdots&&\ddots&\mu_k\\0&\cdots&\cdots&0\end{pmatrix},\ f\mapsto\begin{pmatrix}0&\cdots&\cdots&0\\1&\ddots&&\vdots\\\vdots&\ddots&\ddots&\vdots\\0&\cdots&1&0\end{pmatrix},$$

where $\mu_i=i(k-i+1)$, $i=1,\dots,k$. It turns out that any irreducible representation of $sl(2,\mathbb{R})$ of dimension $k+1$ is equivalent (under a linear change of coordinates) to the one just described. An arbitrary representation of $sl(2,\mathbb{R})$ is a direct sum of irreducible ones.

When working with $gl(2,\mathbb{R})$ rather than $sl(2,\mathbb{R})$, one also has the 2×2 identity matrix $I_{2\times 2}$. It corresponds to the operator $x\frac{\partial}{\partial x}+y\frac{\partial}{\partial y}$ on $P^k[x,y]$, whose associated matrix is $kI_{(k+1)\times(k+1)}$. One can thus naturally extend the above representation to $gl(2,\mathbb{R})$. The complementary subalgebras $\mathbb{R}I$ and $sl(2,\mathbb{R})$ are invariant under the resulting action.

A.3. Nilpotent and solvable Lie algebras. If $\mathfrak{g}_1$ and $\mathfrak{g}_2$ are linear subspaces of a Lie algebra $\mathfrak{g}$, one writes $[\mathfrak{g}_1,\mathfrak{g}_2]$ for the linear space spanned by all the products $[g_1,g_2]$ with $g_1\in\mathfrak{g}_1$ and $g_2\in\mathfrak{g}_2$. Given a Lie algebra $\mathfrak{g}$, the sequence $\mathfrak{g}^{(k)}$ is defined inductively as follows: $\mathfrak{g}^{(1)}:=\mathfrak{g}$, $\mathfrak{g}^{(k+1)}:=[\mathfrak{g}^{(k)},\mathfrak{g}^{(k)}]\subset\mathfrak{g}^{(k)}$. If $\mathfrak{g}^{(k)}=0$ for k sufficiently large, then $\mathfrak{g}$ is called *solvable*. Similarly, one defines the sequence $\mathfrak{g}^k$ by $\mathfrak{g}^1:=\mathfrak{g}$, $\mathfrak{g}^{k+1}:=[\mathfrak{g},\mathfrak{g}^k]\subset\mathfrak{g}^k$ and calls $\mathfrak{g}$ *nilpotent* if $\mathfrak{g}^k=0$ for k sufficiently large. For example, if $\mathfrak{g}$ is a Lie algebra generated by two matrices A and B, we have: $\mathfrak{g}^{(1)}=\mathfrak{g}^1=\mathfrak{g}=\operatorname{span}\{A,B,[A,B],[A,[A,B]],\dots\}$, $\mathfrak{g}^{(2)}=\mathfrak{g}^2=\operatorname{span}\{[A,B],[A,[A,B]],\dots\}$, $\mathfrak{g}^{(3)}=\operatorname{span}\{[[A,B],[A,[A,B]]],\dots\}\subset\mathfrak{g}^3=\operatorname{span}\{[A,[A,B]],[B,[A,B]],\dots\}$, and so on. Every nilpotent Lie algebra is solvable, but the converse is not true.

The *Killing form* on a Lie algebra $\mathfrak{g}$ is the symmetric bilinear form K given by $K(a,b):=\operatorname{tr}(\operatorname{ad}a\circ\operatorname{ad}b)$ for $a,b\in\mathfrak{g}$. *Cartan's 1st criterion* says that $\mathfrak{g}$ is solvable

if and only if its Killing form vanishes identically on $[\mathfrak{g},\mathfrak{g}]$. Let $\mathfrak{g}$ be a solvable Lie algebra over an algebraically closed field, and let ϕ be a representation of $\mathfrak{g}$ on a vector space V. *Lie's theorem* states that there exists a basis for V with respect to which all the matrices $\phi(g)$, $g \in \mathfrak{g}$, are upper-triangular.

A.4. Semisimple and compact Lie algebras. A subalgebra $\bar{\mathfrak{g}}$ of a Lie algebra $\mathfrak{g}$ is called an *ideal* if $[g,\bar{g}] \in \bar{\mathfrak{g}}$ for all $g \in \mathfrak{g}$ and $\bar{g} \in \bar{\mathfrak{g}}$. Any Lie algebra has a unique maximal solvable ideal $\mathfrak{r}$, the *radical*. A Lie algebra $\mathfrak{g}$ is called *semisimple* if its radical is 0. *Cartan's 2nd criterion* says that $\mathfrak{g}$ is semisimple if and only if its Killing form is nondegenerate (meaning that if for some $g \in \mathfrak{g}$ we have $K(g,a) = 0\ \forall a \in \mathfrak{g}$, then g must be 0).

A semisimple Lie algebra is called *compact* if its Killing form is negative definite. A general *compact Lie algebra* is a direct sum of a semisimple compact Lie algebra and a commutative Lie algebra (with the Killing form vanishing on the latter). This terminology is justified by the facts that the tangent algebra of any compact Lie group is compact according to this definition, and that for any compact Lie algebra $\mathfrak{g}$ there exists a connected compact Lie group $\mathcal{G}$ with tangent algebra $\mathfrak{g}$. Compactness of a semisimple matrix Lie algebra $\mathfrak{g}$ amounts to the property that the eigenvalues of all matrices in $\mathfrak{g}$ lie on the imaginary axis. If $\mathcal{G}$ is a compact Lie group, one can associate to any continuous function $f : \mathcal{G} \to \mathbb{R}$ a real number $\int_{\mathcal{G}} f(G)dG$ so as to have $\int_{\mathcal{G}} 1dG = 1$ and $\int_{\mathcal{G}} f(AGB)dG = \int_{\mathcal{G}} f(G)dG \quad \forall A, B \in \mathcal{G}$ (left and right invariance). The measure dG is called the *Haar measure*.

An arbitrary Lie algebra $\mathfrak{g}$ can be decomposed into the semidirect sum $\mathfrak{g} = \mathfrak{r} \oplus \mathfrak{s}$, where $\mathfrak{r}$ is the radical, $\mathfrak{s}$ is a semisimple subalgebra, and $[\mathfrak{s},\mathfrak{r}] \subseteq \mathfrak{r}$ because $\mathfrak{r}$ is an ideal. This is known as a *Levi decomposition*. To compute $\mathfrak{r}$ and $\mathfrak{s}$, switch to a basis in which the Killing form K is diagonalized. The subspace on which K is not identically zero corresponds to $\mathfrak{s} \oplus (\mathfrak{r} \bmod \mathfrak{n})$, where $\mathfrak{n}$ is the maximal nilpotent subalgebra of $\mathfrak{r}$. Construct the Killing form $\bar{K}$ for the factor algebra $\mathfrak{s} \oplus (\mathfrak{r} \bmod \mathfrak{n})$. This form will vanish identically on $(\mathfrak{r} \bmod \mathfrak{n})$ and will be nondegenerate on $\mathfrak{s}$. The subalgebra $\mathfrak{s}$ identified in this way is compact if and only if $\bar{K}$ is negative definite on it. For more details on this construction and examples, see [7, pp. 256–258].

A.5. Subalgebras isomorphic to $sl(2,\mathbb{R})$. Let $\mathfrak{g}$ be a real, noncompact, semisimple Lie algebra. Our goal here is to show that $\mathfrak{g}$ has a subalgebra isomorphic to $sl(2,\mathbb{R})$. To this end, consider a *Cartan decomposition* $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$, where $\mathfrak{k}$ is a maximal compact subalgebra of $\mathfrak{g}$ and $\mathfrak{p}$ is its orthogonal complement with respect to K. The Killing form K is negative definite on $\mathfrak{k}$ and positive definite on $\mathfrak{p}$. Let $\mathfrak{a}$ be a maximal commuting subalgebra of $\mathfrak{p}$. Then it is easy to check using the Jacobi identity that the operators $\mathrm{ad}a$, $a \in \mathfrak{a}$, are commuting. These operators are also symmetric with respect to a suitable inner product on $\mathfrak{g}$ (for $a, b \in \mathfrak{g}$ this inner product is given by $-K(a,\Theta b)$, where Θ is the map sending $k + p$, with $k \in \mathfrak{k}$ and $p \in \mathfrak{p}$, to $k - p$), and hence they are simultaneously diagonalizable. Thus $\mathfrak{g}$ can be decomposed into a direct sum of subspaces invariant under $\mathrm{ad}a$, $a \in \mathfrak{a}$, on each of which every operator $\mathrm{ad}a$ has exactly one eigenvalue. The unique eigenvalue of $\mathrm{ad}a$ on each of these invariant subspaces is given by a linear function λ on $\mathfrak{a}$, and accordingly the corresponding subspace is denoted by $\mathfrak{g}_\lambda$. Since $\mathfrak{p} \neq 0$ (because $\mathfrak{g}$ is not compact) and since K is positive definite on $\mathfrak{p}$, the subspace $\mathfrak{g}_0$ associated with λ being identically zero cannot be the entire $\mathfrak{g}$. Summarizing, we have

$$\mathfrak{g} = \mathfrak{g}_0 \oplus \big(\bigoplus\nolimits_{\lambda \in \Sigma} \mathfrak{g}_\lambda\big),$$

where Σ is a finite set of nonzero linear functions on $\mathfrak{a}$ (which are called the *roots*) and $\mathfrak{g}_\lambda = \{g \in \mathfrak{g} : \mathrm{ad}a(g) = \lambda(a)g \ \forall a \in \mathfrak{a}\}$. Using the Jacobi identity, one can show that $[\mathfrak{g}_\lambda, \mathfrak{g}_\mu]$ is a subspace of $\mathfrak{g}_{\lambda+\mu}$ if $\lambda + \mu \in \Sigma \cup \{0\}$, and equals 0 otherwise. This implies that the subspaces $\mathfrak{g}_\lambda$ and $\mathfrak{g}_\mu$ are orthogonal with respect to K unless $\lambda + \mu = 0$ (cf. [20, p. 38]). Since K is nondegenerate on $\mathfrak{g}$, it follows that if λ is a root, then so is $-\lambda$. Moreover, the subspace $[\mathfrak{g}_\lambda, \mathfrak{g}_{-\lambda}]$ of $\mathfrak{g}_0$ has dimension 1, and λ is not identically zero on it (cf. [20, pp. 39–40]). This means that there exist some elements $e \in \mathfrak{g}_\lambda$ and $f \in \mathfrak{g}_{-\lambda}$ such that $h := [e, f] \neq 0$. It is now easy to see that, multiplying e, f, and h by constants if necessary, we obtain an $sl(2)$-triple. Alternatively, we could finish the argument by noting that if $g \in \mathfrak{g}_\lambda$ for some $\lambda \in \Sigma$, then the operator $\mathrm{ad}g$ is nilpotent (because it maps each $\mathfrak{g}_\mu$ to $\mathfrak{g}_{\mu+\lambda}$, to $\mathfrak{g}_{\mu+2\lambda}$, and eventually to 0 since Σ is a finite set), and the existence of a subalgebra isomorphic to $sl(2, \mathbb{R})$ is guaranteed by the Jacobson–Morozov theorem.

A.6. Generators for $gl(2, \mathbb{R})$. This subsection is devoted to showing that in an arbitrarily small neighborhood of any pair of $n \times n$ matrices one can find another pair of matrices that generate the entire Lie algebra $gl(n, \mathbb{R})$. This fact demonstrates that Lie-algebraic stability conditions considered in the previous sections are never robust with respect to small perturbations of the matrices that define the switched system. Constructions like the one presented here have certainly appeared in the literature, but we are not aware of a specific reference.

We begin by finding some matrices B_1, B_2 that generate $gl(n, \mathbb{R})$. Let B_1 be a diagonal matrix $B_1 = \mathrm{diag}(b_1, b_2, \ldots, b_n)$ satisfying the following two properties.

1. $b_i - b_j \neq b_k - b_l$ if $(i, j) \neq (k, l)$.
2. $\sum_{i=1}^n b_i \neq 0$.

Denote by $od(n, \mathbb{R})$ the space of matrices with zero elements on the main diagonal. Let B_2 be any matrix in $od(n, \mathbb{R})$ such that all its off-diagonal elements are nonzero. It is easy to check that if $E_{i,j}$ is a matrix whose ijth element is 1 and all other elements are 0, where $i \neq j$, then $[B_1, E_{i,j}] = (b_i - b_j)E_{i,j}$. Thus it follows from property 1 above that B_2 does not belong to any proper subspace of $od(n, \mathbb{R})$ that is invariant with respect to the operator $\mathrm{ad}B_1$. Therefore, the linear space spanned by the iterated brackets $ad^k B_1(B_2)$ is the entire $od(n, \mathbb{R})$. Taking brackets of the form $[E_{i,j}, E_{j,i}]$, we generate all traceless diagonal matrices (cf. the example $[e, f] = h$ in section A.2). Since B_1 has a nonzero trace by property 2 above, we conclude that $\{B_1, B_2\}_{LA} = gl(n, \mathbb{R})$.

Now let A_1 and A_2 be two arbitrary $n \times n$ matrices. Using the matrices B_1 and B_2 just constructed, we can define $A_1(\alpha) := A_1 + \alpha B_1$ and $A_2(\alpha) := A_2 + \alpha B_2$, where $\alpha \geq 0$. The two matrices $A_1(\alpha)$ and $A_2(\alpha)$ generate $gl(n, \mathbb{R})$ for any sufficiently small α, as can be shown by using the same argument as the one employed at the end of the proof of Theorem 4. Thus one can take $(A_1(\alpha), A_2(\alpha))$ as a desired pair of matrices in a neighborhood of (A_1, A_2).

Acknowledgments. The second author is grateful to Steve Morse for constant encouragement and interest in this work and to Victor Protsak for helpful discussions on Lie algebras.

REFERENCES

[1] D. Angeli, *A note on stability of arbitrarily switched homogeneous systems*, Systems Control Lett., to appear.

[2] J. P. Aubin and A. Cellina, *Differential Inclusions*, Springer-Verlag, Berlin, 1984.

[3] P. E. Crouch, *Dynamical realizations of finite Volterra series*, SIAM J. Control Optim., 19 (1981), pp. 177–202.
[4] W. P. Dayawansa and C. F. Martin, *A converse Lyapunov theorem for a class of dynamical systems which undergo switching*, IEEE Trans. Automat. Control, 44 (1999), pp. 751–760.
[5] A. F. Filippov, *Differential Equations with Discontinuous Righthand Sides*, Kluwer, Dordrecht, The Netherlands, 1988.
[6] F. R. Gantmacher, *The Theory of Matrices*, Chelsea Publishing, New York, 1960.
[7] R. Gilmore, *Lie Groups, Lie Algebras, and Some of Their Applications*, John Wiley, New York, 1974.
[8] V. V. Gorbatsevich, A. L. Onishchik, and E. B. Vinberg, *Structure of Lie Groups and Lie Algebras*, Encyclopaedia Math. Sci. 41, Springer-Verlag, Berlin, 1994.
[9] L. Gurvits, *Stability of discrete linear inclusion*, Linear Algebra Appl., 231 (1995), pp. 47–85.
[10] M. Kawski, *Nilpotent Lie algebras of vectorfields*, J. Reine Angew. Math., 388 (1988), pp. 1–17.
[11] H. K. Khalil, *Nonlinear Systems*, Macmillan, New York, 1992.
[12] D. Liberzon, J. P. Hespanha, and A. S. Morse, *Stability of switched systems: A Lie-algebraic condition*, Systems Control Lett., 37 (1999), pp. 117–122.
[13] D. Liberzon and A. S. Morse, *Basic problems in stability and design of switched systems*, IEEE Control Systems Magazine, 19 (1999), pp. 59–70.
[14] A. Marigo, *Constructive necessary and sufficient conditions for strict triangularizability of driftless nonholonomic systems*, in Proceedings of th 38th IEEE Conference on Decision and Control, Phoenix, AZ, 1999, pp. 2138–2143.
[15] A. P. Molchanov and Y. S. Pyatnitskiy, *Criteria of absolute stability of differential and difference inclusions encountered in control theory*, Systems Control Lett., 13 (1989), pp. 59–64.
[16] Y. Mori, T. Mori, and Y. Kuroe, *A solution to the common Lyapunov function problem for continuous-time systems*, in Proceedings of the 36th IEEE Conference on Decision and Control, San Diego, CA, 1997, pp. 3530–3531.
[17] A. S. Morse, *Control using logic-based switching*, in Trends in Control: A European Perspective, A. Isidori, ed., Springer-Verlag, New York, 1995, pp. 69–113.
[18] K. S. Narendra and J. Balakrishnan, *A common Lyapunov function for stable LTI systems with commuting A-matrices*, IEEE Trans. Automat. Control, 39 (1994), pp. 2469–2471.
[19] T. Ooba and Y. Funahashi, *On a common quadratic Lyapunov function for widely distant systems*, IEEE Trans. Automat. Control, 42 (1997), pp. 1697–1699.
[20] H. Samelson, *Notes on Lie Algebras*, Van Nostrand Reinhold, New York, 1969.
[21] H. Shim, D. J. Noh, and J. H. Seo, *Common Lyapunov function for exponentially stable nonlinear systems*, presented at the Fourth SIAM Conference on Control and its Applications, 1998.
[22] R. N. Shorten and K. S. Narendra, *Necessary and sufficient conditions for the existence of a common quadratic Lyapunov function for two stable second order linear time-invariant systems*, in Proceedings of the American Control Conference, San Diego, CA, 1999, pp. 1410–1414.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 1, pp. 270–297

OPTIMAL CONTROL FOR CONTINUOUS-TIME LINEAR QUADRATIC PROBLEMS WITH INFINITE MARKOV JUMP PARAMETERS*

MARCELO D. FRAGOSO† AND JACK BACZYNSKI‡

Abstract. The subject matter of this paper is the optimal control problem for continuous-time linear systems subject to Markovian jumps in the parameters and the usual infinite-time horizon quadratic cost. What essentially distinguishes our problem from previous ones, inter alia, is that the Markov chain takes values on a countably infinite set. To tackle our problem, we make use of powerful tools from semigroup theory in Banach space and a decomplexification technique. The solution for the problem relies, in part, on the study of a countably infinite set of coupled algebraic Riccati equations (ICARE). Conditions for existence and uniqueness of a positive semidefinite solution of the ICARE are obtained via the extended concepts of stochastic stabilizability (SS) and stochastic detectability (SD). These concepts are couched into the theory of operators in Banach space and, parallel to the classical linear quadratic (LQ) case, bound up with the spectrum of a certain infinite dimensional linear operator.

Key words. stochastic control, jump parameter, continuous-time, linear systems, infinite Markov chain

AMS subject classifications. 93E20, 93C05, 93C60, 60J75, 60J27

PII. S0363012900367485

1. Introduction.

Our main concern in this paper is with the so-called class of continuous-time linear systems with Markovian jump parameters (LSMJP). The usual infinite-time horizon quadratic cost is considered, and we assume that both the state and the Markov jump are accessible to the controller. (It is perhaps worth mentioning that although in engineering problems the Markov state is not often at hand, there are enough cases where this indeed happens. An illustrative list of such situations is found in [12].) Recent advances in LSMJP have greatly increased its power and led to new applications in many different fields. Potential applications include, inter alia, safety-critical and high-integrity systems (e.g., aircraft, chemical plants, nuclear power stations, robotic manipulator systems, large-scale flexible structures for space stations such as antennae, solar arrays, etc.). Without any intention of being exhaustive here, we mention [6], [7], [10], [12], [13], [14], [16], [17], [18], [19], [20], [22], [26], [27], [28], [33], [34], [35], [39], [40], [41], [42], [46], and [49], as a small sample of works dealing with different aspects of control problems. We mention also [5], [17], [30], [38], [40] (and references therein), and [46] as works dealing with applications of this class. In addition, the connection between linear dynamically varying (LDV) systems and jump linear systems, which has been exploited in [8] and [9], will certainly give a new impetus to LSMJP in the coming years. LDV controllers have been introduced as a technique to control systems with complicated dynamics (nonlinear dynamical

*Received by the editors February 7, 2000; accepted for publication (in revised form) January 25, 2001; published electronically June 26, 2001. Research for this paper was supported in part by the CNPq, under grants 520169/97.2 and 46.5532/2000-4, and by PRONEX. A preliminary version of this manuscript was presented as a regular paper at the 38th IEEE Conference on Decision and Control, Phoenix, AZ, 1999, pp. 4131–4136.

http://www.siam.org/journals/sicon/40-1/36748.html

†National Laboratory for Scientific Computing—LNCC/CNPq, Av. Getulio Vargas 333, Petrópolis, Rio de Janeiro, CEP 25651-070, Brazil (frag@lncc.br).

‡Federal University of Rio de Janeiro—UFRJ/COPPE, Systems and Computing Engineering Program, Bloco H, Ilha do Fundão, Rio de Janeiro/RJ, CEP 21945-970, Brazil (jack@iota.lncc.br).

systems that run over compact sets and have such features as nontrivial recurrence and periodic and aperiodic orbits).

What essentially distinguishes our problem from previous ones is that the Markov chain takes values in a countably infinite set. This, associated with the continuous-time feature of the problem, requires the use of operator theory, particularly, powerful tools from semigroup theory. Semigroup theory laid the basis for establishing here the equivalence between the conditions for stochastic stabilizability (SS) and stochastic detectability (SD) and the spectrum of a certain infinite dimensional linear operator. Operator theory was essential to allowing us to frame the whole problem into an infinite dimensional Banach space setting that lead us to the optimal solution which, analogously to the classical LQ case (see for instance [49]), boils down essentially to questions of existence and uniqueness of solution to a certain countably infinite set of coupled algebraic Riccati equations (ICARE).

In its content, this paper is closely related to [12]. However, technically they are rather different. For instance, beside the tools mentioned above, we have to introduce a natural adaptation of the decomplexification concept described in [1, section 18] for nonlinear complex functions with range in $\mathbb{R}$. This is required to establish a certain version of the gradient concept and, from that, the linear approximation to nonholomorphic functionals. This has allowed us to conveniently specify the semigroup of the Markov process $\{x, \theta\}$ applied to a certain (nonholomorphic) quadratic functional with domain in the complex space $\mathbb{C}^n$. In this way, we have preserved a more general (complex) framework to the problem, as in [12]. It is noteworthy that this framework is not only important for the sake of generalization, but because it allows us a self-adjoint matrix decomposition (refer to Remarks 2.3 and 2.5), as it is effectively required in the proof of the important Lemma 6.6. In addition, it is tacitly usual to work in the complex setting when using an operator theoretical approach.

Of course, it is too early to predict the full extent to which the theory developed here will be applied. However, from the outset we can envisage some situations in which model (3.1) can be naturally applied. For instance, applications to a nonlinear plant for which there is a conceivably infinite countable number of operating points, each of them characterized by a corresponding linearized model, where the abrupt changes would represent the dynamics of the system moving from one operating point to another, are suggested in [12]. This could probably also happen in economics and finance, where the complexity of the system, including the fact that the future is uncertain, is such that you have to consider infinitely many conceivable economical scenarios in order to have a more accurate model. ([3] gives practical-oriented-motivation for the use of infinite dimensional analysis in economics, and [17] gives practical-oriented-motivation for the use of LSJMP in economics.)

Finally, it is worth mentioning that if we specialize our results to the setting in which the state space of the Markov chain is finite, this paper still provides an important contribution in that the conditions in Theorem 6.13 below can be seen as a relaxation of those in Theorem 5 of [33], i.e., Theorem 5 of [33] uses the concept of observability, while we use the concept of SD. Indeed, if we consider the single state case (no jumps), SD reduces to detectability in the usual sense (see Remark 6.14), while the observability concept used in Theorem 5 of [33] (see also Definition 3 of [33]) reduces to observability in the usual sense.

An outline of the content of this paper is as follows. In section 2 we provide the bare essentials of notations and fundamental remarks. The model description is stated in section 3. Some preliminaries are given in section 4. The bulk of the results

in control are exhibited in sections 5 and 6.

2. Notations and initial remarks. As usual, $\mathbb{C}^n$ (respectively, $\mathbb{R}^n$) stands for the n-dimensional Euclidean space over the field of complex (respectively, real) numbers $\mathbb{C}$ (respectively, $\mathbb{R}$) and $\mathbb{N} = \{1, 2, \dots\}$. We denote $\mathbb{M}(\mathbb{C}^m, \mathbb{C}^n)$ as the normed linear space of all n by m complex matrices and, for simplicity, write $\mathbb{M}(\mathbb{C}^n)$ whenever $n = m$. We use the superscripts $^-$, $'$, and * for complex conjugate, transpose, and conjugate transpose, respectively. The notation $L \geq 0$ and $L > 0$ is adopted if a self-adjoint matrix is positive semidefinite or positive definite, respectively. We denote $\mathbb{M}(\mathbb{C}^n)^+ = \{L \in \mathbb{M}(\mathbb{C}^n); L = L^* \geq 0\}$ and write $\|\cdot\|_L$ for the norm in $\mathbb{C}^n$ induced by the inner product $\langle x, y\rangle_L = x^*Ly$ whenever the matrix $L = L^* \geq 0$. Furthermore, $\|\cdot\|_Y$ indicates a norm in the space Y. Except when otherwise mentioned, $\|\cdot\|$ represents either the Euclidean norm in $\mathbb{C}^n$ or the spectral induced norm in $\mathbb{M}(\mathbb{C}^n)$. To avoid notational confusion with the summation index i and j, we denote by ι the pure imaginary complex number. For $z \in \mathbb{C}$, we write z_{Re}(and sometimes $\mathrm{Re}(z)$) and z_{Im} for the real and imaginary parts of z, respectively, so that $z = z_{\mathrm{Re}} + \iota z_{\mathrm{Im}}$. For $x \in \mathbb{C}^n$ we denote the real vectors $x_{\mathrm{Re}} := (x_{1\,\mathrm{Re}}, \dots, x_{n\,\mathrm{Re}})'$ and $x_{\mathrm{Im}} := (x_{1\,\mathrm{Im}}, \dots, x_{n\,\mathrm{Im}})'$, which we call the real and imaginary parts of $x \in \mathbb{C}^n$, and we may write $x = x_{\mathrm{Re}} + \iota x_{\mathrm{Im}}$. (The notation $x_{j\,\mathrm{Re}}$ and $x_{j\,\mathrm{Im}}$ abbreviates the more precise notation $(x_j)_{\mathrm{Re}}$ and $(x_j)_{\mathrm{Im}}$, respectively, $j = 1, \dots, n$.) In addition, by the decomplexification of an arbitrary $x \in \mathbb{C}^n$, we mean the operation $\mathbb{C}^n \ni x \mapsto {}^{\mathrm{R}}x = (x_{\mathrm{Re}}\ x_{\mathrm{Im}})' \in \mathbb{R}^{2n}$, and by the decomplexification of a generic operator $g : [0, \infty) \times \mathbb{C}^n \mapsto \mathbb{R}$, we mean the operator ${}^{\mathrm{R}}g : [0, \infty) \times \mathbb{R}^{2n} \mapsto \mathbb{R}$ that coincides with g pointwise, i.e.,

$$ {}^{\mathrm{R}}g\left(t, {}^{\mathrm{R}}x\right) = g(t, x) \text{ for all } t \in [0, \infty) \text{ and } x \in \mathbb{C}^n, \tag{2.1} $$

which is a natural adaptation of the concept devised in [1].

Remark 2.1. From the above definition, ${}^{\mathrm{R}}\mathbb{C}^n = \mathbb{R}^{2n}$, ${}^{\mathrm{R}}(x + y) = {}^{\mathrm{R}}x + {}^{\mathrm{R}}y$, ${}^{\mathrm{R}}(cx) = c\,{}^{\mathrm{R}}(x)$, and $\|x\| = \left\|{}^{\mathrm{R}}x\right\|$ for $x, y \in \mathbb{C}^n$, c a real number.

Remark 2.2. For every $L \in \mathbb{M}(\mathbb{C}^n)^+$, there is a unique $L^{1/2} \in \mathbb{M}(\mathbb{C}^n)^+$ such that $(L^{1/2})^2 = L$. The absolute value of $L \in \mathbb{M}(\mathbb{C}^n)$, denoted by $|L|$, is defined as $|L| = (L^*L)^{1/2}$. It is easy to verify that $\|L\| = \||L|\|$.

Remark 2.3. Every element in $\mathbb{M}(\mathbb{C}^n)$ has a Cartesian self-adjoint decomposition (see, e.g., [44, p. 376]) and every self-adjoint operator in $\mathbb{M}(\mathbb{C}^n)$ can be decomposed in positive and negative parts [44, p. 464]. Thus, for any $L \in \mathbb{M}(\mathbb{C}^n)$, there exist X^+, X^-, Y^+, Y^- in $\mathbb{M}(\mathbb{C}^n)^+$ such that $L = (X^+ - X^-) + \iota\,(Y^+ - Y^-)$. Moreover, $X^+ \leq X^+ + X^- = (L + L^*)/2$, and thus $\|X^+\| \leq \|L\|$. Similarly, $\|X^-\| \leq \|L\|$, $\|Y^+\| \leq \|L\|$, and $\|Y^-\| \leq \|L\|$.

Set $\mathcal{H}_1^{m,n}$ (respectively, $\mathcal{H}_\infty^{m,n}$) as the linear space made up of all infinite sequences of complex matrices $H = (H_1, H_2, \dots)$, $H_i \in \mathbb{M}(\mathbb{C}^m, \mathbb{C}^n)$, such that $\sum_{i=1}^\infty \|H_i\| < \infty$ (respectively, $\sup\{\|H_i\|,\ i = 1, 2, \dots\} < \infty$). For $H \in \mathcal{H}_1^{m,n}$ (respectively, $H \in \mathcal{H}_\infty^{m,n}$) we define a norm in $\mathcal{H}_1^{m,n}$ (respectively, $\mathcal{H}_\infty^{m,n}$) by $\|H\|_1 = \sum_{i=1}^\infty \|H_i\|$ (respectively, $\|H\|_\infty = \sup\{\|H_i\|,\ i = 1, 2, \dots\}$). We shall write $\mathcal{H}_1^n$ and $\mathcal{H}_\infty^n$ whenever $n = m$ and denote $\mathcal{H}_1^{n^+} = \{H \in \mathcal{H}_1^n,\ H_i \in \mathbb{M}(\mathbb{C}^n)^+,\ i = 1, 2, \dots\}$ and $\mathcal{H}_\infty^{n^+} = \{H \in \mathcal{H}_\infty^n,\ H_i \in \mathbb{M}(\mathbb{C}^n)^+,\ i = 1, 2, \dots\}$ as the class of positive semidefinite elements of $\mathcal{H}_1^n$ and $\mathcal{H}_\infty^n$, respectively. For $H = (H_1, H_2, \dots)$ and $L = (L_1, L_2, \dots)$ in $\mathcal{H}_1^{n^+}$ we shall use the notation $H \leq L$ to indicate that $H_i \leq L_i$ for each i in $\mathbb{N}$. It is clear that

$$ H \leq L \Rightarrow \|H\|_1 \leq \|L\|_1 . \tag{2.2} $$

Furthermore, we shall use the notation H^* to indicate that each component H_i^* of H^* is the adjoint of H_i, $i = 1, 2, \dots$.

We shall denote $(l_1, \|\cdot\|_1)$, $(l_2, \|\cdot\|_2)$, and $(l_\infty, \|\cdot\|_\infty)$, respectively, as the sets made up of all infinite sequences of complex numbers $x = (x_1, x_2, \dots)$ such that $\sum_{i=1}^{\infty} |x_i| < \infty$, $\sum_{i=1}^{\infty} |x_i|^2 < \infty$, and $\sup\{|x_i|,\ i = 1, 2, \dots\} < \infty$, equipped with the usual norm $\|x\|_1 = \sum_{i=1}^{\infty} |x_i|$, $\|x\|_2 = \sum_{i=1}^{\infty} |x_i|^2$, and $\|x\|_\infty = \sup\{|x_i|,\ i = 1, 2, \dots\}$ and, in the case of $(l_2, \|\cdot\|_2)$, equipped with the usual internal product $\langle \cdot, \cdot \rangle$.

Remark 2.4. It is easy to verify that $(\mathcal{H}_1^{m,n}, \|\cdot\|_1)$ and $(l_1, \|\cdot\|_1)$ are uniformly homeomorphic. Similarly, $(\mathcal{H}_\infty^{m,n}, \|\cdot\|_\infty)$ and $(l_\infty, \|\cdot\|_\infty)$ can be shown to be uniformly homeomorphic. Since $(l_1, \|\cdot\|_1)$ and $(l_\infty, \|\cdot\|_\infty)$ are Banach spaces, we have that $(\mathcal{H}_1^{m,n}, \|\cdot\|_1)$ and $(\mathcal{H}_\infty^{m,n}, \|\cdot\|_\infty)$ are also Banach spaces.

Remark 2.5. Consider $Q = (Q_1, Q_2, \dots) \in \mathcal{H}_1^n$. From Remark 2.3, $Q_i = \left(X_i^+ - X_i^-\right) + \iota \left(Y_i^+ - Y_i^-\right)$, where X_i^+, X_i^-, Y_i^+, and Y_i^- belong to $\mathbb{M}(\mathbb{C}^n)^+$. Now define $X^+ = (X_1^+, X_2^+, \dots)$, $X^- = (X_1^-, X_2^-, \dots)$, $Y^+ = (Y_1^+, Y_2^+, \dots)$, and $Y^- = (Y_1^-, Y_2^-, \dots)$. Since $Q \in \mathcal{H}_1^n$, it follows, again from Remark 2.3, that X^+, X^-, Y^+, and Y^- also belong to $\mathcal{H}_1^n$. Therefore, Q can always be decomposed as

$$Q = \left(X^+ - X^-\right) + \iota \left(Y^+ - Y^-\right)$$

with X^+, X^-, Y^+, and Y^- in $\mathcal{H}_1^{n^+}$.

To support the sketch of the proof of Proposition 4.9, we define $\mathcal{W}_\infty^{m,n}$ as the Banach space made up of all infinite dimensional complex matrices of the type $\mathcal{C} = \operatorname{diag}(C_i)$, where $C_i \in \mathbb{M}(\mathbb{C}^m, \mathbb{C}^n)$, $i \in \mathcal{S}$, and $\sup_{i \in \mathcal{S}} \|C_i\| < \infty$. For $\mathcal{C} \in \mathcal{W}_\infty^{m,n}$ we define $\|\mathcal{C}\|_{\mathcal{W}_\infty} = \sup_{i \in \mathcal{S}} \|C_i\|$ as the norm in $\mathcal{W}_\infty^{m,n}$ and $\mathcal{C}^* = \operatorname{diag}(C_i^*) \in \mathcal{W}_\infty^{n,m}$, where C_i^* is the adjoint of C_i. We write $\mathcal{W}_\infty^n$ whenever $n = m$ and denote $\mathcal{W}_\infty^{n^+} = \{\mathcal{C} \in \mathcal{W}_\infty^n, C_i \in \mathbb{M}(\mathbb{C}^n)^+,\ i \in \mathcal{S}\}$ as the class of positive semidefinite elements of $\mathcal{W}_\infty^n$. For $\mathcal{B}, \mathcal{C} \in \mathcal{W}_\infty^{n+}$, we say that $\mathcal{B} \le \mathcal{C}$ if $B_i \le C_i$, $i \in \mathcal{S}$.

For any complex Banach space X, we denote by $Blt(X)$ the Banach space of all bounded linear transformations of X into X equipped with the uniform induced norm represented by $\|\cdot\|$, and for $L \in Blt(X)$ we denote by $\sigma(L)$ the spectrum of L.

Finally, we denote by $1_A\ \{.\}$ the Dirac measure, we write $\{\eta\}$ for any process $\{\eta(t),\ 0 \le t \le T\}$, whenever it is clear whether T is finite or not, and we adopt $E[\cdot]$ for the usual expectation. In addition, a function $f : Y \to \mathbb{R}$, Y a finite dimensional space, is denoted $o(\|r\|)$ if $\lim_{r \to 0} \frac{f(r)}{\|r\|_Y} = 0$ with r approaching zero by any path in Y. A function $f : [0, \infty) \to \mathbb{E}$, $\mathbb{E}$ standing for $\mathbb{R}$ or $\mathbb{C}$, is said to be $o(\delta)$ if $\lim_{\delta \downarrow 0} \frac{|f(\delta)|}{\delta} = 0$. A similar notation, namely, $o^n(\delta)$ (respectively, $o^{nn}(\delta)$), stands for a vector (respectively, matrix) valued function if the above limit holds for each entry. For $g : \mathbb{R}^n \to \mathbb{R}$ with partial derivatives $\frac{\partial g(x)}{\partial x_i}$, $i = 1, \dots, n$, we denote by $\nabla_x g(x) = (\frac{\partial g(x)}{\partial x_1} \dots \frac{\partial g(x)}{\partial x_n})'$ the gradient of g.

3. Problem statement. Let us fix an underlying complete probability space $(\Omega, \mathcal{F}, \mathcal{P})$ and, for arbitrary $s, T \in [0, \infty)$, consider the class of stochastic differential equations

$$\dot{x}(t) = A_{\theta(t)} x(t) + B_{\theta(t)} u(t), \quad s < t < T, \tag{3.1}$$

$$x(s) = x_s, \theta(s) = \theta_s, \tag{3.2}$$

where $x(t) \in \mathbb{C}^n$ denotes the state vector and $u(t) \in \mathbb{C}^m$ the control input. The system parameters are functions of a homogeneous Markov process $\{\theta(t), t \in [s, T]\}$ with right continuous trajectories and an infinite countable state space which, for convenience, we assign to the set $\mathcal{S} = \{1, 2, \dots\}$. We assume that $\{\theta\}$ has a stationary standard

transition probability matrix function (see [36, p. 138]) $\{P_\tau(i,j)\}_{i,j\in\mathcal{S}}$ in that, for $0 \le \tau \le T-t$,

$$P_\tau(i,j) = \mathcal{P}\{\theta(t+\tau) = j \mid \theta(t) = i\} = \begin{cases} \lambda_{ij}\tau + o_{ij}(\tau), & i \neq j, \\ 1 + \lambda_{ii}\tau + o_{ii}(\tau), & i = j, \end{cases} \tag{3.3}$$

with infinitesimal matrix $\Lambda = [\lambda_{ij}]_{i,j\in\mathcal{S}}$, where $\lambda_{ij} \ge 0$ for $i \neq j$. The Markov process $\{\theta\}$ is conservative and stable in that

$$\sum_{j=1, j\neq i}^{\infty} \lambda_{ij} = -\lambda_{ii} \le c < \infty, \quad i = 1, 2, \dots, \tag{3.4}$$

where c does not depend on i. We assume that $\left\{A_{(.)}, B_{(.)}\right\}$ are such that for any $j \in \mathcal{S}$ and for $\theta(t) = j$, $A_{\theta(t)} = A_j$ and $B_{\theta(t)} = B_j$, with A_j, B_j being constant matrices in $\mathbb{M}(\mathbb{C}^n)$ and $\mathbb{M}(\mathbb{C}^m, \mathbb{C}^n)$, respectively. In addition, the parameters are supposed norm-bounded in that $A = (A_1, A_2, \dots) \in \mathcal{H}_\infty^n$ and $B = (B_1, B_2, \dots) \in \mathcal{H}_\infty^{m,n}$. We consider x_s a second order random variable (r.v.) which may depend on the r.v. θ_s, and we shall denote $\vartheta_s = \vartheta_s(x_s, \theta_s)$ as the joint initial distribution of x_s and θ_s. By its turn, we assume the r.v.'s $\theta(t+\tau)$ are conditionally independent of x_s, given $\theta(t)$, for $s \le t < T$, $0 < \tau < T - t$.

In order to tackle our main problem, we begin by studying the auxiliary finite-time control problem as defined below.

We assume that the class of admissible control policies, $\mathcal{U}^{s,T}$ ($\mathcal{U}^{0,T} \equiv \mathcal{U}^T$ for short), is the class of all Borel measurable functions $u : \{[s,T], \mathbb{C}^n, \mathcal{S}\} \to \mathbb{C}^m$ such that, for some constant c, which might depend on u, the following hold.

C1. For every z, $y \in \mathbb{C}^n$, $t \in [s,T]$, and each $i \in \mathcal{S}$,
   (a) $\|u(t,z,i) - u(t,y,i)\| \le c\,\|(z-y)\|$ (Lipschitz condition), and
   (b) $\|u(t,y,i)\| \le c(1 + \|y\|)$ (growth condition).

For starting time $0 \le s < T$, terminal cost condition $L \in \mathcal{H}_\infty^{n^+}$, and for each policy $u \in \mathcal{U}^{s,T}$, define the cost functional

$$\mathcal{J}_{[s,T],L}(\vartheta_s, u) = E\left[\int_s^T \left\|\mathcal{Q}^{1/2}x(t)\right\|^2 + \left\|\mathcal{R}^{1/2}u(t)\right\|^2 dt + x(T)^* L_{\theta(T)} x(T)\right], \tag{3.5}$$

where $x(t)$ is subject to (3.1), $\mathcal{R}, \mathcal{Q} \in \mathbb{M}(\mathbb{C}^n)^+$, and $\mathcal{R} > 0$. For the sake of simplicity, the cost matrices $\mathcal{Q}$ and $\mathcal{R}$ are jump independent. However, the results derived here carry over verbatim to the case $\mathcal{Q}_{\theta(t)}$ and $\mathcal{R}_{\theta(t)}$ when conveniently norm-bounded.

The finite-time optimal control problem consists in finding $\hat{u}^T \in \mathcal{U}^{s,T}$, which minimizes $\mathcal{J}_{[s,T],L}(\vartheta_s, u)$.

Our main problem consists then in analyzing the infinite-time control problem by considering the setup as above, where now $t \ge 0$ and $\mathcal{U} \equiv \mathcal{U}^{0,\infty}$ with the following additional conditions.

C2. Model (3.1) with $t \ge 0$ is mean square stable (MSS), i.e., $E[\|x(t)\|^2] \to 0$ as $t \to \infty$ for any distribution ϑ_0.

C3. The cost functional $\mathcal{J}(\vartheta_0, u)$, defined below, is finite for any distribution ϑ_0.

For each policy $u \in \mathcal{U}$, define the cost functional

$$\mathcal{J}(\vartheta_0, u) := E\left[\int_0^\infty \left(\left\|\mathcal{Q}^{1/2}x(t)\right\|^2 + \left\|\mathcal{R}^{1/2}u(t)\right\|^2\right) dt\right], \tag{3.6}$$

where $x(t)$ is subject to (3.1) with $t > 0$.

The infinite-time problem is then to derive an optimal control policy $\hat{u}$, within the class $\mathcal{U}$, that minimizes the above cost, i.e., such that

$$\mathcal{J}(\vartheta_0, \hat{u}) := \inf_{u \in \mathcal{U}} \mathcal{J}(\vartheta_0, u), \tag{3.7}$$

and to establish structural conditions under which the existence and uniqueness of such a solution are ensured.

Remark 3.1. Since $\{x(t), \theta(t)\}$ is a Markov process, (3.1) and (3.5) set a "Markov" problem, and therefore control policies of the form $u(t) = u(t, x(t), \theta(t))$ suffice vis-à-vis the more expanded class consisting of policies of the form $u(t) = u(t, \{x(s), \theta(s), s \le t\})$.

4. Preliminaries. The following propositions from semigroup theory are essential tools in this work (see, for instance, [45]).

PROPOSITION 4.1. *Let Y be a Banach space, and consider the homogeneous differential equation*

$$\begin{cases} \dot{y}(t) = \mathcal{A}y(t), & t > 0, \\ y(0) = y \end{cases} \tag{4.1}$$

with arbitrary initial data $y \in Y$, where $\mathcal{A}$ is a bounded linear operator defined on Y into Y. Then (4.1) *is satisfied by a unique Y-valued function, continuous for $t \ge 0$ and continuously differentiable for $t > 0$, given by*

$$t \to y(t) = T(t)y \in Y, \quad t \ge 0, \tag{4.2}$$

where $T(t) : Y \to Y$, $t \ge 0$, is the C_0-semigroup (actually a uniformly continuous semigroup) of bounded linear transformations generated by $\mathcal{A}$, its infinitesimal operator.

PROPOSITION 4.2. *Let Y be a Banach space, and, for finite $T > 0$, consider the inhomogeneous differential equation*

$$\begin{cases} \dot{y}(t) = \mathcal{A}y(t) + f(t), & 0 < t < T, \\ y(0) = y \end{cases} \tag{4.3}$$

with arbitrary initial data $y \in Y$, where $\mathcal{A}$ is a bounded linear operator defined from Y into Y and $f \in L_1([0,T], Y)$. Then, for every $y \in Y$, the initial value problem (4.3) *has at most one solution. If, for a certain $y \in Y$, it has a solution, it is given by*

$$y(t) = T(t)y + \int_0^t T(t-s)f(s)ds, \qquad t \in [0,T],$$

where $T(t)$ is the uniformly continuous semigroup as defined in the above proposition. Moreover, if $f \in L_1([0,T], Y)$ is continuously differentiable, then (4.3) *has, for every $y \in Y$, a unique continuous and continuously differentiable solution on $[0,T]$.*

The following lemma is, essentially, a combination of results from [45].

LEMMA 4.3. *Let $\mathcal{A} : \mathfrak{D}(\mathcal{A}) \to Y$ be the infinitesimal generator of a C_0 semigroup $T(t) : Y \to Y$, let Y be a Banach space, and let $\mathfrak{D}(\mathcal{A})$ be the domain of $\mathcal{A}$, and consider the following assertions.*

1. $\sup\{\operatorname{Re}\lambda : \lambda \in \sigma(\mathcal{A})\} < 0$.
2. *There are constants $M \ge 1$ and $\omega > 0$ such that $\|T(t)\| \le M \exp(-\omega t)$.*
3. $\int_0^\infty \|T(t)y\|\, dt < \infty$ *for every $y \in Y$.*

4. $\int_0^\infty \|T(t)\|\,dt < \infty$.

Then 2, 3, *and* 4 *are equivalent assertions and imply* 1. *Moreover, if* $T(t)$ *is analytic, all assertions are equivalent.*

Proof (sketch of proof). Implications ($2 \Rightarrow 4$) and ($4 \Rightarrow 3$) are straightforward and ($3 \Rightarrow 2$) is obtained, e.g., from Theorem 4.4.1 of [45]. For ($2 \Rightarrow 1$), note that the C_0 semigroup $S(t) = \exp(\omega t)T(t)$, $t \geq 0$, is such that $\|S(t)\| \leq M$ and so $\rho(\mathcal{A}_s) \supset \{\lambda \in \mathbb{C} : \operatorname{Re}\lambda > 0\}$, where $\rho(\mathcal{A}_s)$ is the resolvent set of the infinitesimal generator $\mathcal{A}_s$ of $S(t)$ (see, e.g., [45]). Since $\mathcal{A} = \mathcal{A}_s - \omega I$ and $\mathfrak{D}(\mathcal{A}) = \mathfrak{D}(\mathcal{A}_s)$, it follows that $\rho(\mathcal{A}) = \rho(\mathcal{A}_s) - \omega$ and so $\rho(\mathcal{A}) \supset \{\lambda \in \mathbb{C} : \operatorname{Re}\lambda > -\omega\}$, i.e., $\sup\{\operatorname{Re}\lambda : \lambda \in \sigma(\mathcal{A})\} < 0$. Finally, if $T(t)$ is an analytic semigroup, the exponential decay implication ($1 \Rightarrow 2$) follows, e.g., from Theorem 4.4.3 of [45]. For details, see [4]. □

Corollary 4.4. *If Assertion* 2 *of Lemma* 4.3 *holds, then* $\beta = \frac{M}{\omega}$ *(respectively,* $\beta\,\|y\|$*) is an upper bound for Assertion* 4 *(respectively, Assertion* 3*).*

For arbitrary initial condition (x_0, θ_0), let us now consider the homogeneous dynamic system

$$\begin{cases} \dot{x}(t) = F_{\theta(t)}x(t), \quad t > 0, \\ x(0) = x_0,\ \theta(0) = \theta_0, \end{cases} \tag{4.4}$$

where $F = (F_1, F_2, \dots) \in \mathcal{H}_\infty^n$, and define $Q(t) = (Q_1(t), Q_2(t), \dots)$, $t \geq 0$, where

$$Q_i(t) = E[x(t)x(t)^* 1_{\{\theta(t)=i\}}] \in \mathbb{M}(\mathbb{C}^n)^+, \quad i \in \mathcal{S}. \tag{4.5}$$

Furthermore, for $H = (H_1, H_2, \dots) \in \mathcal{H}_1^n$, let us define the operator $\mathcal{D}$, with $\mathcal{D}(H) = (\mathcal{D}_1(H), \mathcal{D}_2(H), \dots)$, such that

$$\mathcal{D}_i(H) = F_i H_i + H_i F_i^* + \sum_{j=1}^{\infty} \lambda_{ji} H_j, \quad i \in \mathcal{S}. \tag{4.6}$$

Proposition 4.5. $\mathcal{D} \in Blt\,(\mathcal{H}_1^n)$.

Proof. The proof follows standard arguments concerning bounded linear transformations in Banach spaces. □

Proposition 4.6. *Let* $x(t)$ *be given by* (4.4) *with arbitrary* $F \in \mathcal{H}_\infty^n$. *Then* $Q(t)$, $t \geq 0$, *defined as in* (4.5), *belongs to* $\mathcal{H}_1^{n^+}$ *and satisfies the Banach space linear differential equation*

$$\begin{cases} \dot{Q}(t) = \mathcal{D}(Q(t)), \quad t > 0, \\ Q(0) = Q^0 \in \mathcal{H}_1^{n^+}, Q_i^0 = E[x_0 x_0^* 1_{\{\theta_0=i\}}], \quad i \in \mathcal{S}, \end{cases} \tag{4.7}$$

or, equivalently, the infinite countable set of interconnected linear differential equations

$$\begin{cases} \dot{Q}_i(t) = \mathcal{D}_i(Q(t)), \quad t > 0, \\ Q_i(0) = Q_i^0 = E[x_0 x_0^* 1_{\{\theta_0=i\}}], \quad i \in \mathcal{S}, \end{cases} \tag{4.8}$$

where $\mathcal{D}$ *is given as in* (4.6).

Proof. We use (4.4) and (4.5) and the fact that $Q_i(t) \in \mathbb{M}(\mathbb{C}^n)^+$ to obtain that $Q(t) \in \mathcal{H}_1^{n^+}$, $t \geq 0$, and that, with probability one,

$$x(t+\delta) = x(t) + F_{\theta(t)}x(t)\delta + o^n(\delta), \quad t, \delta \geq 0. \tag{4.9}$$

Furthermore, some algebraic manipulation and standard results with respect to the existence of limits lead us to $\frac{d^+Q(t)}{dt} = \lim_{\delta \downarrow 0} \frac{Q(t+\delta)-Q(t)}{\delta} = \mathcal{D}(Q(t))$, $t > 0$. We notice further that $\mathcal{D}$ is the infinitesimal generator of a C_0-semigroup, and from an argument of the proof of Theorem 1.2.4 of [45], the proposition follows. ∎

Let us now define, for $H = (H_1, H_2, \dots) \in \mathcal{H}_\infty^{n^+}$, the linear operators $\mathcal{E}(H) = (\mathcal{E}_1(H), \mathcal{E}_2(H), \dots)$ and $\mathcal{G}(H) = (\mathcal{G}_1(H), \mathcal{G}_2(H), \dots)$ as well as the nonlinear operator $\mathcal{T}(H) = (\mathcal{T}_1(H), \mathcal{T}_2(H), \dots)$, where, for each $i \in \mathcal{S}$,

$$\mathcal{E}_i(H) = \sum_{j=1, j\neq i}^{\infty} \lambda_{ij} H_j, \quad \mathcal{G}_i(H) = \mathcal{R}^{-1} B_i^* H_i, \tag{4.10}$$

and

$$\mathcal{T}_i(H) = \mathcal{Q} + A_i^* H_i + H_i A_i - H_i B_i \mathcal{R}^{-1} B_i^* H_i + \lambda_{ii} H_i + \mathcal{E}_i(H). \tag{4.11}$$

PROPOSITION 4.7. *The operator $\mathcal{E}$ maps $\mathcal{H}_\infty^{n+}$ into $\mathcal{H}_\infty^{n^+}$, $\mathcal{T}$ maps $\mathcal{H}_\infty^{n+}$ into $\{Z \in \mathcal{H}_\infty^n : Z^* = Z\}$, and $\mathcal{G}$ maps $\mathcal{H}_\infty^{n^+}$ into $\mathcal{H}_\infty^{n,m}$.*

Proof. The proof is straightforward from standard arguments concerning normed spaces. ∎

With the operators $\mathcal{T}$, $\mathcal{E}$, and $\mathcal{G}$ in hand, let us now define, for finite $T$and $L \in \mathcal{H}_\infty^{n^+}$ arbitrarily fixed, the Banach space Riccati differential equation

$$\begin{cases} \dot{S}^T(t) + \mathcal{T}(S^T(t)) = 0, & t \in (0, T), \\ S^T(T) = L, \end{cases} \tag{4.12}$$

where $S^T(t) = (S_1^T(t), S_2^T(t), \dots)$.

Equation (4.12) may be written as the following infinite countable set of interconnected Riccati differential equations:

$$\begin{cases} \dot{S}_i^T(t) + \mathcal{T}_i(S^T(t)) = 0, & t \in (0, T), \\ S_i^T(T) = L_i, & i \in \mathcal{S}. \end{cases} \tag{4.13}$$

Remark 4.8. Although it may appear, prima facie, that the above equivalence is a tautology, we would like to point out that it relies on the fact that the induced norm by $\mathcal{H}_\infty^n$ on the linear subspace $\{(0, \dots, 0, H, 0, 0, \dots),\ H \in \mathbb{M}(\mathbb{C}^n)\}$ coincides with the usual norm $\|\cdot\|$ of $\mathbb{M}(\mathbb{C}^n)$ ((4.12) ⇒ (4.13)) and that $(\dot{S}_1^T(t), \dot{S}_2^T(t), \dots) \in \mathcal{H}_\infty^n$ ((4.13) ⇒ (4.12)). (For more information on the matter, see [43].) To see that $(\dot{S}_1^T(t), \dot{S}_2^T(t), \dots) \in \mathcal{H}_\infty^n$, note that for $S^T(t) \in \mathcal{H}_\infty^{n^+}$, we have from Proposition 4.7 and the definition of $\mathcal{T}$ that $\mathcal{T}(S^T(t)) = (\mathcal{T}_1(S^T(t)), \mathcal{T}_2(S^T(t)), \dots) \in \mathcal{H}_\infty^n$. In turn, this and (4.13) yield $(\dot{S}_1^T(t), \dot{S}_2^T(t), \dots) \in \mathcal{H}_\infty^n$.

The proposition that follows shows that the solution of the Riccati differential equation given by (4.12) exists and is unique.

PROPOSITION 4.9. *For $T \in [0, \infty)$ and terminal data $L \in \mathcal{H}_\infty^{n^+}$, arbitrarily fixed, there exists a solution $S^T(\cdot) : [0, T] \to \mathcal{H}_\infty^{n^+}$ for (4.12), continuous for $t \in [0, T]$ and continuously differentiable for $t \in (0, T)$. Moreover, this solution is unique (within the class of solutions with these properties).*

Proof (sketch of proof). Besides existence and uniqueness, we have to show that the solution to (4.12) is positive in the sense defined in section 2. This led us to an approach inspired, in part, by that in [49] for the finite dimensional scenario, in conjunction with standard results from the literature on semigroup, evolution equation,

and the Volterra equation in Banach space (see, e.g., [45], [37]). So, we define the infinite dimensional square matrix $\mathfrak{S}(t) = \text{diag}(S_i^T(t))$, $t \in [s,T]$, and shape (4.12) as

$$\dot{\mathfrak{S}}^T(t) + \Psi(\mathfrak{S}(t)) = 0, \qquad \mathfrak{S}(T) = \mathcal{L}, \ \ t \in (s,T), \tag{4.14}$$

where $\Psi(\mathfrak{S}) = \overline{\mathcal{A}}^* \mathfrak{S} + \mathfrak{S}\overline{\mathcal{A}} + \Pi(\mathfrak{S}) + \mathfrak{Q} - \mathfrak{S}\mathcal{B}\mathfrak{R}^{-1}\mathcal{B}^*\mathfrak{S}$, $\overline{\mathcal{A}} = \mathcal{A} + \frac{1}{2}\text{diag}(\lambda_{ii} I_n)$, $\mathcal{A} = \text{diag}(A_i) \in \mathcal{W}_\infty^n$, $\mathcal{B} = \text{diag}(B_i) \in \mathcal{W}_\infty^{m,n}$, $\mathfrak{R}^{-1} = \text{diag}(\mathcal{R}^{-1}) \in \mathcal{W}_\infty^{n^+}$, $\mathfrak{Q} = \text{diag}(\mathcal{Q}) \in \mathcal{W}_\infty^{m^+}$, and $\mathcal{L} = \text{diag}(L_i) \in \mathcal{W}_\infty^{n^+}$. In addition, $\Pi : \mathcal{W}_\infty^n \to \mathcal{W}_\infty^n$ is a bounded linear operator with $\Pi = \mathfrak{D} \circ \chi \circ \mathfrak{D}^{-1}$, where, for every $H \in \mathcal{H}_\infty^n$, $\chi = (\chi_1, \chi_2, \dots) : \mathcal{H}_\infty^n \to \mathcal{H}_\infty^n$ is such that $\chi_i(H) = \sum_{j=1, j\neq i}^{\infty} \lambda_{ij} H_j$, and $\mathfrak{D} : \mathcal{H}_\infty^n \to \mathcal{W}_\infty^n$ is such that $\mathfrak{D}(H) = \text{diag}(H_i)$. We proceed ensuring that every element in $\mathcal{W}_\infty^{:,:}$ is well defined as an operator in $Blt\,(\mathcal{W}_\infty^{:,:}, \mathcal{W}_\infty^{:,:})$, namely, that $\mathcal{C} \in \mathcal{W}_\infty^{m,n}$ implies $\mathcal{C} \in Blt\,(\mathcal{W}_\infty^{m,q}, \mathcal{W}_\infty^{q,n})$, $\|\mathcal{C}\| \leq \|\mathcal{C}\|_{\mathcal{W}_\infty}$, and $\|\mathcal{C}\| = \|\mathcal{C}\|_{\mathcal{W}_\infty}$ for $n = m = q$. Now, in the spirit of the technique of [49], we define a version of the differential equation (4.14), which is linear in $\mathfrak{S}$, and we obtain its Volterra equivalent. Picard's successive approximation method and the positiveness of Π (in that $\Pi(H) \geq 0$ if $H \geq 0$) give us that the unique solution to the Volterra equation is positive semidefinite, i.e., it belongs to $\mathcal{W}_\infty^{n^+}$. An important step now is to build a sequence of solutions $\mathfrak{S}_i(t)$ to the corresponding set of Volterra equations equipped with parameters $\mathcal{K}_1(t)$ (arbitrary) and $\mathcal{K}_i(t) = \mathfrak{R}^{-1}\mathcal{B}^*\mathfrak{S}_{i-1}(t)$, $i = 2, 3, \dots$, in that order. Then, exploring (i) a property of the minimum, (ii) a comparison theorem, and (iii) a standard result on semigroup theory, and applying the dominated convergence theorem, we show that, for each t, $\{\mathfrak{S}_i(t)\}$ is a monotone nonincreasing sequence of positive semidefinite elements that converges to a solution of (4.14). A Lipschitz condition gives us that this positive solution is unique. For details, see [4], [27]. □

5. The finite-time optimal control problem. Referring to the finite-time optimal control problem defined in section 3, we have from system (3.1) that $\{x(t), \theta(t)\}_{t\in[s,T]}$, is a Markov process evolving in $(\mathbb{C}^n, \mathcal{S})$ with sample paths that are continuous from the right. From this fact, and bearing in mind an argument from [1, p. 37], we have that $\{x(t), \theta(t)\}_{t\in[s,T]}$ has a stochastically continuous transition probability and consequently is characterized uniquely in terms of its infinitesimal generator, as follows. Let $\mathbb{B}(\mathcal{X}, \mathbb{R})$, $\mathcal{X} = ([s,T] \times \mathbb{C}^n \times \mathcal{S})$, be the Banach space of all bounded real valued measurable functions g, defined on $\mathcal{X}$, equipped with the norm $\|g\| := \sup\{|g(z)| : z \in \mathcal{X}\}$. The semigroup of linear Markov transition operators $T(h) : \mathbb{B}(\mathcal{X}, \mathbb{R}) \to \mathbb{B}(\mathcal{X}, \mathbb{R})$, $h \in [0, T-t]$, which characterizes $\{x(t), \theta(t)\}_{t\in[s,T]}$, is given by

$$(T(h)g)\,(t, x(t), \theta(t)) := E_{x(t),\theta(t)}[g(t+h, x(t+h), \theta(t+h))] \tag{5.1}$$

for every $(t, x(t), \theta(t)) \in \mathcal{X}$.

By the infinitesimal generator of a family of transition probabilities of the Markov process $\{x(t), \theta(t)\}_{t\in[s,T]}$, we mean the operator $\mathcal{L} : \mathfrak{D}(\mathcal{L}) \mapsto \mathbb{B}(\mathcal{X}, \mathbb{R})$, such that

$$(\mathcal{L}g)\,(t, x(t), \theta(t)) = \lim_{h \downarrow 0} \frac{(T(h)g)\,(t, x(t), \theta(t)) - (T(0)g)\,(t, x(t), \theta(t))}{h} \tag{5.2}$$

for every $(t, x(t), \theta(t)) \in \mathcal{X}$ and $g \in \mathfrak{D}(\mathcal{L})$ with $(T(h)g)\,(t, x(t), \theta(t))$ defined as in (5.1), where $\mathfrak{D}(\mathcal{L})$ is the set of functions $g \in \mathbb{B}(\mathcal{X}, \mathbb{R})$ for which the above limit exists. The limit required is the uniform limit with respect to $\mathcal{X}$. Furthermore, Dynkin's

formula reads as (see [32])

$$\begin{aligned} &g(s,x(s),\theta(s)) - E_{x(s),\theta(s)}\left[g(t,x(t),\theta(t))\right] \\ &\qquad = E_{x(s),\theta(s)}\left[\int_s^t -\left(\mathcal{L}g\right)(r,x(r),\theta(r))dr\right] \end{aligned} \tag{5.3}$$

for $g \in \mathfrak{D}(\mathcal{L})$. (Notice that, since all terms above are measurable and bounded, the integral and expectations above exist and are finite.)

Bearing in mind now the decomplexification concept in section 2, we get, in our scenario, a further derivation for the infinitesimal generator of the Markov process $\{x(t),\theta(t)\}_{t\in[s,T]}$. For this, let us start with the following proposition.

PROPOSITION 5.1. *For any continuous bounded function $g{:}\mathcal{X} \mapsto \mathbb{R}$, we have that*

$$\lim_{h\downarrow 0}\left(T(h)g\right)(t,x(t),\theta(t)) = g(t,x(t),\theta(t)). \tag{5.4}$$

Proof. The proof follows, mutatis mutandis, from [29] . □

We now define $C_b^{1,\mathrm{R}}(\mathcal{X})$ as the set of all functions $g \in \mathbb{B}(\mathcal{X},\mathbb{R})$ such that, for each $i \in \mathcal{S}$, the decomplexification ${}^{\mathrm{R}}g$ is (Fréchet-) continuously differentiable in the variables $t \in (s,T)$ and ${}^{\mathrm{R}}x \in \mathbb{R}^{2n}$.

PROPOSITION 5.2. *Consider system* (3.1) *with $u \in \mathcal{U}^{s,T}$, and let $g \in C_b^{1,\mathrm{R}}(\mathcal{X})$. Then $g \in \mathfrak{D}(\mathcal{L})$, and the infinitesimal operator* (5.2) *reads as*

$$\begin{aligned} \left(\mathcal{L}^u g\right)(t,x(t),\theta(t)) &= \frac{\partial}{\partial t}g(t,x(t),\theta(t)) + \nabla_{{}^{\mathrm{R}}x}\,{}^{\mathrm{R}}g(t,{}^{\mathrm{R}}x(t),\theta(t))'\,{}^{\mathrm{R}}(A_{\theta(t)}x(t) \\ &\quad + B_{\theta(t)}u(t)) + \sum_{j=1}^{\infty} g(t,x(t),j)\lambda_{\theta(t)j} \end{aligned} \tag{5.5}$$

for any $(t,x(t),\theta(t)) \in ((s,T)\times\mathbb{C}^n\times\mathcal{S})$.

Proof. The proof follows from (5.2), bearing in mind (3.1), (3.3), (5.1), and Lemma 7.1. That $g \in \mathfrak{D}(\mathcal{L})$ follows along the same lines of [21, p. 159]. □

Remark 5.3. Actually, $\mathfrak{D}(\mathcal{L})$ now is the set $C_b^{1,\mathrm{R}}(\mathcal{X})$ (see [1, p. 38]).

With $\mathcal{X}_o = ((s,T)\times X_o\times\mathcal{S})$ and X_o an arbitrary open and bounded set in $\mathbb{C}^n$, let us define $C^{1,\mathrm{R}}(\mathcal{X}_o)$ as the set of all real valued measurable functions g, well-defined on $\bar{\mathcal{X}}_o$ (the closure of $\mathcal{X}_o$), such that, for each $i \in \mathcal{S}$, the decomplexification ${}^{R}g$ is (Fréchet-) continuously differentiable in the variables $t \in (s,T)$ and ${}^{\mathrm{R}}x \in {}^{\mathrm{R}}X_o$. In this case, mutatis mutandis, as in [23, Ch. V, Lemma 5.1], the integral and expectations in the above equations exist and are finite, so that Dynkin's formula (5.3) may be applied. This is the setting of our next proposition.

PROPOSITION 5.4. *Let $g \in C^{1,\mathrm{R}}(\mathcal{X}_o)$ be such that*

$$g(t,x,i) = x^* S_i^T(t)x, \tag{5.6}$$

where $t \mapsto S^T(t) = (S_1^T(t), S_2^T(t)\dots) \in \mathcal{H}_\infty^{n^+}$ satisfies the Banach space Riccati differential equation given by (4.12) *with terminal condition $S^T(T) = L \in \mathcal{H}_\infty^{n^+}$. Then, for system* (3.1) *with $u \in \mathcal{U}^{s,T}$, the infinitesimal operator $\mathcal{L}^u$ is given by*

$$\begin{aligned} \left(\mathcal{L}^u g\right)(t,x(t),\theta(t)) &= x(t)^*\{-\mathcal{Q} + S_{\theta(t)}^T(t)B_{\theta(t)}\mathcal{R}^{-1}B_{\theta(t)}^* S_{\theta(t)}^T(t)\}x(t) \\ &\quad + u(t)^* B_{\theta(t)}^* S_{\theta(t)}^T(t)x(t) + x(t)^* S_{\theta(t)}^T(t)B_{\theta(t)}u(t) \end{aligned} \tag{5.7}$$

for any $(t, x(t), \theta(t)) \in ((s,T) \times X_o \times \mathcal{S})$. *Furthermore, Dynkin's formula* (5.3) *reads as*

$$\begin{aligned} & x(s)^* S^T_{\theta(s)}(s)x(s) - E_{x(s),\theta(s)}[x(t)^* S^T_{\theta(t)}(t)x(t)] \\ & = E_{x(s),\theta(s)} \left[\int_s^t x(r)^* \mathcal{Q} x(r) - x(r)^* S^T_{\theta(r)}(r) B_{\theta(r)} \mathcal{R}^{-1} B^*_{\theta(r)} S^T_{\theta(r)}(r) x(r) \right. \\ & \qquad \left. - u(r)^* B^*_{\theta(r)} S^T_{\theta(r)}(r) x(r) - x(r)^* S^T_{\theta(r)}(r) B_{\theta(r)} u(r) dr \right]. \end{aligned} \tag{5.8}$$

Proof. Bearing in mind [23, Chap. V, Lemma 5.1], we use (5.5) as well as Lemma 7.4. □

PROPOSITION 5.5. *The above result also holds if we consider Proposition* 5.4 *with* g *now defined on the hole domain* $\mathcal{X}$.

Proof. Note that g trivially satisfies a polynomial growth condition, namely, $\|g(t,x,i)\| \le c_1(1 + \|x\|^k)$ for every $(t,x,i) \in \mathcal{X}$ and some constants c_1 and k. Also note that g with $S^T(t)$ given by (4.12) is continuous on $\bar{\mathcal{X}}_o$. These facts, together with Lemma 7.7 in the Appendix, fulfill the conditions given in [23, Ch. V, Theorem 5.1] so that, along the same lines as in the proof of this theorem, Dynkin's formula (5.3) may still be applied (the integral and expectations therein exist) and the proposition follows. □

We now derive a cost expression for an arbitrary $u \in \mathcal{U}^{s,T}$ and the optimal solution for the finite-time case.

PROPOSITION 5.6. *For arbitrary* $u \in \mathcal{U}^{s,T}$, *the cost defined in* (3.5) *reads as follows.*

$$\begin{aligned} \mathcal{J}_{[s,T],L}(\vartheta_s, u) = E \Bigg[& x(s)^* S^T_{\theta(s)}(s) x(s) \\ & + \int_s^T \left\| B^*_{\theta(r)} S^T_{\theta(r)}(r) x(r) + \mathcal{R} u(r) \right\|^2_{\mathcal{R}^{-1}} dr \Bigg] \end{aligned} \tag{5.9}$$

with $S^T(r) \in \mathcal{H}^{n^+}_\infty$ *(uniquely) satisfying* (4.12).

Proof. From (3.5), we have that

$$\begin{aligned} \mathcal{J}_{[s,T],L}(\vartheta_s, u) = E \Bigg[& E_{x(s),\theta(s)} \left[\int_s^T (x(r)^* \mathcal{Q} x(r) + u(r)^* \mathcal{R} u(r)) dr \right] \\ & + E_{x(s),\theta(s)}[x(T)^* L_{\theta(T)} x(T)] \Bigg]. \end{aligned} \tag{5.10}$$

Now, from Proposition 5.5, setting $t = T$ in Dynkin's formula (5.8), we get

$$\begin{aligned} \mathcal{J}_{[s,T],L}(\vartheta_s, u) = E \Bigg[& x(s)^* S^T_{\theta(s)}(s) x(s) + E_{x(s),\theta(s)} \left[\int_s^T (u(r)^* \mathcal{R} u(r) \right. \\ & + x(r)^* S^T_{\theta(r)}(r) B_{\theta(r)} \mathcal{R}^{-1} B^*_{\theta(r)} S^T_{\theta(r)}(r) x(r) \\ & \left. + u(r)^* B^*_{\theta(r)} S^T_{\theta(r)}(r) x(r) + x(r)^* S^T_{\theta(r)}(r) B_{\theta(r)} u(r)) dr \right] \Bigg]. \end{aligned} \tag{5.11}$$

Now, since $\mathcal{R} = \mathcal{R}^*$, the expression under integration may be written as

$$u(r)^*\mathcal{R}^*\mathcal{R}^{-1}\mathcal{R}u(r) + x(r)^*S^T_{\theta(r)}(r)B_{\theta(r)}\mathcal{R}^{-1}B^*_{\theta(r)}S^T_{\theta(r)}(r)x(r)$$
$$+ (\mathcal{R}u(r))^*\mathcal{R}^{-1}B^*_{\theta(r)}S^T_{\theta(t)}(r)x(r) + x(r)^*S^T_{\theta(r)}(r)B_{\theta(r)}\mathcal{R}^{-1}\mathcal{R}u(r),$$

and, denoting $y = B^*_{\theta(r)}S^T_{\theta(r)}(r)x(r)$ and $w = \mathcal{R}u(r)$, it becomes

$$w^*\mathcal{R}^{-1}w + y^*\mathcal{R}^{-1}y + w^*\mathcal{R}^{-1}y + y^*\mathcal{R}^{-1}w = (y+w)^*\mathcal{R}^{-1}(y+w)$$
$$= \|y+w\|^2_{\mathcal{R}^{-1}} = \left\|B^*_{\theta(r)}S^T_{\theta(r)}(r)x(r) + \mathcal{R}u(r)\right\|^2_{\mathcal{R}^{-1}}.$$

Thus substitution in (5.11) yields

$$\mathcal{J}_{[s,T],L}(\vartheta_s, u) = E\left[x(s)^*S^T_{\theta(s)}(s)x(s) + \int_s^T \left\|B^*_{\theta(r)}S^T_{\theta(r)}(r)x(r) + \mathcal{R}u(r)\right\|^2_{\mathcal{R}^{-1}} dr\right],$$

(5.12)

which completes the proof.

Remark 5.7. Note that, instead of Proposition 5.5, we could also use Proposition 5.4 alone to deduce (5.11), bearing in mind that the process $\{x\}$ satisfies the differential equation (3.1), with $u \in \mathcal{U}^{s,T}$, and $E[\|x(r)\|^k]$ is bounded for each $k > 0$ and $s \le r \le T$ (see [23, p. 156]). □

PROPOSITION 5.8. *The optimal control in the admissible class $\mathcal{U}^{s,T}$ is given by*

$$\hat{u}^T(t) = -G^T_{\theta(t)}(t)x(t), \tag{5.13}$$

where $G^T(t) = (G^T_1(t), G^T_2(t), \dots) = \mathcal{G}(S^T(t)) \in \mathcal{H}^{n,m}_\infty$ ($G^T_i(t) = \mathcal{R}^{-1}B^*_iS^T_i(t)$) with $S^T(t) \in \mathcal{H}^{n^+}_\infty$ (uniquely) satisfying (4.12), *$t \in [s,T]$. Furthermore, the minimum cost reads as follows:*

$$\mathcal{J}_{[s,T],L}(\vartheta_s, \hat{u}^T) = \inf_{u\in\mathcal{U}^{s,T}} \mathcal{J}_{[s,T],L}(\vartheta_s, u) = E[x(s)^*S^T_{\theta(s)}(s)x(s)]. \tag{5.14}$$

Proof. The proof is immediate from (5.12). □

6. The infinite-time optimal control problem. In this section conditions for solving the infinite horizon optimal control problem are established. Parallel to the classical LQ problem, when dealing with the infinite-time optimal control problem(infinite horizon) two structural concepts turn out to be essential: SS and SD, defined as follows.

DEFINITION 6.1 (SS). *We say that the system (A, B, Λ) is SS if there exists $G \in \mathcal{H}^{n,m}_\infty$ such that for any joint initial distribution ϑ_0, we have that*

$$\int_0^\infty E[\|x(t)\|^2]dt < \infty, \tag{6.1}$$

where $x(t)$ is given by (3.1) *with $t \ge 0$ and $u(t) = -G_{\theta(t)}x(t)$, i.e.,*

$$\dot{x}(t) = F_{\theta(t)}x(t), \quad t > 0, \tag{6.2}$$

with $F_{\theta(t)} = A_{\theta(t)} - B_{\theta(t)}G_{\theta(t)}$. In this case we say that (6.2) *is stochastically stable and G stabilizes (A, B, Λ).*

DEFINITION 6.2 (SD). *Consider* $C = (C_1, C_2, \dots) \in \mathcal{H}_\infty^{n,r}$. *We say that the system* (C, A, Λ) *is SD if there exists* $K \in \mathcal{H}_\infty^{r,n}$ *such that for any joint initial distribution* ϑ_0, *we have that*

$$\int_0^\infty E[\|x(t)\|^2]dt < \infty, \tag{6.3}$$

where $x(t)$ *is given by*

$$\dot{x}(t) = F_{\theta(t)}x(t), \quad t > 0, \tag{6.4}$$

with $F_{\theta(t)} = A_{\theta(t)} - K_{\theta(t)}C_{\theta(t)}$.

Remark 6.3. The system (C, A, Λ) refers to

$$\begin{cases} \dot{r}(t) = A_{\theta(t)}r(t), \quad t > 0, \\ y(t) = C_{\theta(t)}r(t), \\ r(0) = x_0,\ \theta(0) = \theta_0, \end{cases} \tag{6.5}$$

and $x(t)$, given by (6.4), assigns the error of a K-based estimate of $r(t)$. We say then that this estimate detects $r(\cdot)$ in the sense of (6.3) and that K turns (C, A, Λ) detectable.

Remark 6.4 (SS versus MSS.). It has been shown in [22] that MSS and SS are equivalent if the Markov chain has a finite state space. For the countably infinite case, however, this equivalence is no longer true, as we can notice from the counterexample that follows. We denote $-b_{\theta(t)} = A_{\theta(t)}$ and consider the infinite-time scalar version of the random differential equation (3.1) with $u \equiv 0$, $\{\theta\}$ a Poisson process with parameter λ, and $b_i = \frac{\lambda}{2}\ln\left(\frac{i+1}{i}\right)$, $i \in \mathcal{S}$. In this case (see [11], [48]) the discontinuities of the sample paths are ordinary jumps with probability one, the sequence of jump times $\tau_0 = 0 < \tau_1 < \tau_2 \dots$ is infinite and such that $\lim_{n\to\infty}\tau_n = \infty$ almost surely (a.s.), and the sojourn times $\tau_n - \tau_{n-1}$, $n \in \mathbb{N}$, are independent r.v.'s with average $E[\tau_n - \tau_{n-1}] = 1/\lambda$ and density function given by $\lambda e^{-\lambda s}$, $s \geq 0$. Furthermore, the trajectories of the state process $\{x\}$ are decreasing and connected solution pieces of (3.1) given by $x(t) = a_n \exp(\ -b_{n+\theta_0-1}(\ t - \tau_{n-1}))$ a.s., $\tau_{n-1} \leq t < \tau_n$, $n \in \mathbb{N}$, where $a_1 = x_0$, and, from continuity, $a_n = x_0 \exp(\ -\sum_{i=1}^{n-1} b_{i+\theta_0-1}\ (\tau_i - \tau_{i-1}))$, $n = 2, 3, \dots$.

Now, for deterministic $x(0) = x_0 \in \mathbb{R}^n$, $x_0 \neq 0$, and $\theta(0) = \ell \in \mathcal{S}$, we have, using the Jensen inequality and Fubini, that

$$\int_0^\infty E[x(t)^2]dt = E\sum_{n=1}^\infty \int_{\tau_{n-1}}^{\tau_n^-} x(t)^2 dt \geq x_0^2 \sum_{n=1}^\infty \left(\frac{\theta_0}{n+\theta_0}\exp\left\{\int_0^\infty (\ln s)\lambda e^{-\lambda s}ds\right\}\right) = \infty.$$

By its turn, we have that $x(\tau_n) = a_{n+1}$ a.s., so that

$$\begin{aligned} E_{x_0,\theta_0}[x(\tau_n)^2] &= x_0^2 \prod_{i=1}^n \left\{\int_0^\infty \exp(-2b_{i+\theta_0-1}s).\lambda\exp(-\lambda s)ds\right\} \\ &= x_0^2\left(\prod_{i=1}^n\left(1 + \ln\left(\frac{i+\theta_0}{i+\theta_0-1}\right)\right)\right)^{-1} \to 0 \text{ as } n \to \infty. \end{aligned}$$

Moreover, since almost all trajectories of $\{x\}$ are decreasing, we have that $\lim_{t\to\infty} x(t)^2 = \lim_{n\to\infty} x(\tau_n)^2$ a.s.. Hence, using the Lebesgue monotone convergence theorem, we can write that $\lim_{t\to\infty} E[x(t)^2] = E[\lim_{n\to\infty} E_{x_0,\theta_0}[x(\tau_n)^2]] = 0$ for any joint distribution of (x_0, θ_0).

An important tool for solving our optimal control problem is an *equivalence lemma* which draws the connection between the above concepts and the spectrum of a certain infinite dimensional linear operator. We start with the following auxiliary proposition.

PROPOSITION 6.5. *Define $F = (F_1, F_2, \dots)$, where, referring to the SS (respectively, the SD) case, $F_i = A_i - B_iG_i$ (respectively, $F_i = A_i - K_iC_i$), $i \in \mathcal{S}$. Then $F \in \mathcal{H}_\infty^n$.*

Proof. See [4] for details. □

Now we have the equivalence lemma.

LEMMA 6.6. *Consider the operator $\mathcal{D}$ given by (4.6), some $G = (G_1, G_2, \dots) \in \mathcal{H}_\infty^{n,m}$, $C = (C_1, C_2, \dots) \in \mathcal{H}_\infty^{n,r}$, and $K = (K_1, K_2, \dots) \in \mathcal{H}_\infty^{r,n}$. Then the following hold.*

(E1) *The system (A, B, Λ) is SS with stabilizing G if and only if*

$$\sup\{\operatorname{Re}\lambda : \lambda \in \sigma(\mathcal{D})\} < 0, \quad F_i = A_i - B_iG_i, \quad i \in \mathcal{S}. \tag{6.6}$$

(E2) *Similarly, the system (C, A, Λ) is SD with K turning (C, A, Λ) detectable if and only if*

$$\sup\{\operatorname{Re}\lambda : \lambda \in \sigma(\mathcal{D})\} < 0, \quad F_i = A_i - K_iC_i, \quad i \in \mathcal{S}. \tag{6.7}$$

Proof. In order to carry out indistinctly the proof for (E1) or (E2), we shall consider $\mathcal{D}$ and (4.4) equipped either with $F_i = A_i - B_iG_i$ or with $F_i = A_i - K_iC_i$.

$\Rightarrow$*(Sufficiency).* We have that $Q(t)$ given by (4.5), with $x(t)$ satisfying (4.4), is continuous and continuously differentiable in $\mathcal{H}_1^n$. Also, from Proposition 4.6, $Q(t)$ satisfies (4.7). So, using Proposition 4.1, $Q(t) = T(t)Q^0$, $t \geq 0$, where $T(t)$ is the uniformly continuous semigroup generated by $\mathcal{D}$. Now, $\int_0^\infty \left\|T(t)Q^0\right\|_1 dt < \infty$ (see Lemma 4.3) and, since $E[\|x(t)\|^2] = tr(\sum_{i=1}^\infty E[x(t)x(t)^* 1_{\{\theta(t)=i\}}]) \leq n \|Q(t)\|_1$, it follows that $\int_0^\infty E[\|x(t)\|^2]dt < \infty$ for every initial condition r.v. (x_0, θ_0).

$\Leftarrow$ *(Necessity).* From the hypothesis, $\int_0^\infty E[\|x(t)\|^2]dt < \infty$ for every initial condition r.v. (x_0, θ_0) and $x(t)$ given by (4.4). This implies $\int_0^\infty \|Q(t)\|_1 dt < \infty$, where $Q(t)$ is given by (4.5). (Notice that $\|Q(t)\|_1 \leq \sum_{i=1}^\infty E[\|x(t)x(t)^*\| 1_{\{\theta(t)=i\}}] \leq E[\|x(t)\|^2]$.) Also, $Q(t)$ is continuous and continuously differentiable and, from Proposition 4.6, satisfies (4.7). So, using Proposition 4.1, we have that $Q(t) = T(t)Q^0$, $t \geq 0$, where $T(t)$ is the uniformly continuous semigroup generated by $\mathcal{D}$. Hence $\int_0^\infty \left\|T(t)Q^0\right\|_1 dt < \infty$ for every $Q^0 \in \mathcal{H}_1^{n^+}$. (This is justified since, for any $Q^0 \in \mathcal{H}_1^{n^+}$, we can always find r.v.'s x_0 and θ_0 that, subject to (4.5), produce Q^0.) Now, since our variables are defined in the field of the complex numbers, we may appeal to the Cartesian decomposition of Remark 2.5 so that there exist X^+, X^-, Y^+, and Y^- in $\mathcal{H}_1^{n^+}$ such that, for every $Q^0 \in \mathcal{H}_1^n$, $\int_0^\infty \|T(t)Q^0\|_1 dt = \int_0^\infty \|T(t)\{(X^+ - X^-) + \iota\,(Y^+ - Y^-)\}\|_1 dt < \infty$. The result follows from Lemma 4.3. □

Let us now turn our attention to the infinite-time optimal control problem. Since in this case there is no fixed "time horizon," we expect, parallel to the classical LQ problem, that the minimum cost (see (5.14)) should not depend on the starting time s whenever we preserve the same initial condition r.v.'s $(x(s), \theta(s))$. This suggests finding a constant function $[0, T] \ni t \mapsto S^T(t) = S$ satisfying (4.12) with terminal condition $S^T(T) = S$, a "matched" solution for the finite-time problem in the sense that it simulates the infinite-time case. Now we may notice that S satisfies (4.12) with terminal condition $S^T(T) = S$ if and only if it satisfies the ICARE $\mathcal{T}(S) = 0$. From this *liaison*, we should expect that the solution of our infinite-time problem

should hinge on the solution of the ICARE. This is indeed the case, as we shall show throughout the following propositions.

PROPOSITION 6.7. *$[0,T] \ni t \mapsto S^T(t) = S \in \mathcal{H}_1^n$ satisfies*

$$\begin{cases} \dot{S}^T(t) + \mathcal{T}(S^T(t)) = 0, \quad t \in (0,T), \\ S^T(T) = S \end{cases} \tag{6.8}$$

if and only if S satisfies $\mathcal{T}(S) = 0$.

Proof. The proof is a straightforward exercise. □

DEFINITION 6.8. *We say that $S = (S_1, S_2, \dots)$ is a positive semidefinite solution to the ICARE if $S \in \mathcal{H}_\infty^{n^+}$ and satisfies the ICARE*

$$\mathcal{T}(S) = 0. \tag{6.9}$$

Furthermore, S is a stabilizing solution to the ICARE if it is a positive semidefinite solution and $G = (G_1, G_2, \dots) = \mathcal{G}(S)$ stabilizes (A, B, Λ).

Proposition 6.9 below provides sufficient conditions for the existence of a solution to (6.9).

PROPOSITION 6.9. *Suppose (A, B, Λ) is SS. Then, for $L = 0 \in \mathcal{H}_\infty^{n^+}$, the value $S_i^T(0)$ of the (unique) solution $S^T(t) \in \mathcal{H}_\infty^{n^+}$, $t \in [0,T]$, to (4.12), converges to some $S_i \in \mathbb{M}(\mathbb{C}^n)^+$ as $T \to \infty$ for each $i \in \mathrm{S}$. Furthermore, $S = (S_1, S_2, \dots)$ belongs to $\mathcal{H}_\infty^{n^+}$ and satisfies the ICARE (6.9).*

Proof. From Proposition 4.9, the solution to (4.12) indeed exists and is unique for $T \in (0,\infty)$ and $L = 0 \in \mathcal{H}_\infty^{n^+}$. Let us now consider the finite-time control problem of section 5 with $s = 0$ and initial conditions $x(0) = x$ and $\theta(0) = i$, x and i deterministic and arbitrary in $\mathbb{C}^n$ and $\mathcal{S}$, respectively, time horizons $T_1, T_2 \in (0,\infty)$, $T_1 < T_2$, and terminal cost conditions $S^{T_1}(T_1) = S^{T_2}(T_2) = L = 0$. In this case, applying the definition (3.5) in (5.14), we have that

$$\begin{aligned} x^* S_i^{T_2}(0) x &\geq \min_{u \in \mathcal{U}^{T_2}} E\left[\int_0^{T_1} \left\|\mathcal{Q}^{1/2} x(t)\right\|^2 + \left\|\mathcal{R}^{1/2} u(t)\right\|^2 dt\right] \\ &\quad + \min_{u \in \mathcal{U}^{T_2}} E\left[\int_{T_1}^{T_2} \left\|\mathcal{Q}^{1/2} x(t)\right\|^2 + \left\|\mathcal{R}^{1/2} u(t)\right\|^2 dt\right] \\ &\geq \min_{u \in \mathcal{U}^{T_1}} E\left[\int_0^{T_1} \left\|\mathcal{Q}^{1/2} x(t)\right\|^2 + \left\|\mathcal{R}^{1/2} u(t)\right\|^2 dt\right] = x^* S_i^{T_1}(0) x, \end{aligned} \tag{6.10}$$

and since the above expression holds for every $x \in \mathbb{C}^n$,

$$0 \leq S_i^{T_1}(0) \leq S_i^{T_2}(0) \tag{6.11}$$

for every $T_1, T_2 \in (0,\infty)$, $T_1 < T_2$, and $i \in \mathcal{S}$. Let us assume for the moment that, for every $T \in (0,\infty)$ and $i \in \mathcal{S}$,

$$S_i^T(0) \leq dI \tag{6.12}$$

for some constant d which does not depend on i and T. This together with (6.11) allows us to apply Lemma 7.8 in the appendix, which proves the two first assertions of the proposition.

We shall now show that S satisfies $\mathcal{T}(S) = 0$. Bearing in mind Proposition 4.9 and Lemma 7.10 and arbitrarily fixing $a > 0$ and $T \in (a, \infty)$, let $S^T(t)$, $t \in [0, T]$, and $S^{T,-a}(t)$, $t \in [-a, T-a]$, respectively, be the solution to

$$\begin{cases} \dot{S}^T(t) + \mathcal{T}(S^T(t)) = 0, & t \in (0, T), \\ S^T(T) = 0 \end{cases} \tag{6.13}$$

and to

$$\begin{cases} \dot{S}^{T,-a}(t) + \mathcal{T}(S^{T,-a}(t)) = 0, & t \in (-a, T-a), \\ S^{T,-a}(T-a) = 0. \end{cases} \tag{6.14}$$

Since $S_i^T(0) \to S_i$ as $T \to \infty$, we have that $S_i^{T-a}(0) \to S_i$ as $T \to \infty$. Renaming T by $T - a$ in (6.13), it is clear that

$$S_i^{T,-a}(t) = S_i^{T-a}(t), \quad t \in [0, T-a], \tag{6.15}$$

so that $S_i^{T,-a}(0) \to S_i$ as $T \to \infty$. Now, from Lemma 7.10, we have that

$$S_i^T(a) = S_i^{T,-a}(0), \quad i \in \mathcal{S}. \tag{6.16}$$

Hence

$$\lim_{T\to\infty} S_i^T(a) = S_i, \quad i \in \mathcal{S}. \tag{6.17}$$

Rewriting (6.11) and (6.12) as $0 \le S_i^{T_1-a}(0) \le S_i^{T_2-a}(0)$ and $S_i^{T-a}(0) \le dI$, respectively, we have, from (6.15) and (6.16), that $0 \le S_i^{T_1}(a) \le S_i^{T_2}(a)$ and $S_i^T(a) \le dI$ for every $T, T_1, T_2 \in (a, \infty)$, $T_1 < T_2$, and $i \in \mathcal{S}$.

Now, from these two expressions in conjunction with (6.17), the fact that $S = (S_1, S_2, \dots) \in \mathcal{H}_\infty^{n^+}$, and assuming (6.12), we get via Lemma 7.9 of the appendix that

$$\lim_{T\to\infty} \mathcal{T}(S^T(a)) = \mathcal{T}(S). \tag{6.18}$$

Let us now define, for $(0, T) \ni t \mapsto S_i^T(t)$, the differential operator D such that $S_i^T(\cdot) \mapsto DS_i^T(\cdot) = \dot{S}_i^T(\cdot)$. From (6.17) and since $a > 0$ is arbitrary, we have that $\lim_{T\to\infty} S_i^T(t) = S_i$, $t \in (0, T)$. Therefore, from the continuity of D and viewing S_i as a constant function of t, we obtain $\lim_{T\to\infty} \dot{S}_i^T(t) = \lim_{T\to\infty} DS_i^T(t) = D(S_i) = 0$, $i \in \mathcal{S}$. Choosing $t = a$ in the above expression and in (6.13), we obtain

$$\lim_{T\to\infty} \dot{S}^T(a) = 0 \tag{6.19}$$

and

$$\dot{S}^T(a) + \mathcal{T}(S^T(a)) = 0. \tag{6.20}$$

Passing the above expression to the limit and using (6.18) and (6.19), we have that

$$0 = \lim_{T\to\infty} \dot{S}^T(a) + \lim_{T\to\infty} \mathcal{T}(S^T(a)) = \mathcal{T}(S). \tag{6.21}$$

Finally, let us show that $S_i^T(0) \le dI$, $i \in \mathcal{S}$, $T \in (0, \infty)$. This follows from the hypothesis that (A, B, Λ) is SS. Indeed, Definition 6.1 says that, in this case, there

exists $G \in \mathcal{H}_\infty^{m,n}$, which stabilizes (A, B, Λ). So let us define the stabilizing control policy $\overline{u}(t) = -G_{\theta(t)}x(t)$, $t \geq 0$, so that the dynamic (3.1) with $t \geq 0$ reads as $\dot{x}(t) = F_{\theta(t)}x(t)$, $t > 0$, with $F_{\theta(t)} = A_{\theta(t)} - B_{\theta(t)}G_{\theta(t)}$. We shall be interested in the specialized initial condition $x(0) = x$ and $\theta(0) = i$, where x and i are deterministic and arbitrary in $\mathbb{C}^n$ and $\mathcal{S}$, respectively.

From Lemma 6.6, we have that $\sup\{\mathrm{Re}\,\lambda : \lambda \in \sigma(\mathcal{D})\} < 0$ with $\mathcal{D}$ given by (4.6). Recalling that $\mathcal{D}$ generates a uniformly continuous semigroup, say, $T(t)$, we may invoke the equivalence among assertions 1, 2, and 3 of Lemma 4.3, as well as Corollary 4.4, so that, for some constant $\beta \in (0, \infty)$,

$$\int_0^\infty \left\|T(t)Q^0\right\|_1 dt \leq \beta \left\|Q^0\right\|_1 < \infty \tag{6.22}$$

for every $Q^0 \in \mathcal{H}_1^n$. Now, from semigroup theory, $Q(t) = T(t)Q^0, t \geq 0$, is the solution to the differential equation

$$\begin{cases} \dot{Q}(t) = \mathcal{D}(Q(t)), \quad t > 0, \\ Q(0) = Q^0 = (Q_1^0, Q_2^0, \dots) \in \mathcal{H}_1^n, \end{cases}$$

which, from Proposition 4.6, is expressed by (4.5) whenever the initial condition is such that

$$Q_i^0 = E[x(0)x(0)^* 1_{\{\theta(0)=i\}}] = xx^* 1_{\{\theta(0)=i\}}, \quad i \in \mathcal{S}. \tag{6.23}$$

Hence (6.22) reads as

$$\int_0^\infty \|Q(t)\|_1 \, dt \leq \beta \left\|Q^0\right\|_1 < \infty \tag{6.24}$$

for Q^0 satisfying (6.23). Now we have that $E[\|x(t)\|^2] \leq n \|Q(t)\|_1$and $\|Q^0\|_1 \leq E[\|x(0)\|^2] = \|x\|^2$, so that (6.24) becomes

$$\int_0^\infty E[\|x(t)\|^2]dt \leq n\beta \|x\|^2 < \infty. \tag{6.25}$$

Using Schwarz's inequality and Fubini and denoting $d = (\|\mathcal{Q}\| + \|\mathcal{R}\| \, \|G\|_\infty^2)n\beta$, the expression for the cost of policy $\overline{u}$ may be dominated from above as follows:

$$E\left[\int_0^\infty x(t)^*\mathcal{Q}x(t) + \overline{u}(t)^*\mathcal{R}\overline{u}(t)dt\right] \leq E\left[\int_0^\infty (\|\mathcal{Q}\| + \|\mathcal{R}\| \, \|G\|_\infty^2) \|x(t)\|^2 \, dt\right]$$
$$\leq (\|\mathcal{Q}\| + \|\mathcal{R}\| \, \|G\|_\infty^2)n\beta \|x\|^2 \leq d \|x\|^2 = x^* dI x.$$

Now, turning back to the finite-time control problem (see (6.10)), we have that

$$x^* S_i^T(0)x = \min_{u \in \mathcal{U}^T} E\left[\int_0^T x(t)^*\mathcal{Q}x(t) + u(t)^*\mathcal{R}u(t)dt\right]$$
$$\leq E\left[\int_0^\infty x(t)^*\mathcal{Q}x(t) + \overline{u}(t)^*\mathcal{R}\overline{u}(t)dt\right] \leq x^* dI x.$$

Since $T \in (0, \infty)$, $i \in \mathcal{S}$, and $x \in \mathbb{C}^n$ are arbitrary, $S_i^T(0) \leq dI$ for every T and i. □

Defining $\bar{\mathcal{Q}}^{1/2} = (\mathcal{Q}^{1/2}, \mathcal{Q}^{1/2}, \dots) \in \mathcal{H}_\infty^{n^+}$, the following proposition provides sufficient conditions for a solution of the ICARE to be stabilizing.

PROPOSITION 6.10. *Suppose* $(\bar{\mathcal{Q}}^{1/2}, A, \Lambda)$ *is SD and* S *is a positive semidefinite solution to the ICARE* (6.9). *Then* S *is a stabilizing solution to* (6.9).

Proof. For $x(t)$ given by (3.1) with $t \geq 0$ and $u(t) = -G_{\theta(t)}x(t)$, $G = (G_1, G_2, \dots) = \mathcal{G}(S) \in \mathcal{H}_\infty^{n,m}$, and arbitrary initial data (x_0, θ_0), let us consider the statistic $Q(t)$, $t \geq 0$, given by (4.5) with F replaced by $\bar{F} = (\bar{F}_1, \bar{F}_2, \dots)$, $\bar{F}_i = A_i - B_iG_i$. Furthermore, let us define the operators $\bar{\mathcal{D}}$ and $\hat{\mathcal{D}}$ as in (4.6), replacing F by $\bar{F}$ in the former and by $\hat{F} = (\hat{F}_1, \hat{F}_2, \dots)$ in the latter, where $\hat{F}_i = A_i - K_i\mathcal{Q}^{1/2}$ with $K = (K_1, K_2, \dots) \in \mathcal{H}_\infty^{r,n}$ such that $\sup\{\operatorname{Re}\lambda : \lambda \in \sigma(\hat{\mathcal{D}})\} < 0$ (such K exists, bearing in mind the SD hypothesis of the proposition).

The idea of the proof runs as follows. We must prove that $\int_0^\infty E[\,\|x(t)\|^2\,]dt < \infty$ for any initial data (x_0, θ_0). Proving this is tantamount to proving that $\int_0^\infty \|Q(t)\|_1\,dt < \infty$ for arbitrary initial data $Q^0 \in \mathcal{H}_1^{n^+}$, where $Q(t)$, given by (4.5), also satisfies the differential equation $\dot{Q}(t) = \bar{\mathcal{D}}(Q(t))$. This amounts then, essentially, to obtaining an adequate function that dominates $\|Q(t)\|_1$.

From Proposition 4.6 we may write

$$\begin{aligned}\dot{Q}_i(t) = \bar{\mathcal{D}}_i(Q(t)) &= \bar{F}_iQ_i(t) + Q_i(t)\bar{F}_i^* + \sum_{j=1}^{\infty}\lambda_{ji}Q_j(t)\\ &= \hat{F}_iQ_i(t) + Q_i(t)\hat{F}_i^* + \sum_{j=1}^{\infty}\lambda_{ji}Q_j(t) + \Delta_iQ_i(t) + Q_i(t)\Delta_i^*\\ &= \hat{\mathcal{D}}_i(Q(t)) + \Delta_iQ_i(t) + Q_i(t)\Delta_i^*, \qquad (6.26)\end{aligned}$$

where

$$\Delta_i = K_i\mathcal{Q}^{1/2} - B_iG_i. \qquad (6.27)$$

Now, for arbitrary $\varepsilon > 0$, $0 \leq \left(\varepsilon I - \frac{1}{\varepsilon}\Delta_i\right)Q_i(t)\left(\varepsilon I - \frac{1}{\varepsilon}\Delta_i\right)^*$, so that $\Delta_iQ_i(t) + Q_i(t)\,\Delta_i^* \leq \varepsilon^2Q_i(t) + \frac{1}{\varepsilon^2}\Delta_iQ_i(t)\Delta_i^*$. Thus, from (6.26),

$$\dot{Q}_i(t) \leq \hat{\mathcal{D}}_i(Q(t)) + \varepsilon^2Q_i(t) + \frac{1}{\varepsilon^2}\Delta_iQ_i(t)\Delta_i^*. \qquad (6.28)$$

Now, for $H = (H_1, H_2, \dots) \in \mathcal{H}_1^n$, let us define the operators $\tilde{\mathcal{D}}(H) = (\tilde{\mathcal{D}}_1(H), \tilde{\mathcal{D}}_2(H), \dots)$, $\Gamma(H) = (\Gamma_1(H), \Gamma_2(H), \dots)$, and $\mathcal{V}(H) = (\mathcal{V}_1(H), \mathcal{V}_2(H), \dots)$ in $Blt\,(\mathcal{H}_1^n)$, such that

$$\tilde{\mathcal{D}}_i(H) = \hat{\mathcal{D}}_i(H) + \varepsilon^2H_i, \quad \Gamma_i(H) = \Delta_iH_i\Delta_i^*, \qquad (6.29)$$

and

$$\mathcal{V}_i(H) = \left(\varepsilon I - \frac{1}{\varepsilon}\Delta_i\right)H_i\left(\varepsilon I - \frac{1}{\varepsilon}\Delta_i^*\right), \ \ i \in \mathcal{S}. \qquad (6.30)$$

From (6.29), we have that

$$\tilde{\mathcal{D}} = \hat{\mathcal{D}} + \varepsilon^2 I \qquad (6.31)$$

with I being the identity operator associated to $\mathcal{H}_1^n$. We can rewrite (6.28) as

$$\dot{Q}_i(t) \leq \tilde{\mathcal{D}}_i(Q(t)) + \frac{1}{\varepsilon^2}\Gamma_i(Q(t)). \qquad (6.32)$$

In order to use a comparison theorem, we consider now the nonhomogeneous differential equation

$$\begin{cases} \dot{R}_i(t) = \tilde{\mathcal{D}}_i(R(t)) + \frac{1}{\varepsilon^2}\Gamma_i(Q(t)), \\ R_i(0) = Q_i(0), \quad i \in \mathcal{S}, \end{cases} \tag{6.33}$$

or, equivalently, the Banach space differential equation

$$\begin{cases} \dot{R}(t) = \tilde{\mathcal{D}}(R(t)) + \frac{1}{\varepsilon^2}\Gamma\left(Q(t)\right), \\ R(0) = Q(0) \in \mathcal{H}_1^{n^+} \end{cases} \tag{6.34}$$

with $R(t) = (R_1(t), R_2(t), \dots)$. Now, for each finite T and time interval $[0,T]$, $Q(\cdot)$ and consequently $\frac{1}{\varepsilon^2}\Gamma\left(Q(\cdot)\right)$ belong to $L_1([0,T], \mathcal{H}_1^n)$ and are continuously differentiable. Thus Proposition 4.2 tells us that the unique solution $R(t) \in \mathcal{H}_1^n$ to (6.34) is given by

$$R(t) = T_{\tilde{\mathcal{D}}}(t)\left(Q(0)\right) + \frac{1}{\varepsilon^2}\int_0^t T_{\tilde{\mathcal{D}}}(t-s)(\Gamma(Q(s)))ds, \qquad t \in [0,T], \tag{6.35}$$

for any $Q(0) \in \mathcal{H}^{n^+}$, where $T_{\tilde{\mathcal{D}}}$ is the uniformly continuous semigroup generated by $\tilde{\mathcal{D}}$. Let us now define

$$U_i(t) = R_i(t) - Q_i(t), \quad i \in \mathcal{S}. \tag{6.36}$$

Then $U(t) = (U_1(t), U_2(t), \dots)$ belongs to $\mathcal{H}_1^n$ and satisfies the differential equation

$$\begin{cases} \dot{U}(t) = \tilde{\mathcal{D}}(U(t)) + \mathcal{V}\left(Q(t)\right), \\ U(0) = 0. \end{cases} \tag{6.37}$$

Indeed,

$$\begin{aligned} \dot{U}_i(t) &= \dot{R}_i(t) - \dot{Q}_i(t) = \tilde{\mathcal{D}}_i(R(t)) + \frac{1}{\varepsilon^2}\Delta_i Q_i(t)\Delta_i^* \ - \bar{\mathcal{D}}_i(Q(t)) \\ &= \hat{\mathcal{D}}_i(R(t)) + \varepsilon^2 R_i(t) + \frac{1}{\varepsilon^2}\Delta_i Q_i(t)\Delta_i^* \ - \left(\hat{\mathcal{D}}_i(Q(t)) + \Delta_i Q_i(t) + Q_i(t)\Delta_i^*\right) \\ &= \tilde{\mathcal{D}}_i(U(t)) + \left(\varepsilon I - \frac{1}{\varepsilon}\Delta_i\right) Q_i(t) \left(\varepsilon I - \frac{1}{\varepsilon}\Delta_i^*\right) = \tilde{\mathcal{D}}_i(U(t)) + \mathcal{V}_i(Q(t)). \end{aligned}$$

Now $Q(\cdot)$ and consequently $\mathcal{V}\left(Q(\cdot)\right)$ belong to $L_1([0,T], \mathcal{H}_1^n)$. Thus, from Proposition 4.2, $U(t) \in \mathcal{H}_1^n$ defined in (6.36) is the unique solution to (6.37) and is given by

$$U(t) = \int_0^t T_{\tilde{\mathcal{D}}}(t-s)\left(\mathcal{V}(Q(s))\right) ds, \qquad t \in [0,T].$$

Since $Q(s)$ and consequently $\mathcal{V}\left(Q(s)\right)$ belong to $\mathcal{H}_1^{n^+}$ and $\tilde{\mathcal{D}}$ is a bounded linear transformation, it follows that $T_{\tilde{\mathcal{D}}}(t-s)\mathcal{V}(Q(s)) = (\exp((t-s))\tilde{\mathcal{D}})\mathcal{V}(Q(s))$ belongs to $\mathcal{H}_1^{n^+}$. Hence $U(t) \in \mathcal{H}_1^{n^+}$, which, together with (6.36), sets our comparison result, i.e., $0 \le Q(t) \le R(t) \in \mathcal{H}_1^{n^+}$, $t \in [0,T]$, for arbitrary $Q(0) \in \mathcal{H}_1^{n^+}$ and each finite T. Now, using (6.35) and (2.2), we have that

$$\|Q(t)\|_1 \le \|T_{\tilde{\mathcal{D}}}(t)\left(Q(0)\right)\|_1 + \frac{1}{\varepsilon^2}\int_0^t \|T_{\tilde{\mathcal{D}}}(t-s)(\Gamma(Q(s)))\|_1 \, ds.$$

Hence integration on $[0,T]$ yields

(6.38)
$$\int_0^T \|Q(t)\|_1\,dt \le \int_0^T \|T_{\tilde{\mathcal{D}}}(t)\,(Q(0))\|_1\,dt + \frac{1}{\varepsilon^2}\int_0^T\int_0^t \|T_{\tilde{\mathcal{D}}}(t-s)(\Gamma(Q(s)))\|_1\,dsdt.$$

Referring to the last term of (6.38), let us define $l = t - s$ and

$$T^E_{\tilde{\mathcal{D}}}(r) = \begin{cases} T_{\tilde{\mathcal{D}}}(r) & \text{if } r \ge 0, \\ 0 & \text{if } r < 0, \end{cases}$$

so that

$$\int_0^T\int_0^t \|T_{\tilde{\mathcal{D}}}(t-s)(\Gamma(Q(s)))\|_1\,ds\;dt = \int_0^T\int_0^T \left\|T^E_{\tilde{\mathcal{D}}}(t-s)(\Gamma(Q(s)))\right\|_1\,dt\;ds$$

$$\le \int_0^T \|\Gamma(Q(s))\|_1 \int_0^{T-s} \|T_{\tilde{\mathcal{D}}}(l)\|\,dl\;ds \le \int_0^T \|\Gamma(Q(s))\|_1\,ds \int_0^T \|T_{\tilde{\mathcal{D}}}(l)\|\,dl. \tag{6.39}$$

Hence

$$\int_0^T \|Q(t)\|_1\,dt \le \left\{\|(Q(0))\|_1 + \frac{1}{\varepsilon^2}\int_0^T \|\Gamma(Q(s))\|_1\,ds\right\}\int_0^T \|T_{\tilde{\mathcal{D}}}(s)\|\,ds. \tag{6.40}$$

Let us now dominate $\|(\Gamma(Q(s)))\|_1$ from above. Using (6.27), (6.29), and defining $c = \max\{\|K\|^2_\infty, \|B\|^2_\infty\}$,

$$\begin{aligned}\|\Gamma_i(Q(s))\| &= \|(K_i\mathcal{Q}^{1/2} - B_iG_i)Q_i(s)(K_i\mathcal{Q}^{1/2} - B_iG_i)^*\| \\ &\le c(\|\mathcal{Q}^{1/2}Q_i(s)\mathcal{Q}^{1/2}\| + \|G_iQ_i(s)G_i^*\| + 2\|\mathcal{Q}^{1/2}Q_i(s)G_i^*\|).\end{aligned}$$

Now from (4.5),

$$\|\mathcal{Q}^{1/2}Q_i(s)\mathcal{Q}^{1/2}\| = \|\mathcal{Q}^{1/2}E[x(s)x(s)^*1_{\{\theta(s)=i\}}]\mathcal{Q}^{1/2}\| \le E[\|\mathcal{Q}^{1/2}x(s)1_{\{\theta(s)=i\}}\|^2].$$

Similarly, $\|G_iQ_i(s)G_i^*\| \le E[\|G_ix(s)1_{\{\theta(s)=i\}}\|^2]$ and, bearing in mind that $2ab \le a^2 + b^2$ for any real numbers a,b,

$$\begin{aligned}2\|\mathcal{Q}^{1/2}Q_i(s)G_i^*\| &\le 2E[\|\mathcal{Q}^{1/2}x(s)\,(G_ix(s))^*\,1_{\{\theta(s)=i\}}\|] \\ &\le 2E[\|\mathcal{Q}^{1/2}x(s)1_{\{\theta(s)=i\}}\|\|\,(G_ix(s))\,1_{\{\theta(s)=i\}}\|] \\ &\le E[\|\mathcal{Q}^{1/2}x(s)1_{\{\theta(s)=i\}}\|^2] + E[\|\,(G_ix(s))\,1_{\{\theta(s)=i\}}\|^2].\end{aligned}$$

Consequently,

$$\|\Gamma(Q(s))\|_1 = \sum_{i=1}^\infty \|\Gamma_i(Q(s))\| \le 2cE\left[\sum_{i=1}^\infty \left\{\left\|\mathcal{Q}^{1/2}x(s)\right\|^2 1_{\{\theta(s)=i\}}\right.\right.$$

$$\left.\left. + \left\|G_{\theta(s)}x(s)\right\|^2 1_{\{\theta(s)=i\}}\right\}\right] = 2cE[\|\mathcal{Q}^{1/2}x(s)\|^2 + \|G_{\theta(s)}x(s)\|^2],$$

and (6.40) becomes

$$\int_0^T \|Q(t)\|_1\,dt \le \left\{\|Q(0)\|_1 + \frac{2c}{\varepsilon^2}\int_0^T E[\|\mathcal{Q}^{1/2}x(s)\|^2\right.$$

$$\left. + \|G_{\theta(s)}x(s)\|^2]ds\right\}\int_0^T \|T_{\tilde{\mathcal{D}}}(t)\|\,dt \tag{6.41}$$

for arbitrary $T \in (0, \infty)$. At this point we shall use the two hypotheses of the proposition to obtain an adequate bound to the first integral on the right-hand side of (6.41). So let us first consider the finite-time optimal control problem in section 5 with matched termination cost $S^T(T) = L = S$ such that $\mathcal{T}(S) = 0$. Then it follows from Propositions 5.8 (with $s = 0$) and 6.7 that the optimal control reads as $\hat{u}^T(t) = -G_{\theta(t)}x(t)$ with $G = (G_1, G_2, \dots) = \mathcal{G}(S^T(t)) = \mathcal{G}(S) \in \mathcal{H}_\infty^{n,m}$. Moreover, from (3.5) and (5.14), and since norms are equivalent in finite dimensional spaces, we have that

$$\begin{aligned} &\|S\|_\infty E[\|x(0)\|^2] \geq E[x(0)^* S_{\theta(0)} x(0)] \\ &= E\left[\int_0^T (\|\mathcal{Q}^{1/2}x(t)\|^2 + \|G_{\theta(t)}x(t)\|_{\mathcal{R}}^2)dt + \|x(T)\|_{S_{\theta(T)}}^2\right] \\ &\geq \int_0^T E[\|\mathcal{Q}^{1/2}x(t)\|^2 + m_{\mathcal{R}}\|G_{\theta(t)}x(t)\|^2]dt \end{aligned} \tag{6.42}$$

for some $m_{\mathcal{R}} > 0$. Or else,

$$\int_0^T E[\|\mathcal{Q}^{1/2}x(t)\|^2 + \|G_{\theta(t)}x(t)\|^2]dt \leq d_1 E[\|x(0)\|^2] < \infty$$

for $d_1 = \|S\|_\infty (1 + \frac{1}{m_{\mathcal{R}}})$ which do not depend on T. Hence, using this expression to majorize the right-hand side of (6.41) and passing to the limit, we have that

$$\lim_{T\to\infty} \int_0^T \|Q(t)\|_1 \, dt \leq \left\{ \|(Q(0))\|_1 + \frac{2c}{\varepsilon^2} d_1 (E[\|x(0)\|]) \right\} \lim_{T\to\infty} \int_0^T \|T_{\tilde{\mathcal{D}}}(t)\| \, dt. \tag{6.43}$$

Now $\sup\{\operatorname{Re}\lambda : \lambda \in \sigma(\hat{\mathcal{D}})\} < 0$, and so, from continuity of the spectrum of $\tilde{\mathcal{D}}$ on ε (note, from (6.31), that $\sigma(\tilde{\mathcal{D}}) = \sigma(\hat{\mathcal{D}}) + \varepsilon^2$), we have that $\sup\{\operatorname{Re}\lambda : \lambda \in \sigma(\tilde{\mathcal{D}})\} < 0$ for some $\varepsilon > 0$ sufficiently small. Thus, from Lemma 4.3, $\lim_{T\to\infty} \int_0^T \|T_{\tilde{\mathcal{D}}}(t)\| \, dt < \infty$, where $T_{\tilde{\mathcal{D}}}(t)$ is the semigroup generated by the bounded linear operator $\tilde{\mathcal{D}}$, or else (see (6.43)) $\lim_{T\to\infty} \int_0^T \|Q(t)\|_1 \, dt < \infty$ whenever $Q(0) \in \mathcal{H}_1^{n^+}$. Moreover, $E[\|x(t)\|^2] \leq n \|Q(t)\|_1$ so that $\lim_{T\to\infty} \int_0^T E[\|x(t)\|^2]dt < \infty$ for any initial condition (x_0, θ_0). Hence $G = \mathcal{G}(S)$ stabilizes (A, B, Λ). □

The next proposition shows the uniqueness of stabilizing solutions to the ICARE (6.9) as well as the optimality of this solution.

PROPOSITION 6.11. *Suppose $S = (S_1, S_2, \dots)$ is a stabilizing solution to the ICARE* (6.9). *Then S is the unique stabilizing solution to* (6.9) *and, for any initial condition* $(x(0), \theta(0))$,

$$\begin{aligned} \inf_{u\in\mathcal{U}} \mathcal{J}(\vartheta_0, u) = \mathcal{J}(\vartheta_0, \hat{u}) &:= E\left[\int_0^\infty (\|\mathcal{Q}^{1/2}x(t)\|^2 + \|\mathcal{R}^{1/2}\hat{u}(t)\|^2)dt\right] \\ &= E[x(0)^* S_{\theta(0)} x(0)], \end{aligned} \tag{6.44}$$

where $\hat{u}(t) = -G_{\theta(t)}x(t)$ with $G = (G_1, G_2, \dots) = \mathcal{G}(S) \in \mathcal{H}_\infty^{n,m}$ and $x(t)$ is given by (3.1) *with $t \geq 0$ plugged with $\hat{u}$.*

Proof. First notice that $\hat{u}$, as defined above, stabilizes (A, B, Λ). Thus $\hat{u} \in \mathcal{U}$, so that $\mathcal{U}$ is nonempty. Now let us pick an arbitrary control strategy $u \in \mathcal{U}$ and, focusing on the finite time horizon case, let us define, for an arbitrary T, the matched cost termination $S^T(T) = L = S$ and the control policy u^T such that $[0, T] \ni t \mapsto u^T(t) = u(t) \in \mathbb{R}^m$. Clearly, $x^T(t) = x(t)$ for $t \in [0, T]$, where $x^T(t)$ satisfies system

(3.1) with $s = 0$ plugged with u^T, $x(t)$ satisfies (3.1) with $t \geq 0$ plugged with u, and the same initial data $(x(0), \theta(0))$ stands for both cases. Now, as a consequence of having $S^T(T) = L = S$, the solution to the Riccati equation (4.12) is $S^T(t) = S$, $t \in [0, T]$ (see Proposition 6.7), and, consequently, $S^T_{\theta(0)}(0) = S_{\theta(0)}$. Thus, from the cost definition (3.5) and Proposition 5.6, we have that

$$\begin{aligned} \mathcal{J}_{[0,T],S}(\vartheta_0, u^T) &= E\left[\int_0^T (\|\mathcal{Q}^{1/2}x(t)\|^2 + \|\mathcal{R}^{1/2}u(t)\|^2)dt\right] + E[x(T)^* S_{\theta(T)} x(T)] \\ &= E[x(0)^* S_{\theta(0)} x(0)] + E\left[\int_0^T \|B^*_{\theta(r)} S_{\theta(r)} x(r) + \mathcal{R}u(r)\|^2_{\mathcal{R}^{-1}} dr\right]. \end{aligned} \tag{6.45}$$

Now, since $u \in \mathcal{U}$, we have, from condition C2 in section 3, $0 \leq E[x(T)^* S_{\theta(T)}\, x(T)] \leq \|S\|_\infty E[\|x(T)\|^2] \to 0$ as $T \to \infty$. Thus, passing (6.45) to the limit and by inspection of (3.6), it turns out that

$$\begin{aligned} \mathcal{J}(\vartheta_0, u) &= \lim_{T\to\infty} \mathcal{J}_{[0,T],S}(\vartheta_0, u^T) = E\left[\int_0^\infty (\|\mathcal{Q}^{1/2}x(t)\|^2 + \|\mathcal{R}^{1/2}u(t)\|^2)dt\right] \\ &= E[x(0)^* S_{\theta(0)} x(0)] + E\left[\int_0^\infty \|B^*_{\theta(r)} S_{\theta(r)} x(r) + \mathcal{R}u(r)\|^2_{\mathcal{R}^{-1}} dr\right] \end{aligned} \tag{6.46}$$

for arbitrary $u \in \mathcal{U}$ and initial condition ϑ_0. Hence, bearing in mind that a stabilizing solution S to the ICARE (6.9) belongs to $\mathcal{U}$, the minimum of (6.46) over $u \in \mathcal{U}$ is achieved with $\hat{u}$, in which case the second term on the right-hand side of (6.46) is zero. (Recall from (4.10) that $\hat{u}(t) = \mathcal{R}^{-1} B^*_{\theta(t)} S_{\theta(t)} x(t)$.) Expression (6.44) then follows. Finally, let us suppose that there exists a stabilizing solution $V \neq S$ to the ICARE. As in the case of the stabilizing solution S, we shall arrive at the conclusion that $E[x(0)^* V_{\theta(0)} x(0)]$ is the minimum of $\mathcal{J}(\vartheta_0, u)$ over $u \in \mathcal{U}$. But the minimum clearly does not depend on S and V, so it follows that

$$E[x(0)^* S_{\theta(0)} x(0)] = E[x(0)^* V_{\theta(0)} x(0)]$$

for any initial condition r.v. $(x(0), \theta(0))$. Making $x(0) = x$ and $\theta(0) = i$, x and i deterministic and arbitrary in $\mathbb{C}^n$ and $\mathcal{S}$, respectively, the above equation becomes $x^* S_i x = x^* V_i x$. Since S_i and V_i are Hermitian for every $i \in \mathcal{S}$, $S = V$. ∎

PROPOSITION 6.12. *Suppose* (A, B, Λ) *is SS and* $(\bar{\mathcal{Q}}^{1/2}, A, \Lambda)$ *is SD. Then, for arbitrary terminal condition* $L \in \mathcal{H}^{n^+}_\infty$, *the value* $\tilde{S}^T_i(0)$ *of the unique solution* $\tilde{S}^T(t) \in \mathcal{H}^{n^+}_\infty$, $t \in [0, T]$, *to* (4.12) *converges to* $S_i \in \mathbb{M}(\mathbb{C}^n)^+$ *as* $T \to \infty$ *for each* $i \in \mathcal{S}$, *and* $S = (S_1, S_2, \dots)$ *is the stabilizing solution to the ICARE* (6.9).

Proof. The idea of the proof is to recast an essential result of Proposition 6.9, namely, that $\lim_{T\to\infty} S^T_i(0) = S_i$, $i \in \mathcal{S}$, where $S = (S_1, S_2, \dots)$ is a positive semidefinite solution to the ICARE (6.9) and $S^T(t)$, $t \in [0, T]$, satisfies (4.12) with null terminal condition.

Bearing in mind the finite-time control problem of section 5 with arbitrary terminal cost condition $L \in \mathcal{H}^{n^+}_\infty$ and specialized initial condition $x(0) = x$ and $\theta(0) = i$, x and i deterministic and arbitrary in $\mathbb{C}^n$ and $\mathcal{S}$, respectively, and using, in this order, Proposition 5.8 coupled with definition (3.5) with $s = 0$ and the last equality of (6.10),

we obtain that

$$x^*\tilde{S}_i^T(0)x = \min_{u\in\mathcal{U}^T} E\left[\int_0^T (\|\mathcal{Q}^{1/2}x(t)\|^2 + \|\mathcal{R}^{1/2}u(t)\|^2)dt + x(T)^*L_{\theta(T)}x(T)\right]$$

$$\geq \min_{u\in\mathcal{U}^T} E\left[\int_0^T (\|\mathcal{Q}^{1/2}x(t)\|^2 + \|\mathcal{R}^{1/2}u(t)\|^2)dt\right] = x^*S_i^T(0)x, \tag{6.47}$$

where $S_i^T(0)$, $i \in \mathcal{S}$, is as defined in Proposition 6.9. Taking the limit on both sides of (6.47), we have that

$$\liminf_{T\to\infty} x^*\tilde{S}_i^T(0)x \geq \liminf_{T\to\infty} x^*S_i^T(0)x = \lim_{T\to\infty} x^*S_i^T(0)x = x^*S_ix, \tag{6.48}$$

where, from Propositions 6.9, 6.10, and 6.11, $\lim_{T\to\infty} x^*S_i^T(0)x$ exists, and it is such that $S = (S_1, S_2, \dots)$ is the unique stabilizing solution to the ICARE (6.9).

Let us now select the policy $u(t) = -G_{\theta(t)}x(t)$, $t \geq 0$, where $G = (G_1, G_2, \dots) = \mathcal{G}(S) \in \mathcal{H}_\infty^{n,m}$. Since by assumption $\tilde{S}^T(T) = L$, $x^*\tilde{S}_i^T(0)x$ is the minimum cost for the finite-time control problem above. (See Proposition 5.8 with $s = 0$.) Hence, considering the restriction of $u(t)$ to the interval $[0, T]$ (which assigns an admissible policy in $\mathcal{U}^T$) and using definition (3.5), we have that

$$x^*\tilde{S}_i^T(0)x \leq E\left[\int_0^T (\|\mathcal{Q}^{1/2}x(t)\|^2 + \|\mathcal{R}^{1/2}u(t)\|^2)dt + x(T)^*L_{\theta(T)}x(T)\right]$$

$$\leq E\left[\int_0^T (\|\mathcal{Q}^{1/2}x(t)\|^2 + \|\mathcal{R}^{1/2}u(t)\|^2)dt\right] + \|L\|_\infty E[\|x(T)\|^2]. \tag{6.49}$$

Now, bearing in mind the infinite-time control problem, and since $S = (S_1, S_2, \dots)$ is stabilizing, we have that the limit of the integral in (6.49) exists and $E[\|x(T)\|^2] \to 0$ as $T \to \infty$. Therefore, taking limits in (6.49), it follows that

$$\limsup_{T\to\infty} x^*\tilde{S}_i^T(0)x \leq \lim_{T\to\infty} E\left[\int_0^T (\|\mathcal{Q}^{1/2}x(t)\|^2 + \|\mathcal{R}^{1/2}u(t)\|^2)dt\right]$$

$$= E\left[\int_0^\infty (\|\mathcal{Q}^{1/2}x(t)\|^2 + \|\mathcal{R}^{1/2}u(t)\|^2)dt\right] = x^*S_ix, \tag{6.50}$$

where the last equality above comes from Proposition 6.11. Now, (6.48) and (6.50) tell us that $\lim_{T\to\infty} x^*\tilde{S}_i^T(0)x = x^*S_ix$. Since x and i are arbitrary and S_i is Hermitian, $\lim_{T\to\infty} \tilde{S}_i^T(0) = S_i$ for each $i \in \mathcal{S}$ and arbitrary $L \in \mathcal{H}_\infty^{n^+}$. □

THEOREM 6.13. *Suppose (A, B, Λ) is SS. Then, for $L = 0 \in \mathcal{H}_\infty^{n^+}$, the value $S_i^T(0)$ of the unique solution $S^T(t) \in \mathcal{H}_\infty^{n^+}$, $t \in [0, T]$, to (4.12) converges to some $S_i \in \mathbb{M}(\mathbb{C}^n)^+$ as $T \to \infty$ for each $i \in \mathcal{S}$. Moreover, $S = (S_1, S_2, \dots)$ belongs to $\mathcal{H}_\infty^{n^+}$ and satisfies the ICARE (6.9). Suppose, in addition, that $(\bar{\mathcal{Q}}^{1/2}, A, \Lambda)$ is SD. Then S is stabilizing and the unique positive semidefinite solution to (6.9), and, for arbitrary terminal condition $L \in \mathcal{H}_\infty^{n^+}$, the value $\tilde{S}_i^T(0)$ of the unique solution $\tilde{S}^T(t) \in \mathcal{H}_\infty^{n^+}$, $t \in [0, T]$, to (4.12) converges to S_i as $T \to \infty$ for each $i \in \mathcal{S}$. Furthermore, the optimal control policy $\hat{u}$ is given by $\hat{u}(t) = -G_{\theta(t)}x(t)$ with $G = (G_1, G_2, \dots) = \mathcal{G}(S) \in \mathcal{H}_\infty^{n,m}$, and produces the cost $\mathcal{J}(\vartheta_0, \hat{u}) = \inf_{u\in\mathcal{U}} \mathcal{J}(\vartheta_0, u) = E[x(0)^*S_{\theta(0)}x(0)]$.*

Proof. The proof is a straightforward consequence of Propositions 4.9, 5.6, 5.8, 6.9, 6.10, 6.11, and 6.12. □

Remark 6.14. If we specialize our framework to the nonjump case, the definitions of SS and SD in section 6 recast the definitions of stabilizability and detectability of the standard linear/quadratic/deterministic case. Indeed, the definition of SS duly specialized to the single state case and with deterministic $x_0 \in \mathbb{C}^n$ means that $\int_0^\infty \|x(t)\|^2 \, dt < \infty$, where $\dot{x}(t) = Fx(t)$ with $F = A - BG \in \mathbb{M}(\mathbb{C}^n)$ for some $G \in \mathbb{M}(\mathbb{C}^n, \mathbb{C}^m)$ and arbitrary $x_0 \in \mathbb{C}^n$. Since F does not depend on $\{\theta\}$, and to remain in the operator theoretical context, let us view it as the infinitesimal generator of the semigroup $T_F(t)$ so that $x(t) = T_F(t)x_0$. We may then write that $\int_0^\infty \|T_F(t)x_0\|^2 \, dt < \infty$. Invoking Lemma 4.3 for this finite dimensional application, the latter expression is the same as saying that every eigenvalue of F is placed on the open left complex halfplane, i.e., (A, B) is stabilizable in the usual sense. The case of stochastic stability follows an analogous procedure.

7. Appendix.

7.1. Linear approximation of nonnecessarily holomorphic functionals, via a decomplexification concept. Based on the decomplexification concept defined in section 2, the following lemma provides the linear approximation of a complex function g.

LEMMA 7.1. *Assume the decomplexification* ${}^{\mathrm{R}}g : [0,\infty) \times \mathbb{R}^{2n} \mapsto \mathbb{R}$, *of* $g : [0,\infty) \times \mathbb{C}^n \mapsto \mathbb{R}$ *is (Fréchet-) differentiable. Then, for every* $(t,x) \in [0,\infty) \times \mathbb{C}^n$, g *has the linear approximation*

$$g(t+s, x+w) = g(t,x) + \frac{\partial}{\partial t} g(t,x)s + \nabla_{{}^{\mathrm{R}}x} {}^{\mathrm{R}}g(t, {}^{\mathrm{R}}x)' \, {}^{\mathrm{R}}w + o(\|(s,w)\|). \tag{7.1}$$

Proof. See [4] for details. □

Remark 7.2. Differentiability of ${}^{\mathrm{R}}g$ suffices to guarantee the existence of the linear approximation of g. More stringent conditions, such as g being holomorphic, are not required.

Remark 7.3. From the definition of decomplexification of $x \in \mathbb{C}^n$, as stated in section 2, it is clear that $\nabla_{{}^{\mathrm{R}}x} {}^{\mathrm{R}}g(t, {}^{\mathrm{R}}x)' {}^{\mathrm{R}}w = \nabla_{x_{\mathrm{Re}}} {}^{\mathrm{R}}g(t, {}^{\mathrm{R}}x)' \, w_{\mathrm{Re}} + \nabla_{x_{\mathrm{Im}}} {}^{\mathrm{R}}g(t, {}^{\mathrm{R}}x)' w_{\mathrm{Im}}$.

LEMMA 7.4. *Let* g *be the (nonholomorphic) function given by*

$$[0,T] \times \mathbb{C}^n \ni (t,x) \to g(t,x) = x^* S_m(t) x \in \mathbb{R} \tag{7.2}$$

with $S_m(t) = S_m(t)^* \in \mathbb{M}(\mathbb{C}^n)$ *differentiable for all* $t \in [0,T]$. *Then* $\frac{\partial}{\partial t} g(t,x) = x^* \dot{S}_m(t) x$, *and the differentiable function* ${}^{\mathrm{R}}g$ *is such that*

$$\nabla_{{}^{\mathrm{R}}x} {}^{\mathrm{R}}g(t, {}^{\mathrm{R}}x) = \begin{pmatrix} \nabla_{x_{\mathrm{Re}}} {}^{\mathrm{R}}g(t, {}^{\mathrm{R}}x) \\ \nabla_{x_{\mathrm{Im}}} {}^{\mathrm{R}}g(t, {}^{\mathrm{R}}x) \end{pmatrix} = 2 \begin{pmatrix} (S_m(t)x)_{\mathrm{Re}} \\ (S_m(t)x)_{\mathrm{Im}} \end{pmatrix}, \tag{7.3}$$

or else

$$\nabla_{{}^{\mathrm{R}}x} {}^{\mathrm{R}}g(t, {}^{\mathrm{R}}x)' \, {}^{\mathrm{R}}w = w^* S_m(t) x + x^* S_m(t) w. \tag{7.4}$$

Proof. See [4]. □

7.2. Support for the proof of Proposition 5.5. The following results are essential to the proof of Proposition 5.5.

LEMMA 7.5. *For arbitrary $\mathcal{Q},\mathcal{R} \in \mathbb{M}(\mathbb{C}^n)^+$ and $T \in \mathbb{M}(\mathbb{C}^m,\mathbb{C}^n)$, there exists $d_1 \geq 0$ such that, for every $(x,u) \in \mathbb{C}^n \times \mathbb{C}^m$, we have that*

$$d_1(x^*\mathcal{Q}x + u^*\mathcal{R}u) \geq u^*Tx + x^*T^*u.$$

Proof. See [4]. □

LEMMA 7.6. *$H \in \mathcal{H}_\infty^{n+} \Rightarrow H_i \leq H_0$ for some $H_0 \in \mathbb{M}(\mathbb{C}^n)^+$.*

Proof. Suppose, by contradiction, that there exists some sequence $\{H_{i_j}\}_{j\in\mathbb{N}}$, increasing and unbounded in the self-adjoint partial ordering. Hence there is $H_0^1 \in \mathbb{M}(\mathbb{C}^n)^+$ with $\left\|H_0^1\right\| \geq \|H\|_\infty$ and some $j_0 \in \mathbb{N}$ such that $H_{i_{j_0}} > H_0^1$, which leads us to $\left\|H_{i_{j_0}}\right\| > \left\|H_0^1\right\| \geq \|H\|_\infty \geq \left\|H_{i_{j_0}}\right\|$, which is a contradiction. □

LEMMA 7.7. *Let $\{x\}$ be given by (3.1) with $u \in \mathcal{U}^{s,T}$. Then, for $g(t,x,i) = x^*S_i^T(t)x$ defined on $\mathcal{X}$, there is some real valued function $M^u(x) \equiv M(x,u(t,x,i))$ defined on $\mathbb{C}^n \times \mathbb{C}^m$ with*

$$E_{s,x,i}\left[\int_s^t \mid M(x(r),u(r))dr \mid\right] < \infty, \tag{7.5}$$

and such that

$$\frac{\partial}{\partial t}g(t,x,i) + \left(\mathcal{L}^u g\right)(t,x,i) + M(x,u) \geq 0$$

for every $(t,x,i) \in ((s,T) \times \mathbb{C}^n \times \mathcal{S})$.

Proof. We use Lemmas 7.5 and 7.6. See [4] for details. □

7.3. Support for the proof of Proposition 6.9. The following results are essential to the proof of Proposition 6.9.

LEMMA 7.8. *For arbitrary $i \in \mathcal{S}$ and $H_i^T \in \mathbb{M}(\mathbb{C}^n)^+$, $T \in (0,\infty)$, suppose that $H_i^{T_1} \leq H_i^{T_2} \leq dI$ for every $T_1 < T_2$ and some constant $0 < d < \infty$ which does not depend on i, T_1, and T_2. Then there exists $H_i \in \mathbb{M}(\mathbb{C}^n)^+$ such that $H_i^T \to H_i$ as $T \to \infty$, $i \in \mathcal{S}$, and $H = (H_1, H_2, \dots) \in \mathcal{H}_\infty^{n+}$.*

Proof. The first assertion follows from a standard monotonicity result concerning positive semidefinite matrices. Now $\left\|H_i^T\right\| \leq d$ for every finite T. Hence $\|H_i\| \leq d$ for every $i \in \mathcal{S}$, which proves the second assertion. □

LEMMA 7.9. *For $H^T \in \mathcal{H}_\infty^{n+}$, $T \in (0,\infty)$, let us assume that*

1. *$H_i^T \to H_i \in \mathbb{M}(\mathbb{C}^n)^+$ as $T \to \infty$, $i \in \mathcal{S}$.*
2. *$H = (H_1, H_2, \dots) \in \mathcal{H}_\infty^{n+}$.*
3. *$H_i^{T_1} \leq H_i^{T_2} \leq dI$, $T_1, T_2 \in (0,\infty)$, $T_1 < T_2$, $i \in \mathcal{S}$, and some constant $0 < d < \infty$, which does not depend on i, T_1, and T_2.*

Then, with $\mathcal{T}$ given by (4.11), we get that

$$\lim_{T\to\infty} \mathcal{T}_i(H^T) = \mathcal{T}_i(H), \quad i \in \mathcal{S}, \quad \text{and} \quad \lim_{T\to\infty} \mathcal{T}(H^T) = \mathcal{T}(H). \tag{7.6}$$

Proof. From (4.11) we have that

$$\begin{aligned}
&\lim_{T\to\infty} \mathcal{T}_i(H^T) \\
&= \mathcal{Q} + A_i^* \lim_{T\to\infty} H_i^T + \left(\lim_{T\to\infty} H_i^T\right) A_i - \left(\lim_{T\to\infty} H_i^T\right) B_i\mathcal{R}^{-1}B_i^* \left(\lim_{T\to\infty} H_i^T\right) \\
&+ \lambda_{ii} \lim_{T\to\infty} H_i^T + \lim_{T\to\infty} \mathcal{E}_i(H^T) \\
&= \mathcal{Q} + A_i^*H_i + H_iA_i - H_iB_i\mathcal{R}^{-1}B_i^*H_i + \lambda_{ii}H_i + \lim_{T\to\infty} \mathcal{E}_i(H^T).
\end{aligned} \tag{7.7}$$

If we assume for the moment that $\lim_{T\to\infty}\mathcal{E}_i(H^T)=\mathcal{E}_i(H)$, then (7.6) follows immediately from (7.7). So, let us prove that the previous equation indeed holds, by first showing that, for arbitrary $x\in\mathbb{C}^n$, $\lim_{T\to\infty}x^*\mathcal{E}_i(H^T)x=x^*\mathcal{E}_i(H)x$. With this aim, we have from assertion 3 that $0\le\lambda_{ij}x^*H_j^{T_1}x\le\lambda_{ij}x^*H_j^{T_2}x\le\lambda_{ij}x^*dIx$ for every positive T_1,T_2, $T_1<T_2$, $i,j\in\mathcal{S}$, which implies that

$$0\le\lim_{M\to\infty}\sum_{j=1,j\neq i}^{M}\lambda_{ij}x^*H_j^{T_1}x\le\lim_{M\to\infty}\sum_{j=1,j\neq i}^{M}\lambda_{ij}x^*H_j^{T_2}x$$
$$\le\lim_{M\to\infty}\sum_{j=1,j\neq i}^{M}\lambda_{ij}x^*dIx=d\left\|x\right\|^2\left|\lambda_{ii}\right|.$$

Hence, from the monotonicity of the bounded function $g(M,T)=\sum_{j=1,j\neq i}^{M}\lambda_{ij}x^*H_j^Tx$ on M and T, and bearing in mind assertions 1 and 2, we may write that

$$\lim_{T\to\infty}x^*\mathcal{E}_i(H^T)x=\lim_{T\to\infty}x^*\left(\lim_{M\to\infty}\sum_{j=1,j\neq i}^{M}\lambda_{ij}H_j^T\right)x$$
$$=\lim_{M\to\infty}\lim_{T\to\infty}\sum_{j=1,j\neq i}^{M}\lambda_{ij}x^*H_j^Tx=\lim_{M\to\infty}\sum_{j=1,j\neq i}^{M}\lambda_{ij}x^*H_jx=x^*\mathcal{E}_i(H)x.$$

Since the above equation holds for every $x\in\mathbb{C}^n$ and $\mathcal{E}_i(H)$ is self-adjoint, $\lim_{T\to\infty}\mathcal{E}_i(H^T)=\mathcal{E}_i(H)$, so that the first expression of (7.6) follows. Consequently, bearing in mind Proposition 4.7, we get

$$\mathcal{H}_\infty^n\ni\mathcal{T}(H)=\left(\lim_{T\to\infty}\mathcal{T}_1(H^T),\lim_{T\to\infty}\mathcal{T}_2(H^T),\dots\right)$$
$$=\lim_{T\to\infty}(\mathcal{T}_1(H^T),\mathcal{T}_2(H^T),\dots)=\lim_{T\to\infty}\mathcal{T}(H^T).\qquad\square$$

LEMMA 7.10. *For finite T, $\Delta\in\mathbb{R}$, and $\mathcal{K}$ an operator from the Banach space X into X, let us consider the Banach space differential equation*

$$\begin{cases}\dot V(t)+\mathcal{K}(V(t))=0,\quad t\in(-\Delta,T-\Delta),\\ V(T-\Delta)=0,\end{cases}\tag{7.8}$$

and let $S^T(\cdot)\equiv S^{T,0}(\cdot)$ and $S^{T,-\Delta}(\cdot)$ be functions such that

$$S^{T,-\Delta}(t)=S^T(t+\Delta),\quad t\in[-\Delta,T-\Delta].\tag{7.9}$$

Then $S^{T,-\Delta}(\cdot)$ is a solution to (7.8) *if and only if $S^T(\cdot)$ is a solution to* (7.8) *with $\Delta=0$. If a solution to one system is unique, then it is the case of the other system too, and both solutions satisfy* (7.9).

Proof. The lemma is intuitive since it corresponds to a shift of T. Considering the "only if " part, let $S^{T,-\Delta}(\cdot)$ be such that

$$\begin{cases}\dot S^{T,-\Delta}(t)+\mathcal{K}(S^{T,-\Delta}(t))=0,\quad t\in(-\Delta,T-\Delta),\\ S^{T,-\Delta}(T-\Delta)=0.\end{cases}\tag{7.10}$$

Now, from (7.9), it follows that $\dot S^{T,-\Delta}(t)=\dot S^T(t+\Delta)$ and

$$S^{T,-\Delta}(T-\Delta)=S^T(T)=0,\tag{7.11}$$

so that substitution in (7.10) yields $\dot{S}^T\ (t+\Delta) + \mathcal{K}(S^T(\ t+\Delta)) = 0, \quad t \in (-\Delta, T-\Delta)$, i.e., $\dot{S}^T(t) + \mathcal{K}(S^T(t)) = 0, \quad t \in (0,T)$. This and (7.11) show that $S^T(\cdot)$ satisfies (7.8) with $\Delta = 0$. For the "if" part, an analogous procedure holds, and the uniqueness part of the proof is easily shown by contradiction. □

Acknowledgments. The authors would like to express their gratitude to the referees for their suggestions and helpful comments. The authors would also like to thank Professor J. B. do Val for invaluable conversation on the applications of our results in economics.

REFERENCES

[1] V. I. Arnold, *Ordinary Differential Equations*, MIT Press, Cambridge, MA, 1978.
[2] L. Arnold, *Stochastic Differential Equations Theory and Applications*, John Wiley and Sons, New York, 1974.
[3] C. D. Aliprantis and K. C. Border, *Infinite Dimensional Analysis: A Hitchhiker's Guide*, Stud. Econom. Theory 4, Springer-Verlag, Berlin, 1994.
[4] J. Baczynski, *Optimal Control for Continuous Time LQ-Problems with Infinite Markov Jump Parameters via Semigroup*, Ph.D. thesis, Federal University of Rio de Janeiro—UFRJ/COPPE, 2000.
[5] W. P. Blair, Jr. and D. D. Sworder, *Continuous-time regulation of a class of econometric models*, IEEE Trans. Systems Man. Cyber., 5 (1975), pp. 341–346.
[6] W. P. Blair, Jr. and D. D. Sworder, *Feedback control of a class of linear discrete systems with jump parameters and quadratic cost criteria*, Internat. J. Control, 21 (1975), pp. 833–844.
[7] H. A. P. Blom and Y. Bar-Shalom, *The interacting multiple model algorithm for systems with Markovian switching coefficients*, IEEE Trans. Automat. Control, 33 (1988), pp. 780–783.
[8] S. Bohacek and E. Jonckheere, *Linear Dynamically Varying LQ Control of Systems with Complicated Dynamics*, preprint, 1998.
[9] S. Bohacek and E. Jonckheere, *Linear Dynamically Varying Systems versus Jump Linear Systems*, preprint, 1998.
[10] E. Boukas and P. Shi, *Stochastic stability and guaranteed cost control of discrete-time uncertain systems with Markovian jumping parameters*, Internat. J. Robust Nonlinear Control, 8 (1998), pp. 1155–1167.
[11] K. L. Chung, *Markov Chains with Stationary Probabilities*, Springer-Verlag, New York, 1967.
[12] O. L. V. Costa and M. D. Fragoso, *Discrete-time LQ-optimal control problems for infinite Markov jump parameter systems*, IEEE Trans. Automat. Control, 40 (1995), pp. 2076–2088.
[13] O. L. V. Costa and M. D. Fragoso, *Stability results for discrete-time linear systems with Markovian jumping parameters*, J. Math. Anal. Appl., 179 (1993), pp. 154–178.
[14] O. L. V. Costa and R. P. Marques, *Mixed H_2 /H^∞ control of discrete-time Markovian jump linear systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 95–100.
[15] R. Curtain, *A semigroup approach to the LQG problem for infinite-dimensional systems*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 25 (1978), pp. 713–720.
[16] C. E. de Souza and M. D. Fragoso, *H^∞ control for linear systems with Markovian jumping parameters*, Control Theory Adv. Tech., 9 (1993), pp. 457–466.
[17] J. B. do Val and T. Basar, *Receding horizon control of jump linear systems and a macroeconomic policy problem*, J. Econom. Dynam. Control, 23 (1999), pp. 1099–1131.
[18] F. Dufour and R. J. Elliot, *Adaptive control of linear systems with Markov perturbations*, IEEE Trans. Automat. Control, 43 (1998), pp. 351–372.
[19] F. Dufour and P. Bertrand, *The filtering problem for continuous-time linear systems with Markovian switching coefficients*, Systems Control Lett., 23 (1994), pp. 453–461.
[20] R. J. Elliot and D. D. Sworder, *Control of a hybrid conditionally linear Gaussian process*, J. Optim. Theory Appl., 74 (1992), pp. 75–85.
[21] E. B. Dynkin, *Markov Processes,* Vol. 1, Academic Press, New York, 1965.
[22] X. Feng, K. A. Loparo, Y. Ji, and H. J. Chizeck, *Stochastic stability properties of jump linear systems*, IEEE Trans. Automat. Control, 37 (1992), pp. 1884–1892.
[23] W. H. Fleming and R. V. Rishel, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[24] M. D. Fragoso, *On a partially observable LQG problem for systems with Markovian jumping parameters*, Systems Control Lett., 10 (1988), pp. 349–356.
[25] M. D. Fragoso, *A small random perturbation analysis of a partially observable LQG problem for systems with Markovian jumping parameters*, IMA J. Math. Control Inform., 7 (1990), pp. 293–305.
[26] M. D. Fragoso, J. B. R. do Val, and D. L. Pinto, Jr., *Jump linear H^∞ control: The discrete-time case*, Control Theory Adv. Tech., 10 (1995), pp. 1459–1474.
[27] M. D. Fragoso and J. Baczynski, *On an Infinite Dimensional Perturbed Riccati Differential Equation Arising in Stochastic Control*, Internal report, National Laboratory for Scientific Computing—LNCC, 43/00, 2000; European Control Conference, Porto, Portugal, 2001, invited paper, to appear.
[28] M. D. Fragoso and E. M. Hemerly, *Optimal control for a class of noisy linear systems with Markovian jumping parameters and quadratic cost*, Internat. J. Systems Sci., 22 (1991), pp. 2553–2561.
[29] I. G. Gihman and A. V. Skorohod, *Introduction of the theory of random process*, in Stochastic Differential Equations, Springer-Verlag, Berlin, 1972.
[30] W. S. Gray and O. Gonzalez, *Modelling electromagnetic disturbances in closed-loop computer controlled flight systems*, in Proceedings of the 37th IEEE Conference on Decision and Control, Philadelphia, 1998, pp. 359–364.
[31] J. B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* I, Springer-Verlag, Berlin, Heidelberg, 1993.
[32] K. Itô and H. P. McKean, Jr., *Diffusion Processes and Their Sample Paths*, Springer-Verlag, Berlin, New York, 1974.
[33] Y. Ji and H. J. Chizeck, *Controllability, stabilizability, and continuous-time Markovian jumping linear quadratic control*, IEEE Trans. Automat. Control, 35 (1990), pp. 777–788.
[34] Y. Ji and H. J. Chizeck, *Jump linear quadratic Gaussian control: Steady-state solution and testable conditions*, Control Theory Adv. Tech., 6 (1990), pp. 289–319.
[35] Y. Ji, H. J. Chizek, X. Feng, and K. A. Loparo, *Stability and control of discrete-time jump linear systems*, Control Theory Adv. Tech., 7 (1991), pp. 247–270.
[36] S. Karlin and H. M. Taylor, *A Second Course in Stochastic Process*, Academic Press, New York, 1981.
[37] M. A. Krall, *Applied Analysis*, D. Reidel, Dordrecht, The Netherlands, 1986.
[38] R. Malhame and C. Y. Chong, *Electric load model synthesis by diffusion approximation in a high order hybrid state stochastic system*, IEEE Trans. Automat. Control, 30 (1985), pp. 854–860.
[39] M. Mariton, *Almost sure and moments stability of jump linear systems*, Systems Control Lett., 11 (1988), pp. 393–397.
[40] M. Mariton, *Jump Linear Systems in Automatic Control*, Marcel Dekker, New York, 1990.
[41] M. Mariton and P. Bertrand, *Output feedback for a class of linear systems with stochastic jump parameters*, IEEE Trans. Automat. Control, 30 (1985), pp. 898–903.
[42] T. Morozan, *Optimal stationary control for dynamic systems with Markov perturbations*, Stochastic Anal. Appl., 1 (1983), pp. 219–225.
[43] L. Nachbin, *Introdução à Análise Funcional: Espaços de Banach e Cálculo Diferencial*, The General Secretariat of the Organization of American States, Washington, D.C., 1976.
[44] A. W. Naylor and G. R. Sell, *Linear Operator Theory in Engineering and Science*, Springer-Verlag, New York, 1983.
[45] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
[46] D. D. Sworder, *Feedback control for a class of linear systems with jump parameters*, IEEE Trans. Automat. Control, 14 (1969), pp. 9–14.
[47] D. D. Sworder and R. O. Rogers, *An LQ solution to a control problem associated with a solar thermal central receiver*, IEEE Trans. Automat. Control, 28 (1983), pp. 971–978.
[48] Z. Wang and X. Yang, *Birth and Death Process and Markov Chains*, Springer-Verlag, Berlin, Science Press, Beijing, 1992.
[49] W. M. Wonham, *On a matrix Riccati equation of stochastic control*, SIAM J. Control, 6 (1968), pp. 681–697.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 1, pp. 298–327

UNIFORM ROBUST PERFORMANCE AGAINST QUASI-LTI UNCERTAINTY IN SAMPLED-DATA SYSTEMS*

SEAN E. BOURDON† AND GEIR E. DULLERUD‡

Abstract. Uniform robust performance of sampled-data systems is studied in the context of arbitrarily slowly time-varying structured perturbations. Exact conditions for robustness are obtained. These conditions are convex but inherently infinite dimensional in nature.

Key words. sampled-data systems, quasi-LTI, structured uncertainty, periodic systems

AMS subject classifications. 93C57, 93D09

PII. S0363012900366777

1. Introduction.

The main objective of this paper is to provide an exact characterization for uniform robust performance in *sampled-data systems* against a class of structured linear quasi-time-invariant (quasi-LTI) perturbations for systems with L_2 inputs. Motivated by the work of Poolla and Tikku [23] on standard time-invariant systems, we obtain separate conditions for uniform robust stability and uniform robust performance. The now ubiquitous use of digital hardware in the control of complex processes serves to underscore the importance of sampled-data and multirate systems, which, in turn, motivates the *exact* analysis presented herein. Although we focus primarily on sampled-data systems in what follows, this work has more general application to periodic continuous time systems such as multirate systems and jump systems. We therefore believe that this work may find wider application, in, for instance, control of networked systems.

We choose to work with a perturbation class of arbitrarily slowly time-varying operators. These operators, at least intuitively, closely approximate the set of linear time-invariant (LTI) operators and are therefore referred to as quasi-LTI. Moreover, this set seems to form a natural perturbation class in our framework. Our nominal model consists of a continuous time plant in feedback with a discrete time controller, both of which are LTI. However, this is necessarily an idealization since any physical system exhibits some degree of time-variation, no matter how modest.

Other perturbation classes have previously been considered in the same context. In Thompson et al. [27, 28], Hara, Nakajima, and Kabamba [16], and [13], conditions characterizing robustness to the class of LTI perturbations were obtained. The problem of finding exact conditions characterizing robust performance against periodically time-varying (PTV) perturbations has been worked on by Thompson et al. [27, 28], Sivashankar and Khargonekar [26], and Dullerud and Glover [14]. Also, linear time-varying (LTV) and quasi-PTV perturbations were considered in [12], where exact conditions for robust performance to these perturbation classes are provided.

*Received by the editors January 14, 2000; accepted for publication (in revised form) January 22, 2001; published electronically June 26, 2001.
http://www.siam.org/journals/sicon/40-1/36677.html

†Systems Control Group, Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario, Canada, M5S 3G4 (bourdon@control.toronto.edu, http://control.toronto.edu/∼bourdon).

‡Department of Mechanical and Industrial Engineering, University of Illinois, Urbana, IL 61801 (dullerud@uiuc.edu, http://epic.me.uiuc.edu/∼dullerud). The research of this author was supported by NSF under grant ECS-9875244 CAREER.

The approach used in the paper uses the framework for sampled-data robustness of [12] and appeals to the lifting techniques of [5, 4, 29, 30]. The analytical robustness conditions we obtain are in terms of a convex optimization problem over an infinite dimensional set. Computational issues associated with the conditions are dealt with in a separate paper [9]. In particular, it is shown in [9] that converging upper and lower bounds on the stability radius of a sampled-data system can be computed within any desired accuracy using computations involving only linear (finite dimensional) matrix inequalities.

The exact nature of our robustness conditions is not totally unexpected given the results found in [12] on LTI perturbations. In fact, robustness of sampled-data systems is similar to that of purely continuous time LTI systems in that a greater degree of time variation in the perturbation class results in a simpler robustness test.

Finally, in proving the necessity of our robustness conditions, we generalize a result regarding the so-called S-procedure first proved by Megretski and Treil [18] from a finite dimensional space to one which has a countable basis. Our version is a special case, since it is well known that the result does not hold in general for a countable number of quadratic forms.

2. Mathematical preliminaries. We begin by introducing some concepts from mathematical analysis. Our treatment is kept brief. However, the material presented here is standard; see, e.g., [7, 17] for a more complete introduction. Throughout, we denote the nonnegative integers by $\mathbb{N}_{\mathrm{o}}$ and the real and complex numbers by $\mathbb{R}$ and $\mathbb{C}$, respectively.

Suppose E is a Hilbert space. We denote the norm on E by $\|\cdot\|_E$, although for convenience we frequently suppress the subscript. The space of bounded linear operators on E is written $\mathcal{L}(E)$, on which we will always put the norm topology; if X is in $\mathcal{L}(E)$, we denote the E to E induced norm of X by $\|X\|_{E\to E}$. Furthermore, the adjoint of X is written X^*, its spectrum $\mathrm{spec}(X)$ and its spectral radius $\mathrm{rad}(X)$. Given a subspace $\mathcal{X} \subset \mathcal{L}(E)$, we denote the *open* unit ball by $\mathcal{U}\mathcal{X}$.

We will be primarily concerned with three specific Hilbert spaces. The first of these is $L_2^m[0,\infty)$, which is the standard set of square integrable functions mapping $[0,\infty)$ to the Euclidean space $\mathbb{R}^m$. For simplicity we refer to this space as L_2 when convenient. Given a real number h, we can also define a compressed version of the space $L_2^m[0,\infty)$ on the interval $[0,h)$. We will use $\mathcal{K}_2$ to denote the space $L_2^m[0,h)$.

A third Hilbert space of interest is formed using a given Hilbert space E, the base space, and is denoted $\ell_2(E)$. It is the space of sequences mapping $\mathbb{N}_{\mathrm{o}}$ to E consisting of elements $(x_0, x_1, x_2, \dots)$ which satisfy

$$\sum_{k=0}^{\infty} \|x_k\|_E^2 < \infty.$$

If the base space E is not particularly relevant, we abbreviate further to ℓ_2. We say an operator is LTI on ℓ_2 if it commutes with the unilateral shift.

The *half-plane algebra*, which we denote $\mathcal{A}_{\mathbb{C}^+}$, is a frequency domain space which will also play an important role throughout this paper. It is comprised of functions that map the closed right half-plane $\overline{\mathbb{C}}^+$ to the $m \times p$ complex matrices $\mathbb{C}^{m\times p}$, are continuous on $\overline{\mathbb{C}}^+ \cup \{\infty\}$, and are analytic on the open half-plane $\mathbb{C}^+$, with the norm $\|\widehat{\Delta}\|_\infty := \sup_{\omega\in\mathbb{R}} \bar{\sigma}(\widehat{\Delta}(j\omega))$; here $\bar{\sigma}(\cdot)$ is the maximum singular value. The following lemma about the half-plane algebra is a result we shall appeal to a number of times.

PROPOSITION 2.1. *The set of stable, proper, rational functions $\mathcal{RH}_\infty$ is dense in $\mathcal{A}_{\mathbb{C}^+}$.*

So, given any $\widehat{D} \in \mathcal{A}_{\mathbb{C}^+}$ and $\varepsilon > 0$, there exist matrices $X_0, \ldots, X_n$ for some $n \geq 0$, so that with $\widehat{F}(s) = \sum_{k=0}^{n} X_k \left(\frac{1-s}{1+s}\right)^k$ we have

$$\|\widehat{D} - \widehat{F}\|_\infty < \varepsilon.$$

That is, any element of $\mathcal{A}_{\mathbb{C}^+}$ can be approximated by a finite sum of the above form.

Note that $\mathcal{A}_{\mathbb{C}^+}$ is a subspace of $\mathcal{H}_\infty$, and therefore any function $\widehat{\Delta}$ in $\mathcal{A}_{\mathbb{C}^+}$ defines a causal operator $\mathbf{\Delta}$ on L_2 through multiplication and the Laplace transform. Let $\mathcal{L}_{\mathcal{A}_{\mathbb{C}^+}}$ denote this subspace of $\mathcal{L}(L_2)$, whose elements have such transfer function representations in $\mathcal{A}_{\mathbb{C}^+}$.

We will also frequently make use of the operator-valued space $\mathcal{A}_{\mathbb{D}}$, the discrete time counterpart to $\mathcal{A}_{\mathbb{C}^+}$, called the *disc algebra*. This space consists of functions $\check{G} : \overline{\mathbb{D}} \to \mathcal{L}(\mathcal{K}_2)$, which are analytic in the unit disc $\mathbb{D}$, continuous on $\overline{\mathbb{D}}$, and for which the norm

$$\|\check{G}\|_\infty := \max_{\omega \in \mathbb{R}} \|\check{G}(e^{j\omega})\|_{\mathcal{K}_2 \to \mathcal{K}_2} = \max_{z \in \overline{\mathbb{D}}} \|\check{G}(z)\|_{\mathcal{K}_2 \to \mathcal{K}_2}$$

is finite.

3. Problem formulation. In this section, we pose the problem that is to be the main focus of this paper and provide a technical overview of our results. To begin, a brief introduction to uncertain sampled-data systems is presented. Next, we define the central notions of uniform robust stability and uniform robust performance for the systems under consideration, and then we define the particular perturbation sets of this paper and their associated scalings. Finally, we state the main results of the paper, whose proofs are covered in detail in section 5.

A standard configuration for uncertain sampled-data systems is depicted in Figure 3.1. This paradigm for studying robust performance, in our context of structured perturbations, was first introduced for continuous time systems by Doyle [11] and Safonov [25] and can incorporate many standard models of uncertainty; see also the survey article [21] on the structured singular value.

In the figure, the operator $\mathbf{G}$ represents a finite dimensional linear time-invariant (FDLTI) system with minimal state space realization (A, B, C, D). That is, $\mathbf{G}$ has a transfer function representation $\hat{G}(s) := C(sI - A)^{-1}B + D = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right]$. Moreover, we will assume throughout the paper that the realization for our plant $\mathbf{G}$ has the form

$$\hat{G}(s) = \left[\begin{array}{c|ccc} A & B_1^1 & B_1^2 & B_2 \\ \hline C_1^1 & 0 & 0 & D_{12}^1 \\ C_1^2 & 0 & 0 & D_{12}^2 \\ C_2 & 0 & 0 & 0 \end{array}\right] =: \begin{bmatrix} \hat{G}_{11} & \hat{G}_{12} \\ \hat{G}_{21} & \hat{G}_{22} \end{bmatrix}, \tag{3.1}$$

where each of the matrices B, C, and D is partitioned with respect to its inputs and its outputs. Notice that the matrix $D_{21} = 0$. This ensures that the signal y is low-pass filtered. We have also set $D_{11} = 0$ and $D_{22} = 0$ for simplicity, although these restrictions can be relaxed without affecting our subsequent results.

Our plant $\mathbf{G}$ is in feedback with a discrete time FDLTI controller $\mathbf{K}_d$ through an *ideal sampler* $\mathbf{S}$ and a *zero-order hold operator* $\mathbf{H}$. These mappings satisfy

$$(\mathbf{S}u)[k] := u(kh),$$
$$(\mathbf{H}v)(t) := v[k] \text{ for } t \in [kh, (k+1)h),$$

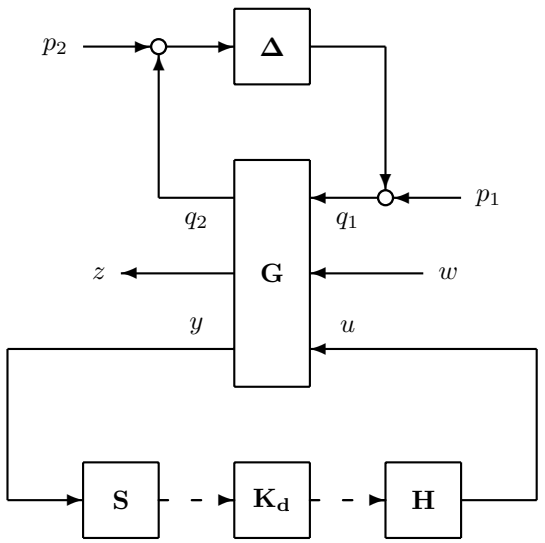


FIG. 3.1. *Uncertain sampled-data system.*

for each $u : [0, \infty) \to \mathbb{C}^n$, every sequence $v : \mathbb{N}_o \to \mathbb{C}^n$, and some real number $h > 0$, called the *sampling period* of the sampled-data system. Throughout the paper, we assume that the sample and hold devices are *synchronized* and that the sampling period h is *fixed*. We further assume that $(A_{K_d}, B_{K_d}, C_{K_d}, D_{K_d})$ is a minimal state space realization for $\mathbf{K}_d$.

The final assumption we place on our controller $\mathbf{K}_d$ is that it asymptotically stabilizes its nominal interconnection with our plant $\mathbf{G}$. For future reference, when $\mathbf{\Delta} = 0$, we call the mapping $\mathbf{M} : \begin{bmatrix} p_1 \\ w \end{bmatrix} \mapsto \begin{bmatrix} q_2 \\ z \end{bmatrix}$ from Figure 3.1 the *nominal sampled-data system*.

Assumption 3.1. Suppose that the signals w, p_1, and p_2 from Figure 3.1 are all zero. Further suppose that the operator $\mathbf{\Delta} = 0$. Then for any initial states $x_G(0)$ and $x_{K_d}[0]$ of the minimal state space realizations for $\mathbf{G}$ and $\mathbf{K}_d$, respectively, the limits $\lim_{t\to\infty} x_G(t) = 0$ and $\lim_{k\to\infty} x_{K_d}[k] = 0$ are both satisfied.

The above condition guarantees input-output stability of the nominal sampled-data system. However, it is also possible to make this condition equivalent to input-output stability; see [10].

The various signals appearing in Figure 3.1 are all physically meaningful. For instance, w represents all exogenous inputs to our system, such as disturbances, noise, and command signals. The regulated output z is the signal which is to be attenuated. The internal inputs to our system are given by p_1 and p_2. Of course, we require that our system be stable with respect to these inputs. The internal outputs are the signals q_1 and q_2. The signal u contains the controlled inputs to our system, whereas the measured outputs of the system are found in y.

Throughout the paper, we assume that the dimension of the signals p_1, p_2, q_1, and q_2 is m and that r denotes the dimension of w and z. Under these assumptions, we notice that the nominal closed-loop system is square. As with many of our other assumptions, this one is made out of convenience and our results will not be affected

if we choose to remove it. The final assumption we make on our signals is that the inputs w, p_1, and p_2 all belong to the space of bounded energy signals L_2.

Finally, the operator $\mathbf{\Delta}$ also appears in our system through a feedback loop and represents a perturbation to the nominal model. Its purpose is to encompass the uncertainty incurred by inaccuracies in the mathematical description. Recall that our nominal model is linear. However, by including uncertainty directly into our model description, our results are applicable to a much wider range of possibilities, including nonlinear and time-varying systems. The exact nature of quasi-LTI perturbations will be discussed below.

Recall that this paper is dedicated primarily to the study of uniform robust stabilization and performance of sampled-data systems against structured quasi-LTI perturbations. Intuitively, the system of Figure 3.1 has uniform robust stability to an uncertainty set $\mathcal{S}$ if it is internally stable given any perturbation $\mathbf{\Delta} \in \mathcal{S}$. This being the case, we say that our system has uniform robust performance to the set $\mathcal{S}$ if a performance inequality is also satisfied. The following definition makes these notions precise.

DEFINITION 3.2. *Suppose $\mathcal{X}$ is a subspace of $\mathcal{L}(L_2)$ and $\rho > 0$. Then the system in Figure* 3.1 *is said to have* uniform robust stability *against perturbations in the set $\rho\mathcal{UX}$ if the maps $\begin{bmatrix} w \\ p_1 \\ p_2 \end{bmatrix} \mapsto \begin{bmatrix} z \\ q_1 \\ q_2 \end{bmatrix}$ exist for each $\mathbf{\Delta} \in \rho\mathcal{UX}$ and are uniformly bounded in norm. If, in addition, the performance inequality $\rho \cdot \|w \mapsto z\| \leq 1$ is satisfied for all $\mathbf{\Delta} \in \rho\mathcal{UX}$, then the system in Figure* 3.1 *is said to have* uniform robust performance *with respect to that same perturbation set.*

Notice that both definitions are made with the same scaling constant ρ. In practice, however, we will usually set $\rho = 1$ for convenience when stating and proving results. This can be done without loss of generality since all of the systems we consider are linear and can hence be scaled appropriately a priori. Also note that a similar argument shows that the radius of the uncertainty set and the bound on the performance inequality can be varied independently by having first scaled $\mathbf{G}$. Having defined uniform robust stability and uniform robust performance, we can now introduce the set of quasi-LTI perturbations, which is the specific uncertainty set $\mathcal{S}$ that is the focus of this paper.

The perturbations we work with are assumed to be members of the spatially structured set

$$\mathcal{X}_s := \{\mathbf{\Delta} = \operatorname{diag}(\mathbf{\Delta}^1, \ldots, \mathbf{\Delta}^d) : \mathbf{\Delta}^k \in \mathcal{L}(L_2^{m_k}) \text{ for } 1 \leq k \leq d\},$$

where $\sum_{k=0}^{d} m_k = m$. Note that it is the Euclidean part of the elements of $\mathcal{X}_s$ on which the structure is imposed; given an operator $\mathbf{\Delta} = \operatorname{diag}(\mathbf{\Delta}^1, \ldots, \mathbf{\Delta}^d) \in \mathcal{X}_s$ and a signal $u = (u_1, \ldots, u_d) \in L_2^m$ with $u_k \in L_2^{m_k}$, we have $\mathbf{\Delta}u = (\mathbf{\Delta}^1 u_1, \ldots, \mathbf{\Delta}^d u_d)$. Also notice that the spatial blocks are all square. Again this is strictly for simplicity, and all of our results hold when this assumption is removed. This structured uncertainty arrangement is particularly useful in that it models uncertainty occurring simultaneously in various parts of the model in a nonconservative fashion. We refer to [3] and [20], which provide additional motivation for using this particular uncertainty arrangement from an engineering perspective.

Now let us define the set of quasi-LTI operators. By this, we mean perturbations $\mathbf{\Delta}$ lying in the set

$$\mathcal{L}_{LTI}(\nu) := \left\{ \mathbf{\Delta} \in \mathcal{L}(L_2^m) : \sup_{T>0} \frac{\|\mathbf{D}_T \mathbf{\Delta} - \mathbf{\Delta} \mathbf{D}_T\|}{T} \leq \nu, \mathbf{\Delta} \text{ causal} \right\},$$

where $\nu > 0$ and $\mathbf{D}_T$ is the T-shift on L_2. This set is precisely the continuous time analogue of the discrete time slowly time-varying set in [23] and is called quasi-LTI since, intuitively, less time variation is permitted as ν is decreased; indeed, the set $\mathcal{L}_{LTI}(0)$ corresponds to the set of all causal LTI operators on L_2. We can then define the set of quasi-LTI *structured* operators via

$$\mathcal{X}_{LTI}(\nu) := \mathcal{X}_s \cap \mathcal{L}_{LTI}(\nu).$$

In the case where $\mathcal{X}_{LTI}(\nu) = \mathcal{L}_{LTI}(\nu)$, we say that our perturbation set is *unstructured*.

Accordingly, we now define a particular class of so-called D-scaling sets, which will appear throughout what follows. Let $\mathfrak{D}^s_{LTI}$ be the set of all *nonsingular* operators in $\mathcal{L}_{\mathcal{A}_{\mathbb{C}^+}}$ whose spatial structure allows them to commute with all members of $\mathcal{X}_{LTI}(0)$. Specifically, we have

$$\mathfrak{D}^s_{LTI} = \{\mathbf{D} \in \mathcal{L}_{\mathcal{A}_{\mathbb{C}^+}} : \mathbf{D}\boldsymbol{\Delta} = \boldsymbol{\Delta}\mathbf{D} \text{ for each } \boldsymbol{\Delta} \in \mathcal{X}_{LTI}(0), \text{ and } 0 \notin \operatorname{spec}(\mathbf{D})\}.$$

That is, every $\mathbf{D} \in \mathfrak{D}^s_{LTI}$ has a corresponding transfer function representation $\hat{D} \in \mathcal{A}_{\mathbb{C}^+}$ of the form $\hat{D} = \operatorname{diag}(\hat{d}_1 I_{m_1}, \ldots, \hat{d}_d I_{m_d})$, where each scalar function $\hat{d}_k \in \mathcal{A}_{\mathbb{C}^+}$. Of course, if $\mathcal{X}_{LTI}(\nu) = \mathcal{L}_{LTI}(\nu)$, the transfer function $\hat{D}$ simply has the form $\hat{D} = \hat{d} I_m$, where $\hat{d} \in \mathcal{A}_{\mathbb{C}^+}$, and we write $\mathfrak{D}^u_{LTI}$ in lieu of $\mathfrak{D}^s_{LTI}$ for this particular case.

Recall that we defined the operator $\mathbf{M}$ to be the nominal closed-loop sampled-data system mapping $\begin{bmatrix} p_1 \\ w \end{bmatrix} \mapsto \begin{bmatrix} q_2 \\ z \end{bmatrix}$ when $\boldsymbol{\Delta} = 0$. Thus the system of Figure 3.1 is exactly that of Figure 3.2 below; the latter is more convenient to work with in our framework. Also notice that by Assumption 3.1, the operator $\mathbf{M}$ is bounded. If we compatibly partition $\mathbf{M} =: \begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{bmatrix}$ with respect to its inputs and outputs, we are in a position to state our first result regarding robust stabilization. The lemma, whose proof is straightforward, says that we need only verify the existence and boundedness of one of the component maps of Figure 3.2 (rather than all four) in appealing to our definition of uniform robust stability.

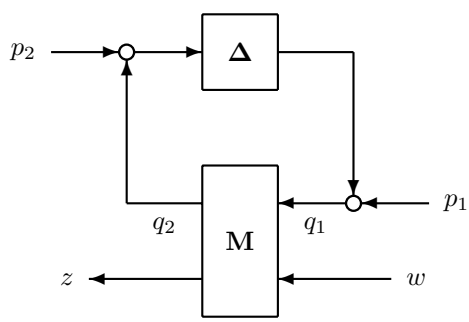


FIG. 3.2. *Robust performance configuration.*

LEMMA 3.3. *Suppose that $\mathcal{X}$ is a subspace of $\mathcal{L}(L_2)$ and $\rho > 0$. Then the system of Figure* 3.1 *has uniform robust stability to the perturbation set $\rho\mathcal{U}\mathcal{X}$ if and only if for each $\boldsymbol{\Delta} \in \rho\mathcal{U}\mathcal{X}$ the mapping $(\mathbf{I} - \mathbf{M}_{11}\boldsymbol{\Delta})^{-1}$ exists in $\mathcal{L}(L_2)$ and the family of maps $(\mathbf{I} - \mathbf{M}_{11}\boldsymbol{\Delta})^{-1}$ is uniformly bounded over $\mathcal{U}\mathcal{X}$.*

In light of the lemma, we see that uniform robust stability of a sampled-data system can be studied using the simplified configuration of Figure 3.3. Of course, we can use this framework to study uniform robust stability of the sampled-data system of Figure 3.2 by setting $\mathbf{M} = \mathbf{M}_{11}$ in Figure 3.3. This brings us to our main results.

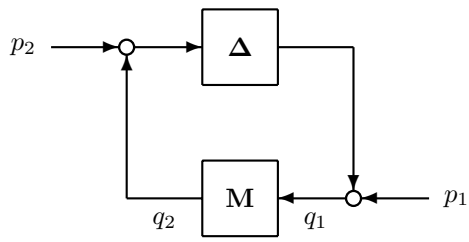


FIG. 3.3. *Robust stabilization configuration.*

THEOREM 3.4. *Suppose the nominal sampled-data closed-loop operator* $\mathbf{M} \in \mathcal{L}_{\mathcal{A}_{\mathbb{D}}}$. *Then, for every* $0 < \rho < 1$, *the system of Figure* 3.3 *has uniform robust stability against perturbations in* $\rho\mathcal{UX}_{LTI}(\nu)$ *for some* $\nu > 0$ *if and only if*

$$\inf_{\mathbf{D} \in \mathfrak{D}^s_{LTI}} \|\mathbf{D}\mathbf{M}\mathbf{D}^{-1}\|_{L_2 \to L_2} \leq 1.$$

An important feature of our robustness paradigm is that robust performance problems can be cast in a robust stability framework. See [8] for the details concerning this conversion. This then allows us to greatly expedite the proofs of our results as the framework established in proving robust stability results can be reused, modulo some technical modifications, to prove results concerning robust performance. To this end, we have Theorem 3.5 below, which constitutes the main result of the paper.

THEOREM 3.5. *Suppose the nominal sampled-data closed-loop operator* $\mathbf{M} \in \mathcal{L}_{\mathcal{A}_{\mathbb{D}}}$. *Then, for every* $0 < \rho < 1$, *the system of Figure* 3.1 *has uniform robust performance against perturbations in* $\rho\mathcal{UX}_{LTI}(\nu)$ *for some* $\nu > 0$ *if and only if*

$$\inf_{\mathbf{D} \in \mathfrak{D}^s_{LTI}} \left\| \begin{bmatrix} \mathbf{D} & 0 \\ 0 & \mathbf{I} \end{bmatrix} \mathbf{M} \begin{bmatrix} \mathbf{D} & 0 \\ 0 & \mathbf{I} \end{bmatrix}^{-1} \right\|_{L_2 \to L_2} \leq 1.$$

The theorem supplies us with a condition that guarantees uniform robust performance of the sampled-data system of Figure 3.1. Written in the operator formulation, the condition is precisely the same as the discrete time condition derived in [23].

Remark 3.6. Theorems 3.4 and 3.5 both involve the minimization of a scaled norm condition over the set $\mathfrak{D}^s_{LTI}$. Notice that by Proposition 2.1 we could have replaced this set by $\mathcal{L}_{\mathcal{RH}_\infty} \cap \mathfrak{D}^s_{LTI}$, where $\mathcal{L}_{\mathcal{RH}_\infty}$ is the subspace of linear operators on L_2 which have stable, proper, rational transfer function representations. Although all of our results still hold with this new set, it is more technically convenient to use $\mathfrak{D}^s_{LTI}$ in our proofs. Further, we point out that the results do *not* hold in general if $\mathcal{A}_{\mathbb{C}^+}$ and $\mathcal{A}_{\mathbb{D}}$ are replaced by their associated H_∞ spaces.

We now make a few comments regarding the results presented in this section. First, notice that our robustness conditions involve a minimization over the D-scaling set $\mathfrak{D}^s_{LTI}$. That we obtain D-scaling problems is not entirely surprising in light

of previous work in the area. In [12], it is shown that a structured singular value calculation is an exact test for robust stability of a sampled-data system to the set of LTI perturbations. It was also shown that if the perturbations are LTV, PTV, or quasi-PTV, then the test for robustness involves a D-scaling problem. Of course, it is well known that D-scaling problems form natural upper bounds to structured singular value calculations.

Having stated the main results of the paper, we now outline the related work contained within the next few sections. We begin by introducing a further set of tools and results required in proving our necessity results. In section 5, we prove a simplified version of Theorem 3.4 with $\mathcal{X}_{LTI}(\nu) = \mathcal{L}_{LTI}(\nu)$. Namely, we limit ourselves to the case where our perturbations are unstructured. The proofs for robustness to the class of structured quasi-LTI perturbations require mostly technical modifications from the results presented in section 5 and are omitted due to space considerations. Complete details concerning these extensions may be found in [8].

4. Analysis of sampled-data systems. In section 2, we focused on presenting mathematical tools which can be used to study any dynamical system. The goal of this section is to develop some further results that will simplify the proofs of the main results stated in the last section. The material presented below is collected in this section because the results are of independent interest. First, we examine some of the tools required in the analysis of sampled-data systems. Namely, we will introduce a lifting formalism for periodic operators as well as another operator-valued representation of our nominal system $\mathbf{M}$ called the sampled-data frequency response. We reserve the last subsection of this section for an introduction to the S-procedure.

4.1. Lifting of periodic systems. We begin our study of sampled-data systems by describing a technique for lifting periodic operators. The aim of this formalism, which was first developed in [5, 29, 30, 4] is to provide a framework in which the periodic system becomes time-invariant. Although the technique introduced applies to any periodic operator in $\mathcal{L}(L_2)$, we focus primarily on sampled-data systems.

The first step is to define the *sampled-data lifting operator* $\mathbf{W}$ as a mapping from $L_2[0,\infty)$ to $\ell_2(\mathcal{K}_2)$. Given $u \in L_2$, the sequence $\tilde{u} = \mathbf{W}u$ is defined via

$$(\tilde{u}[k])(\tau) := u(kh + \tau) \tag{4.1}$$

for $\tau \in [0,h)$ and $k \in \mathbb{N}_o$. From the definition, it is obvious that $\mathbf{W}^{-1}$ exists and that $\mathbf{W}$ is an isomorphism between L_2 and $\ell_2(\mathcal{K}_2)$. Hence, if $\mathbf{F} \in \mathcal{L}(L_2)$, then the mapping $\tilde{F} := \mathbf{W}\mathbf{F}\mathbf{W}^{-1}$ is a bounded linear operator on $\ell_2(\mathcal{K}_2)$; in fact, $\|\mathbf{F}\|_{L_2 \to L_2} = \|\tilde{F}\|_{\ell_2 \to \ell_2}$. Note the convention we have adopted here: if $u \in L_2$ and $\mathbf{F} \in \mathcal{L}(L_2)$, then the lifted signal and the lifted system are denoted by $\tilde{u}$ and $\tilde{F}$, respectively.

Using the sampled-data lifting operator and the Z-transform, we can obtain a transfer function representation for $\mathbf{M}$ of the form $\check{M}(z) = \check{C}z(I - zA_d)^{-1}\check{B} + \check{D}$. The explicit form of the operators on the right-hand side is in Appendix A. Using these expressions, it is then an easy matter to show that $\check{M}(z) \in \mathcal{A}_{\mathbb{D}}$. This property constitutes the starting point for subsection 4.2.1.

4.2. The sampled-data frequency response. The second tool we introduce in this section is the sampled-data frequency response, which provides us with another frequency domain representation for a class of operators in $\mathcal{L}(L_2)$. This new representation is at the heart of the necessity conditions proposed in Theorems 3.4 and 3.5, as is best seen later through Theorem 5.2. As the name suggests, the sampled-data frequency response plays an analogous role to the Fourier frequency response for

standard continuous time systems. This connection will become more obvious in what follows.

Our study of the sampled-data frequency response is split into three parts. We begin by defining the frequency response operator. We then focus on the class of LTI perturbations having transfer function representations in the half-plane algebra $\mathcal{A}_{\mathbb{C}^+}$. This allows us to define the D-scaling sets we will use to scale the frequency response of our sampled-data systems. In the second part, we dwell on an asymptotic property of the frequency response operator in proving two technical lemmas used in the proof of Theorem 3.4. Finally, in subsection 4.2.3, we discuss the continuity of the mappings introduced in defining the sampled-data frequency response operator of a system. Our presentation is based on that of [12], and we refer the reader to this book and the references cited therein for a more complete overview than the one presented here; see also [1, 2, 31].

4.2.1. Lifting in frequency domain. We begin by defining the set $\mathcal{L}_{\mathcal{A}_{\mathbb{D}}}$ to consist of operators $\mathbf{G} \in \mathcal{L}(L_2)$ for which there exists a function $\check{G}(z) \in \mathcal{A}_{\mathbb{D}}$ such that $\mathbf{G} = \mathbf{W}^{-1}Z^{-1}\check{G}Z\mathbf{W}$. Notice that at a fixed point $z_{\mathrm{o}} \in \overline{\mathbb{D}}$ the operator $\check{G}(z_{\mathrm{o}}) : \mathcal{K}_2 \to \mathcal{K}_2$. Clearly, every operator in the set $\mathcal{L}_{\mathcal{A}_{\mathbb{D}}}$ is h-periodic, although the set of all causal h-periodic operators on L_2 is isomorphic to the larger space $\mathcal{H}_\infty(\mathbb{D})$. For such an operator, we also have that

$$\|\mathbf{G}\| = \max_{z\in\mathbb{D}} \|\check{G}(z)\| = \sup_{\omega\in\mathbb{R}} \|\check{G}(e^{j\omega})\|,$$

where the second equality follows from a maximum modulus result. Thus it seems that when dealing with questions about robust stability and performance, we need only concern ourselves with the behavior of $\check{G}(z)$ along the boundary of the unit disc since this is where the function takes its "largest" values. As we will see later, this is precisely the case.

Let us now briefly discuss the space $\mathcal{K}_2$. It is not difficult to show that for any $\omega_{\mathrm{o}} \in (-\pi, \pi]$ the sequence $\{\psi_k\}$ forms a complete orthonormal basis for $\mathcal{K}_2$, where

$$\psi_k(t) := h^{-\frac{1}{2}} e^{jh^{-1}(2\pi\nu_k - \omega_{\mathrm{o}})t} \tag{4.2}$$

for $t \in [0, h)$, and ν_k is the kth element in the sequence $\{0, 1, -1, 2, -2, \dots\}$.

We now define a one-parameter family of operators $J_\omega : \mathcal{K}_2 \to \ell_2$ for $\omega \in (-\pi, \pi]$. Given $\psi \in \mathcal{K}_2$ with Fourier expansion $\psi = \sum_{k=0}^{\infty} a_k\psi_k$, we have

$$J_{\omega_{\mathrm{o}}}\psi := (a_0, a_1, a_2, \dots),$$

where $\omega_{\mathrm{o}} \in (-\pi, \pi]$ is the frequency at which the basis $\{\psi_k\}$ is defined. In this fashion, we can define an operator-valued function $J : \partial\mathbb{D} \to \mathcal{L}(\mathcal{K}_2, \ell_2)$ through the relationship

$$J(e^{j\omega}) := J_\omega.$$

From this definition, it is immediate that $J^{-1} = J^*$ at each point $\omega \in (-\pi, \pi]$ and that J is an isomorphism between the space of square integrable $\mathcal{K}_2$-valued functions on $\partial\mathbb{D}$ and the square integrable ℓ_2-valued functions on $\partial\mathbb{D}$.

The sampled-data frequency response of an operator $\mathbf{G} \in \mathcal{L}_{\mathcal{A}_{\mathbb{D}}}$ is then defined by

$$G(e^{j\omega}) := J_\omega \check{G}(e^{j\omega}) J_\omega^*. \tag{4.3}$$

Our above discussion allows us to conclude that $G : \partial\mathbb{D} \to \mathcal{L}(\ell_2)$ and that

$$\sup_{\omega\in(-\pi,\pi]} \|G(e^{j\omega})\|_{\ell_2\to\ell_2} = \sup_{\omega\in(-\pi,\pi]} \|\check{G}(e^{j\omega})\|_{\mathcal{K}_2\to\mathcal{K}_2} = \|\mathbf{G}\|_{L_2\to L_2}.$$

In the last section, we saw that for the sampled-data system of Figure 3.1, the transfer function $\check{M}(e^{j\omega}) = \check{C}e^{j\omega}(I - e^{j\omega}A_d)^{-1}\check{B} + \check{D}$. Using our mapping J, we can define new operators $\tilde{B} := \check{B}J_\omega^*$, $\tilde{C} := J_\omega\check{C}$, and $\tilde{D} := J_\omega\check{D}J_\omega^*$ so that

$$M(e^{j\omega}) = \tilde{C}e^{j\omega}(I - e^{j\omega}A_d)^{-1}\tilde{B} + \tilde{D}.$$

From these definitions, it is easy to see that $\tilde{B} : \ell_2 \to \mathbb{C}^{\tilde{n}}$, $\tilde{C} : \mathbb{C}^{\tilde{n}} \to \ell_2$, and $\tilde{D} : \ell_2 \to \ell_2$. Hence each operator can be viewed as an infinite dimensional "matrix." For example, we can write

$$\tilde{B} =: \begin{bmatrix}(\tilde{B})_0 & (\tilde{B})_1 & (\tilde{B})_2 & \cdots\end{bmatrix},$$

where the block $(\tilde{B})_k$ is simply a matrix acting on the kth element of a sequence in ℓ_2^m. Similar definitions can be made for $(\tilde{C})_l$ and $(\tilde{D})_{lk}$, the matrix components of $\tilde{C}$ and $\tilde{D}$, respectively. State space formulae for all of the above quantities can be found in Appendix A.

We end our introduction to the sampled-data frequency response with a closer examination of a special case. The results that we state can be found in [12] along with their proofs. Recall that $\mathcal{L}_{\mathcal{A}_{\mathbb{C}^+}}$ denotes the subspace of operators in $\mathcal{L}(L_2)$ which have transfer function representations in the half-plane algebra $\mathcal{A}_{\mathbb{C}^+}$. Suppose our perturbation $\mathbf{\Delta} \in \mathcal{X}_{LTI}(0) \cap \mathcal{L}_{\mathcal{A}_{\mathbb{C}^+}}$. That is, $\mathbf{\Delta}$ is LTI and lies in our spatially structured set, while its transfer function $\hat{\Delta} \in \mathcal{A}_{\mathbb{C}^+}$. By Proposition 2.1, this set is the closure of the FDLTI operators in the spatially structured set $\mathcal{X}_s$. For such a perturbation, it can be shown that

$$\Delta(e^{j\omega_\circ}) = \operatorname{diag}(\hat{\Delta}(j\theta_0), \hat{\Delta}(j\theta_1), \hat{\Delta}(j\theta_2), \dots),$$

where the frequency $\theta_k = \frac{2\pi\nu_k - \omega_\circ}{h}$. See [12]. Thus $\Delta(e^{j\omega_\circ})$ can be viewed as an infinite dimensional block diagonal matrix whose blocks inherit their spatial structure from $\mathcal{X}_s$. More precisely, by defining the set

$$\mathbf{\Delta}_{LTI} := \{\operatorname{diag}(\Delta_0, \Delta_1, \Delta_2, \dots) : \Delta_k \in \mathcal{X}\},$$

where the set of spatially structured matrices

$$\mathcal{X} := \{\operatorname{diag}(Q_1, Q_2, \dots, Q_d) : Q_k \in \mathbb{C}^{m_k\times m_k}\} \subset \mathbb{C}^{m\times m},$$

we see that $\Delta(e^{j\omega}) \in \mathbf{\Delta}_{LTI}$ at each frequency along the unit circle.

Finally, we can define the set of D-scaling operators for $M(e^{j\omega})$. This is done by analogy with our definitions of the sets $\mathfrak{D}^s_{LTI}$ and $\mathfrak{D}^u_{LTI}$. Let $\tilde{\mathfrak{D}}^s$ be the set of nonsingular operators which commute with each member of $\mathbf{\Delta}_{LTI}$. From this definition, it is easy to show that

$$\tilde{\mathfrak{D}}^s = \{\operatorname{diag}(\tilde{D}_0, \tilde{D}_1, \tilde{D}_2, \dots) : \tilde{D}_k = \operatorname{diag}(\tilde{d}_{k,1}I_{m_1}, \dots, \tilde{d}_{k,d}I_{m_d}),\ 0 \neq \tilde{d}_{k,l} \in \mathbb{C}\}.$$

Of course, we can similarly conclude that when $\tilde{\mathfrak{D}}^s = \tilde{\mathfrak{D}}^u$, each block $\tilde{D}_k = \tilde{d}_k I_m$ with $0 \neq \tilde{d}_k \in \mathbb{C}$. We now further restrict this set by appealing to the following elementary result.

PROPOSITION 4.1. *Given bounded linear operators M and D acting on some Hilbert space $\mathcal{H}$ with D invertible. Then for a scalar $\delta > 0$, the following are equivalent:*

(1) $\|DMD^{-1}\|^2 \leq \delta^2$.
(2) $M^*D^*DM - \delta^2 D^*D \leq 0$.

We therefore define the positive subsets

$$\mathfrak{D}^s = \{\mathrm{diag}(D_0, D_1, D_2, \dots) : D_k = \mathrm{diag}(d_{k,1}I_{m_1}, \dots, d_{k,d}I_{m_d}),\ 0 < d_{k,l} \in \mathbb{R}\} \tag{4.4}$$

and

$$\mathfrak{D}^u = \{\mathrm{diag}(d_0 I_m, d_1 I_m, d_2 I_m, \dots) : 0 < d_k \in \mathbb{R}\} \tag{4.5}$$

and will work with them instead of $\tilde{\mathfrak{D}}^s$ and $\tilde{\mathfrak{D}}^u$. With this, we conclude our introduction to the sampled-data frequency response. Further properties of this representation are discussed in the next two subsections.

4.2.2. Frequency response as an asymptotic limit. We now focus on another key property of the sampled-data frequency response operator. The results below provide us with a new means by which we can connect the original operator $\mathbf{M}$ to its frequency response function $M(e^{j\omega})$. The result will be used later when we prove the necessity of the condition of Theorem 3.4.

To begin with, we need to define a two-parameter set of scalar functions

$$\phi^q_{\omega_o}(t) := \begin{cases} \frac{1}{\sqrt{qh}} e^{j\omega_o t}, & 0 \leq t < qh, \\ 0, & t \geq qh, \end{cases} \tag{4.6}$$

where $\omega_o \in \mathbb{R}$ and $q \in \mathbb{N}_o$. Notice that any such function always has unit norm. These functions have a useful property in connection with the frequency response operator, as is seen in the following lemma.

LEMMA 4.2. *Given $b_0, \dots, b_N \in \mathbb{C}^m$, a frequency $\theta_o \in (-\pi, \pi]$, and the sequence $\nu_k = \{0, 1, -1, 2, -2, \dots\}$. Let $\omega_l := \frac{2\pi\nu_l - \theta_o}{h}$ for every $l \in \mathbb{N}_o$. Then*

$$\|z^q - \mathbf{M}w^q\|_{L_2} \longrightarrow 0 \text{ as } q \to \infty,$$

where $\mathbf{M}$ is the nominal sampled-data system of Figure 3.1, $w^q(t) = \sum_{l=0}^N b_l \phi^q_{\omega_l}(t)$, and $z^q(t) = \sum_{l=0}^N \sum_{p=0}^\infty \left(M_{p,l}(e^{j\theta_o})b_l\right) \phi^q_{\omega_p}(t)$.

The lemma states that each of the $N+1$ harmonics $\omega_0, \dots, \omega_N$ generates a countable number of aliased harmonics whose sizes are determined from the frequency response operator $M(e^{j\omega})$; in interpreting this result, observe for a fixed q that the functions $\phi^q_{\omega_l}$ form an orthonormal sequence in L_2.

4.2.3. Continuity properties. We end our introduction to the sampled-data frequency response with a look at an important continuity property of the frequency response operator $M(e^{j\omega})$ defined in subsection 4.2.1. Let us begin by defining the operator $X \in \mathcal{L}(\ell_2)$ via

$$(b_0, b_1, b_2, \dots) \xrightarrow{X} (b_{\eta_0}, b_{\eta_1}, b_{\eta_2}, \dots),$$

where η_k is the sequence $\{2, 0, 4, 1, 6, 3, 8, 5, 10, 7, 12, 9, \dots\}$.

Recall that $M(e^{j\omega})$ provides an alternative representation for the multiplication operator $\check{M}(z)$ on the unit circle $\partial\mathbb{D}$. The interesting thing to note is that although the transfer function $\check{M}(\cdot)$ is continuous on $\partial\mathbb{D}$, Proposition 4.3 implies that its counterpart $M(\cdot)$ does not always share this property in general. Nonetheless, we still have the following result, which says that a discontinuity can occur at only one point on $\partial\mathbb{D}$.

PROPOSITION 4.3. *The frequency response of the nominal sampled-data system* **M** *of Figure* 3.1 *satisfies the following two properties.*

(i) $M(e^{j\omega})$ *is continuous on* $(-\pi, \pi)$ *and left continuous at* π.

(ii) $\lim_{\omega\to-\pi^+} \|M(e^{j\omega}) - X^*M(-1)X\|_{\ell_2\to\ell_2} = 0$.

The proof results from an easy application of the triangle inequality and the continuity of the lifting operator J_ω. See [12].

4.3. The S-procedure. The final topic we will cover in this section is the S-procedure [18]. In our presentation we provide a new result, which generalizes earlier work from a finite to a countable number of quadratic forms under special conditions. It is worth noting that the general result for a countable number of quadratic forms does not hold.

It is in this vein that we introduce time-invariant quadratic forms on ℓ_2: a mapping $\psi : \ell_2 \to \mathbb{R}$ is called a time-invariant quadratic form if there exist two time-invariant operators X and Y in $\mathcal{L}(\ell_2)$ satisfying

$$\psi(u) = \|Xu\|_2^2 - \|Yu\|_2^2$$

for each $u \in \ell_2$.

In the work that follows, we will work exclusively with sequences of time-invariant operators on ℓ_2. However, our study will be limited to those sequences which satisfy the following condition.

Condition 4.4. A sequence $\{X_k\}$ of time-invariant operators on ℓ_2 satisfies this condition if $\sum_{k=0}^{\infty} \|X_k u\|_2^2$ is finite for every $u \in \ell_2$ with $\|u\|_2 = 1$.

Note, in particular, that any such sequence X_k tends strongly to zero.

Suppose we have two sequences $\{X_k\}$ and $\{Y_k\}$ of time-invariant operators on ℓ_2 which satisfy Condition 4.4. We then define the set

$$\nabla := \{(\psi_0(u), \psi_1(u), \dots) : u \in \ell_2,\ \|u\|_2 = 1\} \subset \ell_\infty, \tag{4.7}$$

where $\psi_k(u) = \|X_k u\|_2^2 - \|Y_k u\|_2^2$ for $k \in \mathbb{N}_o$ and ℓ_∞ is the set of bounded real-valued sequences. Lemma 4.5 below states an important property of the set ∇.

LEMMA 4.5. *Suppose* $\{X_k\}$ *and* $\{Y_k\}$ *are sequences of time-invariant operators on* ℓ_2 *which satisfy Condition* 4.4*, and* ∇ *is the corresponding subset of* ℓ_∞ *defined in* (4.7)*. Then* ∇ *is a subset of* ℓ_1*, the set of absolutely summable sequences, and its closure* $\overline{\nabla}$ *is convex.*

The proof of this result can be found in [8] along with the proofs of the other results of this section; it is an extension of the proof of an analogous result found in [18] and adopts the presentation of [22].

In what follows, we denote the positive orthant of ℓ_1 by

$$\Pi^+ := \{x \in \ell_1 : x = (x_0, x_1, \dots), x_k \geq 0 \text{ for each } k \in \mathbb{N}_o\}.$$

This definition in hand, we are now set to present the key result in this section. The proof relies on the strong separation theorem for normed spaces and the fact that the normed dual space of ℓ_1 is isomorphic to ℓ_∞.

THEOREM 4.6. *Suppose* $\{X_k\}$ *and* $\{Y_k\}$ *are sequences of time-invariant operators in* $\mathcal{L}(\ell_2)$ *such that* ∇*, the corresponding set defined by* (4.7)*, is a bounded subset of* ℓ_1*. Then the following statements are equivalent.*

(i) *The inequality* $\inf_{x\in\nabla,\ y\in\Pi^+} \|x-y\|_{\ell_1} > 0$ *is satisfied.*

(ii) *There exists a bounded sequence of real scalars* $d_\infty, d_0, d_1, d_2, \ldots > \beta > 0$ *for some* $\beta > 0$*, such that* $d_0\psi_0(u) + d_1\psi_1(u) + \cdots \le -d_\infty$ *for all* $u \in \ell_2$ *with* $\|u\|_2 = 1$*.*

We now begin the process of narrowing our focus by specializing this last result to our sampled-data framework. For each $l \in \mathbb{N}_{\mathrm{o}}$, we define the projection operator E_l on ℓ_2 to be the operator whose representation is

$$E_l = \operatorname{diag}(\underbrace{0, 0, \ldots, 0}_{l \text{ zeros}}, I_m, 0, \ldots),$$

where I_m is the $m \times m$ identity matrix. Now consider the space $\ell_2(\ell_2^m)$. Suppose $L \in \mathcal{L}(\ell_2^m)$. We define the *memoryless* operator $T_L : \ell_2(\ell_2^m) \to \ell_2(\ell_2^m)$ as the mapping satisfying

$$(T_L u)_k := L u_k$$

for each $k \in \mathbb{N}_{\mathrm{o}}$. Given an LTI operator V on $\ell_2(\ell_2)$ and an integer $N \in \mathbb{N}_{\mathrm{o}}$, we define the special set of quadratic forms

$$\psi_l(u) := \begin{cases} \|T_{E_l} V u\|_2^2 - \|T_{E_l} u\|_2^2, & l = 0, \ldots, N, \\ \|T_{Q_N} V u\|_2^2 - \|T_{Q_N} u\|_2^2, & l = N+1, \\ 0, & l > N+1, \end{cases} \tag{4.8}$$

where $u \in \ell_2(\ell_2)$ and the projection operator $Q_n : \ell_2 \to \ell_2$ is defined by

$$Q_n(a_0, \ldots, a_n, a_{n+1}, \ldots) = (0, \ldots, 0, a_{n+1}, a_{n+2}, \ldots)$$

for each $a \in \ell_2$ and $n \in \mathbb{N}_{\mathrm{o}}$. Having made this definition, we are now able to state our next result. It relates condition (i) from the last theorem to an equivalent D-scaling problem over the set $\mathfrak{D}_n^u \subset \mathfrak{D}^u$ defined by

$$\mathfrak{D}_n^u = \{D = \operatorname{diag}(d_0 I_m, d_1 I_m, \ldots) \in \mathfrak{D}^u : d_l = d_{N+1} \text{ for } l > N+1\},$$

where $\mathfrak{D}^u$ is the D-scaling set defined in (4.5); thus note that $\mathfrak{D}_n^u$ is a subset of $\mathfrak{D}^u$.

COROLLARY 4.7. *Given a time-invariant operator* $V \in \mathcal{L}(\ell_2(\ell_2))$ *and the corresponding set of time-invariant quadratic forms defined in* (4.8)*. Then the inequality* $\inf_{x\in\nabla,\ y\in\Pi^+} \|x-y\|_{\ell_1} > 0$ *holds if and only if*

$$\inf_{D\in\mathfrak{D}_n^u} \|T_D V T_{D^{-1}}\|_{\ell_2(\ell_2)\to\ell_2(\ell_2)} < 1.$$

This simplified version of Theorem 4.6 is sufficient for our purposes since there is only a finite number of nonzero quadratic forms defined in (4.8). However, the full result may have wider application.

The following corollary is the final result of this section. It links the quadratic forms defined in (4.8) to the D-scaling set $\mathfrak{D}^u$ and is precisely the result we will appeal to later in proving the main theorems of the paper.

COROLLARY 4.8. *If the inequality*

$$\inf_{D\in\mathfrak{D}^u} \|DM(e^{j\omega_o})D^{-1}\|_{\ell_2\to\ell_2} > 1$$

holds for some frequency $\omega_o \in (-\pi,\pi]$, *then for each integer* $n > 0$ *there exist* $u \in \ell_2(\ell_2^m)$ *with* $\|u\| = 1$ *and* $\gamma > 1$ *such that*

$$\gamma\|T_{E_l}u\|_{\ell_2(\ell_2)} \le \|T_{E_l}T_M u\|_{\ell_2(\ell_2)}$$

for each $l = 0,\dots,n$, *and*

$$\gamma\|T_{Q_n}u\|_{\ell_2(\ell_2)} \le \|T_{Q_n}T_M u\|_{\ell_2(\ell_2)}.$$

Note here that T_M above, short for $T_{M(e^{j\omega_o})}$, is time-invariant.

We now have a complete set of tools with which to study our problems. We shall make extensive use of the techniques developed here in section 5, where we prove the necessity and sufficiency results of section 3.

5. Uniform robust stability: Unstructured perturbations. The proofs for Theorems 3.4 and 3.5 are quite lengthy. For the purposes of this paper, we will therefore concentrate on the following specialized result. Namely, we shall focus on characterizing uniform robust stability against the set of *unstructured* perturbations, $\mathcal{UL}_{LTI}(\nu)$. The extensions to Theorems 3.4 and 3.5 are routine; details are in [8].

THEOREM 5.1. *Suppose the nominal sampled-data closed-loop operator* $\mathbf{M} \in \mathcal{L}_{\mathcal{A}_\mathbb{D}}$, *and the frequency response* $M(e^{j\omega_o})$ *is a compact operator at each frequency* $\omega_o \in (-\pi,\pi]$. *Then, for every* $0 < \rho < 1$, *the system of Figure* 3.3 *has uniform robust stability against perturbations in* $\rho\mathcal{UL}_{LTI}(\nu)$ *for some* $\nu > 0$ *if and only if*

$$\inf_{\mathbf{D}\in\mathfrak{D}^u_{LTI}} \|\mathbf{DMD}^{-1}\|_{L_2\to L_2} \le 1.$$

The closed loop operator $\mathbf{M}$ defined in section 3 will always satisfy the supposition of the theorem; see the state space formula of Appendix A. Thus the above result holds for a more general class of periodic systems.

5.1. Necessity. We begin our proof of Theorem 5.1 by showing that the D-scaling condition put forward is necessary in order to guarantee uniform robust stability of our sampled-data system. The result stated next is key. It allows us to convert our original D-scaling condition to a D-scaling condition involving the sampled-data frequency response operator introduced last section.

THEOREM 5.2. *Suppose* $M(e^{j\omega})$ *is continuous and compact at each* $\omega \in \mathbb{R}$. *Then*

$$\inf_{\mathbf{D}\in\mathfrak{D}^u_{LTI}} \|\mathbf{DMD}^{-1}\| = \sup_{\omega\in\mathbb{R}} \inf_{D\in\mathfrak{D}^u} \|DM(e^{j\omega})D^{-1}\|.$$

The proof of this result is located in Appendix B. At first, it appears that there is no distinct advantage to this new formulation since the representations involved remain infinite dimensional. However, the properties of the sampled-data frequency response operator that were introduced in the last section make this new representation more amenable to constructing the required proof. Moreover, the computational framework we develop in [9] is entirely based on the interchangeability of the two representations.

Lemma 5.3 below is the main necessity result that we shall prove in this paper. By Theorem 5.2, it is precisely the contrapositive of the desired result. Now, although we

significantly bolstered our set of tools in the last section, we still require an additional technical result which will help us construct the destabilizing perturbation required in the proof of the theorem. We state the main result here and then briefly digress before returning to the proof in subsection 5.1.2.

LEMMA 5.3. *Suppose* $\mathbf{M} \in \mathcal{L}_{\mathcal{A}_{\mathbb{D}}}$, *and* $M(e^{j\omega_o})$ *is a compact operator at each frequency* $\omega_o \in (-\pi, \pi]$. *If*

$$\sup_{\omega\in(-\pi,\pi]} \inf_{D\in\mathfrak{D}^u} \|DM(e^{j\omega})D^{-1}\|_{\ell_2\to\ell_2} > 1,$$

then the sampled-data system in Figure 3.3 *does not have uniform robust stability to the perturbation sets* $\mathcal{UL}_{LTI}(\nu)$ *for any* $\nu > 0$.

5.1.1. Constructing destabilizing perturbations. Here we present a number of lemmas. The key result of this subsection is concerned with the maximal rate of time variation required for an operator to move power across frequencies while maintaining a set of power inequalities. The basic construction parallels that in [23].

Our first result gives an asymptotic property of an operator in the set $\mathcal{L}_{\mathcal{A}_{\mathbb{C}+}}$ and is nothing more than a special case of an asymptotic frequency response result presented in section 4. Since the result is standard in linear systems theory, no proof is presented.

LEMMA 5.4. *Suppose* $\Omega \in \mathbb{R}$ *and* $\mathbf{Q} \in \mathcal{L}_{\mathcal{A}_{\mathbb{C}+}}$. *Then the following limit is satisfied:*

$$\lim_{q\to\infty} \|\hat{Q}(j\Omega)\phi_\Omega^q - \mathbf{Q}\phi_\Omega^q\| = 0,$$

where the function ϕ_Ω^q *is defined in* (4.6).

Our next lemma is concerned with filtering a signal consisting of $N+1$ countable sets of aliased frequencies. The construction is based on a result first proved by Rudin [24] and the asymptotic property presented in our last result.

LEMMA 5.5. *Suppose* (a) *the frequencies* $-\pi < \omega_0 < \cdots < \omega_N < \omega_0 + \pi$, *and* (b) *the corresponding sequences* $a_0, \ldots, a_N$ *are in* ℓ_2. *Then there exists* $\mathbf{Q} \in \mathcal{L}_{\mathcal{A}_{\mathbb{C}+}}$ *with* $\|\mathbf{Q}\| = 1$ *such that*

$$\lim_{q\to\infty} \|\mathbf{Q}z^q - v^q\| = 0,$$

where $z^q = \sum_{k=0}^N \sum_{l=0}^\infty a_k^l \phi_{\omega_k^l}^q$, $v^q = \sum_{k=0}^N \sum_{l=0}^n a_k^l \phi_{\omega_k^l}^q$, *and* $\omega_k^l = \frac{2\pi\nu_l - \omega_k}{h}$.

Lemma 5.6 is the key result in this section. It provides an upper bound for the rate of change required for an operator $\mathbf{\Delta}$ to move power across frequencies. Once again, due to space limitations, the proof is omitted. We refer the reader to [8].

LEMMA 5.6. *Suppose* (i) $a_k^l, b_k^l \in \mathbb{C}^m$ *for* $0 \le l \le n$ *and* $0 \le k \le N$, (ii) $-\pi < \omega_0 < \cdots < \omega_N < \omega_0 + \pi$, $\omega_i \ne -\omega_j$ *for* $i, j = 0, \ldots, N$, *and* (iii) *for some* $\gamma > 1$, *the finite sequences* a^l *and* b^l *satisfy* $|a^l|_2 \ge \gamma |b^l|_2$ *for each* $0 \le l \le n$. *Then there exists* $\mathbf{\Delta} \in \mathcal{UL}_{LTI}(\nu_o)$, *with* $\nu_o := \frac{\omega_N - \omega_0}{h}$, *such that*

$$\lim_{q\to\infty} \|\mathbf{\Delta}z^q - w^q\| = 0,$$

where $z^q = \sum_{k=0}^N \sum_{l=0}^n a_k^l \phi_{\omega_k^l}^q$, $w^q = \sum_{k=0}^N \sum_{l=0}^n b_k^l \phi_{\omega_k^l}^q$, *and* $\omega_k^l = \frac{2\pi\nu_l - \omega_k}{h}$.

Remark 5.7. The convention introduced in the statement of Lemma 5.6 shall be used throughout the remainder of the paper. Namely, given a sequence of sequences, $a = (a_0, a_1, \ldots,)$, we shall use a_k^l to denote the lth component of the kth sequence.

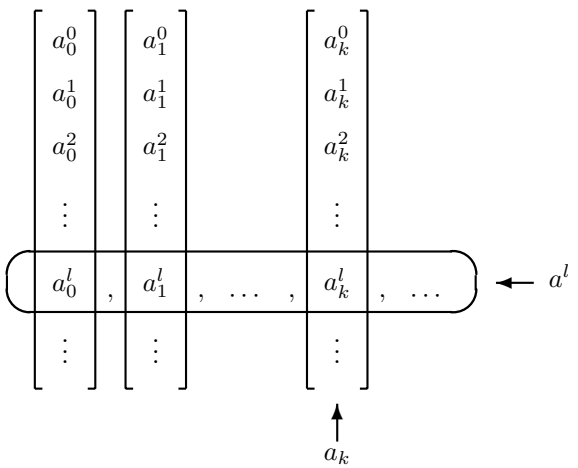


FIG. 5.1. *Typical element of* $\ell_2(\ell_2)$.

Also, we will denote the sequence $(a_0^l, a_1^l, a_2^l, \dots,)$ by a^l. Figure 5.1 illustrates the use of this notation when $a \in \ell_2(\ell_2)$.

Remark 5.8. In the hypotheses of Lemma 5.6, we have assumed that the frequencies $\omega_0, \dots, \omega_N$ are chosen so that $\omega_i \neq -\omega_j$. Although this assumption is never explicitly used in our construction, it is necessary to have it in place when the perturbation $\mathbf{\Delta}$ is required to map real signals back onto real signals.

Corollary 5.9 below is the result needed to construct the destabilizing perturbation required in the proof of Lemma 5.3.

COROLLARY 5.9. *Suppose* (i) $a_k, b_k \in \ell_2$ *for* $0 \leq k \leq N$, (ii) $-\pi < \omega_0 < \cdots < \omega_N < \omega_0 + \pi$, $\omega_i \neq -\omega_j$ *for* $i, j = 0, \dots, N$, *and* (iii) *there exists* $\gamma > 1$ *such that for each* $0 \leq l \leq n$, *we have* $|a^l|_2 \geq \gamma |b^l|_2$. *Then there exists* $\mathbf{\Delta} \in \mathcal{UL}(\nu_o)$ *with* $\nu_o = \frac{\omega_N - \omega_0}{h}$, *such that*

$$\lim_{q \to \infty} \|\mathbf{\Delta} z^q - w^q\| = 0,$$

where $z^q = \sum_{k=0}^{N} \sum_{l=0}^{\infty} a_k^l \phi_{\omega_k^l}^q$, $w^q = \sum_{k=0}^{N} \sum_{l=0}^{n} b_k^l \phi_{\omega_k^l}^q$, *and* $\omega_k^l = \frac{2\pi\nu_l - \omega_k}{h}$.

Having established this last result, we are now in a position to prove Lemma 5.3, which we do next.

5.1.2. Proof of Lemma 5.3. Choose any $\nu_o > 0$. It is sufficient to show that given this choice of ν_o and any $\varepsilon > 0$, we can construct a perturbation $\mathbf{\Delta} \in \mathcal{UL}_{LTI}(\nu_o)$ and a corresponding signal $z \in L_2$ of unit norm such that $\|(\mathbf{I} - \mathbf{M}\mathbf{\Delta})z\| < \varepsilon$.

Let $\varepsilon > 0$. By hypothesis, there exists a $\theta_o \in (-\pi, \pi]$ so that

$$\inf_{D \in \mathfrak{D}^u} \|DM(e^{j\theta_o})D^{-1}\| > 1. \tag{5.1}$$

Next, choose $N \in \mathbb{N}_o$ large enough so that

$$\|P_N M(e^{j\theta_o}) - M(e^{j\theta_o})\| < \frac{\varepsilon}{3\|\mathbf{M}\|}, \tag{5.2}$$

where the projection operator $P_n = I - Q_n$ is defined by

$$P_n(a_0, \dots, a_n, a_{n+1}, \dots) = (a_0, \dots, a_n, 0, 0, \dots)$$

for each $a \in \ell_2$ and $n \in \mathbb{N}_o$. Note that by compactness of $M(e^{j\theta_o})$, such an integer N always exists [15]. Moreover, by continuity of $M(\cdot)$, there exists an interval I about θ_o such that $\|P_N M(e^{j\omega}) - M(e^{j\omega})\| < \frac{\varepsilon}{3\|\mathbf{M}\|}$ for all $\omega \in I$.

Now recall our earlier work on the S-procedure. By Corollary 4.8, condition (5.1) implies that there exist $\gamma > 1$ and $b \in \ell_2(\ell_2)$ with $\|b\| = 1$ so that by defining $a' := T_M b$, we have

$$\gamma \|T_{E_l} b\|_2 \le \|T_{E_l} a'\|_2 = \sqrt{\sum_{k=0}^{\infty} \|E_l M(e^{j\theta_o}) b_k\|_2^2} \tag{5.3}$$

for $l = 0, \dots, N$ and

$$\gamma \|T_{Q_N} b\|_2 \le \|T_{Q_N} a'\|_2 = \sqrt{\sum_{k=0}^{\infty} \|Q_N M(e^{j\theta_o}) b_k\|_2^2}. \tag{5.4}$$

Since T_M is a memoryless operator, we can assume, without loss of generality, that (5.3) and (5.4) are satisfied for b with finite support. Let $K+1$ be the support length of b.

Using (5.3), (5.4), and the continuity of $M(e^{j\omega})$, we can choose $K+1$ distinct frequencies $-\pi < \omega_0 < \cdots < \omega_K < \omega_0 + \pi$, each in the interval I, so that by defining $a_k := M(e^{j\omega_k}) b_k$, we have

(i) $\dfrac{\omega_K - \omega_0}{h} < \nu_o$,

(ii) $\omega_i \neq -\omega_j$ for all $i, j = 0, \dots, K$,

(iii) $$\gamma' \sqrt{\sum_{k=0}^{K} \|E_l b_k\|_2^2} = \gamma' \|T_{E_l} b\|_2 \le \|T_{E_l} a\|_2 = \sqrt{\sum_{k=0}^{K} \|E_l M(e^{j\omega_k}) b_k\|_2^2} \tag{5.5}$$
for each $l = 0, \dots, N$, and

(iv) $$\gamma' \sqrt{\sum_{k=0}^{K} \|Q_N b_k\|_2^2} = \gamma' \|T_{Q_N} b\|_2 \le \|T_{Q_N} a\|_2 = \sqrt{\sum_{k=0}^{K} \|Q_N M(e^{j\omega_k}) b_k\|_2^2} \tag{5.6}$$

for some $1 < \gamma' < \gamma$. Furthermore, without loss of generality, we can also assume that the frequencies $\omega_0, \dots, \omega_K$ are all rational numbers.

We now seek to make use of Corollary 5.9 in order to construct our destabilizing perturbation. Let

$$z^q := \sum_{k=0}^{K} \sum_{l=0}^{\infty} a_k^l \phi_{\omega_k^l}^q, \; u^q := \sum_{k=0}^{K} \sum_{l=0}^{\infty} b_k^l \phi_{\omega_k^l}^q, \text{ and } w^q := \sum_{k=0}^{K} \sum_{l=0}^{N} b_k^l \phi_{\omega_k^l}^q.$$

Henceforth, we shall also assume for simplicity that the integer q in the above definitions is always chosen from the set $\Omega := \{n \in \mathbb{N}_o : n\omega_k \in \mathbb{Z} \text{ for } k = 1, \dots, K\}$. Under

this assumption and the conditions given in (5.6) and (5.2), we then have that

$$\begin{aligned}\|u^q - w^q\|_{L_2}^2 &= \sum_{k=0}^{K} \sum_{l=N+1}^{\infty} \|b_k^l\|_2^2 \\ &= \sum_{k=0}^{K} \|Q_N b_k\|_{\ell_2}^2 \\ &\leq \sum_{k=0}^{K} \|Q_N M(e^{j\omega_k})\|_{\ell_2 \to \ell_2}^2 \cdot \|b_k\|_{\ell_2}^2 \\ &< \left(\frac{\varepsilon}{3\|\mathbf{M}\|}\right)^2 \sum_{k=0}^{K} \|b_k\|_{\ell_2}^2 \\ &= \left(\frac{\varepsilon}{3\|\mathbf{M}\|}\right)^2, \end{aligned} \tag{5.7}$$

since $\|b\| = 1$.

We will now show that $\|\mathbf{M}\boldsymbol{\Delta} z^q - z^q\| < \varepsilon$ for $q \in \Omega$ large enough in order to complete this part of the proof. By Corollary 5.9 and (5.5), we know there exists $\boldsymbol{\Delta} \in \mathcal{UL}_{LTI}(\nu_o)$ so that $\|\boldsymbol{\Delta} z^q - w^q\| \longrightarrow 0$ as $q \to \infty$. Using (5.7) and the triangle inequality, we find that for q sufficiently large $\|\boldsymbol{\Delta} z^q - u^q\| < \frac{\varepsilon}{3\|\mathbf{M}\|}$. The submultiplicative inequality then implies that

$$\|\mathbf{M}\boldsymbol{\Delta} z^q - \mathbf{M}u^q\| < \frac{\varepsilon}{3} \tag{5.8}$$

for $q \in \Omega$ sufficiently large. Also, from the definition of w^q and Lemma 4.2, we see that

$$\left\| \mathbf{M}w^q - \sum_{k=0}^{K} \sum_{l=0}^{N} \sum_{p=0}^{\infty} (M_{p,l}(e^{j\omega_k}) b_k^l) \phi^q_{\omega_k^p} \right\| \longrightarrow 0 \text{ as } q \to \infty.$$

With the help of the triangle and submultiplicative inequalities, it is not difficult to use (5.7) to deduce that

$$\left\| \mathbf{M}u^q - \sum_{k=0}^{K} \sum_{l=0}^{N} \sum_{p=0}^{\infty} (M_{p,l}(e^{j\omega_k}) b_k^l) \phi^q_{\omega_k^p} \right\| < \frac{\varepsilon}{3} \tag{5.9}$$

for q sufficiently large. Finally, from the definition of z^q, we have

$$z^q = \sum_{k=0}^{K} \sum_{l=0}^{\infty} E_l M(e^{j\omega_k}) b_k \phi^q_{\omega_k^l} = \sum_{k=0}^{K} \sum_{p=0}^{\infty} \sum_{l=0}^{\infty} M_{l,p}(e^{j\omega_k}) b_k^p \phi^q_{\omega_k^l},$$

from which it follows that

$$\begin{aligned}\left\| z^q - \sum_{k=0}^{K}\sum_{l=0}^{N}\sum_{p=0}^{\infty}(M_{p,l}(e^{j\omega_k})b_k^l)\phi^q_{\omega_k^p}\right\|_{L_2}^2 &= \left\| \sum_{k=0}^{K}\sum_{l=N+1}^{\infty}\sum_{p=0}^{\infty}(M_{p,l}(e^{j\omega_k})b_k^l)\phi^q_{\omega_k^p}\right\|_{L_2}^2 \\ &= \sum_{k=0}^{K}\|M(e^{j\omega_k})Q_N b_k\|_{\ell_2}^2 \\ &\le \|\mathbf{M}\|^2\sum_{k=0}^{K}\|Q_N b_k\|_{\ell_2}^2 \\ &< \left(\frac{\varepsilon}{3}\right)^2, \qquad (5.10)\end{aligned}$$

just as in the derivation of (5.7).

Using (5.8), (5.9), and (5.10) along with the triangle inequality, we see that

$$\|\mathbf{M}\boldsymbol{\Delta}z^q - z^q\| < \varepsilon$$

for q sufficiently large. Now set $z := \frac{z^q}{\|z^q\|}$, where q is chosen as above. Note that since $\|b\| = 1$, (5.5) and (5.6) guarantee that $1 \le \|z^q\| \le \infty$ and hence that $\|\mathbf{M}\boldsymbol{\Delta}z - z\| < \varepsilon$, as required. □

5.2. Sufficiency. Having established the necessity of our robustness condition, we now focus on proving that the inequality stated in Theorem 5.1 is sufficient in order to guarantee uniform robust stability against perturbations in the class $\mathcal{L}_{LTI}(\nu)$ for some $\nu > 0$.

Our aim in this part is to prove Lemma 5.10, which is a generalization of a similar result from [23].

LEMMA 5.10. *Suppose* $\mathbf{M} \in \mathcal{L}_{\mathcal{A}_{\mathbb{D}}}$. *If*

$$\inf_{\mathbf{D}\in\mathfrak{D}^u_{LTI}} \|\mathbf{D}\mathbf{M}\mathbf{D}^{-1}\| < 1,$$

then the sampled-data system of Figure 3.3 *has uniform robust stability to perturbations in the set* $\mathcal{UL}_{LTI}(\nu)$ *for some* $\nu > 0$.

In proving our result, we will introduce a second definition for the class of quasi-LTI operators and prove that it is weaker than our original definition. This new class, which is introduced below, is more technically convenient to work with in our context. We should also point out that Lemma 5.10 is not specific to our sampled-data arrangement since $\mathbf{M}$ can be any arbitrarily LTV operator.

The proof of the above theorem is rendered much simpler if we choose to work with an expanded class of quasi-LTI operators. Namely, define the set

$$\mathcal{P}(\varepsilon) := \{\boldsymbol{\Delta} \in L_2^m : \|\boldsymbol{\Theta}\boldsymbol{\Delta} - \boldsymbol{\Delta}\boldsymbol{\Theta}\| \le \varepsilon,\ \boldsymbol{\Delta} \text{ causal}\},$$

where $\varepsilon > 0$ and $\boldsymbol{\Theta} \in \mathcal{L}(L_2^m)$ is the operator whose transfer function representation in $\mathcal{A}_{\mathbb{C}^+}$ is $\hat{\Theta}(s) = \frac{1-s}{1+s}$. Lemma 5.11 below states that this new definition is in fact weaker than our first.

LEMMA 5.11. *Given* $\varepsilon > 0$, *there exists* $\nu > 0$ *such that*

$$\mathcal{UL}_{LTI}(\nu) \subseteq \mathcal{UP}(\varepsilon).$$

Proof. We sketch the details for the proof of the above result here and refer the reader to [8] for the complete version. The first step in proving that our new definition for the class of slowly time-varying operators is weaker than the first is to notice that given an $\varepsilon > 0$, the inequality $\|\mathbf{\Theta\Delta} - \mathbf{\Delta\Theta}\| \leq \varepsilon$ holds if and only if $\|\mathbf{\Lambda\Delta} - \mathbf{\Delta\Lambda}\| \leq \varepsilon$, where $\mathbf{\Lambda} \in \mathcal{L}(L_2)$ represents convolution with the function $\lambda(t) := 2e^{-t}$. Now, given $\delta > 0$ and an input $u \in L_2$, we can choose $N \in \mathbb{N}_{\mathrm{o}}$ and $\tau > 0$ large enough so that the inequality $\|\mathbf{\Psi}_{N,\tau}u - \mathbf{\Lambda}u\| < \delta$ is satisfied. The operator $\mathbf{\Psi}_{N,\tau}$ simply represents convolution with the piecewise constant function $\psi_{N,\tau}(t) := 2\sum_{k=0}^{N-1} e^{-k\tau} w_{k\tau,\tau}(t)$, where the family of window functions $w_{T,\tau}$ is defined by

$$w_{T,\tau}(t) := \begin{cases} 1, & T < t < T + \tau, \\ 0 & \text{otherwise} \end{cases}$$

for $T, \tau > 0$. The idea behind this construction can easily be seen through Figure 5.2 below.

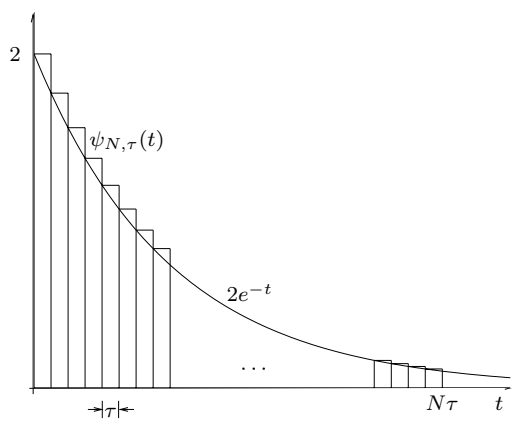


FIG. 5.2. *Piecewise constant function approximation to* $\lambda(t) = 2e^{-t}$.

Now, without loss of generality, we could have chosen the above τ small enough so that the inequality $\|\mathbf{W}_{T,\tau}u - \mathbf{D}_T u\| < \frac{\delta}{4}$ also holds independent of the value of $T > 0$. Here $\mathbf{W}_{T,\tau}$ represents a convolution with the function $\frac{1}{\tau} w_{T,\tau}(t)$, and $\mathbf{D}_T$ is the T-shift on L_2. Having made this choice, we can be assured that $\|\mathbf{\Xi}u - \mathbf{\Psi}_{N,\tau}u\| < \delta$, where the operator $\mathbf{\Xi}$ is defined by $\mathbf{\Xi} := 2\sum_{k=0}^{N-1} \tau e^{-k\tau} \mathbf{D}_{k\tau}$.

From the definition of the set $\mathcal{L}_{LTI}(\nu)$, we can conclude that if $\mathbf{\Delta} \in \mathcal{UL}_{LTI}(\nu)$, then $\|\mathbf{D}_T\mathbf{\Delta} - \mathbf{\Delta}\mathbf{D}_T\| < \nu T$ for every $T > 0$. Let $\delta = \frac{\varepsilon}{5}$. Then, using the above fact, along with the triangle and submultiplicative inequalities, it can be shown that $\|\mathbf{\Theta\Delta} - \mathbf{\Delta\Theta}\| \leq \frac{4\varepsilon}{5} + 2\nu$. Now choose $0 < \nu < \frac{\varepsilon}{10}$ in order to complete the proof. ∎

Finally, we present a technical result which is used in the proof of Lemma 5.10. Its proof is straightforward and is hence omitted.

LEMMA 5.12. *If* $\mathbf{\Delta} \in \mathcal{UP}(\varepsilon)$, *then* $\|\mathbf{\Theta}^k\mathbf{\Delta} - \mathbf{\Delta}\mathbf{\Theta}^k\| \leq k\varepsilon$ *for any* $k \in \mathbb{N}_{\mathrm{o}}$.

Proof of Lemma 5.10. Here we assume that $\mathbf{M}$ is not the zero operator on L_2^m; otherwise, the result is trivial. By hypothesis, there exists $\mathbf{D} \in \mathfrak{D}_{LTI}^u$ satisfying the inequality $\|\mathbf{DMD}^{-1}\| < 1$. Choose $\eta > 0$ so that $\|\mathbf{DMD}^{-1}\| + \eta < 1$.

By definition of the set $\mathfrak{D}^u_{LTI}$, we know that the operator $\mathbf{D}^{-1}$ has a corresponding transfer function $\hat{D}^{-1} \in \mathcal{A}_{\mathbb{C}^+}$. Now functions in the half-plane algebra are isomorphic to those in the disc algebra via the bilinear transformation $z = \frac{1-s}{1+s}$. By Proposition 2.1, we can choose scalars $x_0, x_1, \ldots, x_n$ so that, for any $\eta > 0$, we have

$$\left\| \mathbf{D}^{-1} - \sum_{k=0}^{n} x_k \mathbf{\Theta}^k \right\| = \left\| \hat{D}^{-1} - \sum_{k=0}^{n} x_k \left(\frac{1-s}{1+s} \right)^k \right\|_\infty < \eta,$$

provided, of course, that we choose n large enough.

Set $\mathbf{T}_n = \sum_{k=0}^{n} x_k \mathbf{\Theta}^k$, and choose n sufficiently large so that

$$\|\mathbf{D}^{-1} - \mathbf{T}_n\| < \frac{1}{3}\eta \, \frac{1}{\|\mathbf{DM}\|}.$$

Then $\|\mathbf{DMD}^{-1} - \mathbf{DMT}_n\| \leq \|\mathbf{DM}\| \, \|\mathbf{D}^{-1} - \mathbf{T}_n\| < \frac{1}{3}\eta$. Choose $\varepsilon_{\mathrm{o}} > 0$ so that $\|\mathbf{DM}\| \, \varepsilon_{\mathrm{o}} \sum_{k=1}^{n} k|x_k| < \frac{1}{3}\eta$. Since $\|\mathbf{\Delta}\| \leq 1$ for any $\mathbf{\Delta} \in \mathcal{UP}(\varepsilon_{\mathrm{o}})$, we deduce that

$$\|\mathbf{DMD}^{-1}\mathbf{\Delta} - \mathbf{DMT}_n\mathbf{\Delta}\| < \frac{1}{3}\eta \, \|\mathbf{\Delta}\| \leq \frac{1}{3}\eta. \tag{5.11}$$

Using our last lemma, we find that the following inequalities hold for all perturbations $\mathbf{\Delta} \in \mathcal{UP}(\varepsilon_{\mathrm{o}})$:

$$\begin{aligned} \|\mathbf{T}_n\mathbf{\Delta} - \mathbf{\Delta}\mathbf{T}_n\| &= \Big\| \sum_{k=0}^{n} x_k \mathbf{\Theta}^k \mathbf{\Delta} - \mathbf{\Delta} \sum_{k=0}^{n} x_k \mathbf{\Theta}^k \Big\| \\ &\leq \sum_{k=1}^{n} |x_k| \, \|\mathbf{\Theta}^k\mathbf{\Delta} - \mathbf{\Delta}\mathbf{\Theta}^k\| \\ &\leq \varepsilon_{\mathrm{o}} \sum_{k=1}^{n} k|x_k| \\ &< \frac{1}{3}\eta \, \frac{1}{\|\mathbf{DM}\|}, \end{aligned}$$

where the last inequality follows by our choice of $\varepsilon_{\mathrm{o}} > 0$.

The submultiplicative inequality then allows us to conclude that

$$\|\mathbf{DMT}_n\mathbf{\Delta} - \mathbf{DM\Delta T}_n\| \leq \|\mathbf{DM}\| \, \|\mathbf{T}_n\mathbf{\Delta} - \mathbf{\Delta}\mathbf{T}_n\| < \frac{1}{3}\eta. \tag{5.12}$$

Using (5.11), (5.12), and the triangle inequality, we have

$$\|\mathbf{DMD}^{-1}\mathbf{\Delta} - \mathbf{DM\Delta T}_n\| < \frac{2}{3}\eta. \tag{5.13}$$

Notice from our definition of $\mathbf{T}_n$ that we have

$$\|\mathbf{DM\Delta T}_n - \mathbf{DM\Delta D}^{-1}\| \leq \|\mathbf{DM}\| \, \|\mathbf{\Delta}\| \, \|\mathbf{D}^{-1} - \mathbf{T}_n\| \leq \frac{1}{3}\eta. \tag{5.14}$$

Finally, using (5.13), (5.14), and the triangle inequality, we see that

$$\big| \|\mathbf{DMD}^{-1}\mathbf{\Delta}\| - \|\mathbf{DM\Delta D}^{-1}\| \big| \leq \|\mathbf{DMD}^{-1}\mathbf{\Delta} - \mathbf{DM\Delta D}^{-1}\| < \eta.$$

Thus

$$\|\mathbf{DM\Delta D}^{-1}\| < \|\mathbf{DMD}^{-1}\mathbf{\Delta}\| + \eta \leq \|\mathbf{DMD}^{-1}\| + \eta < 1$$

by our choice of η. Since

$$\mathrm{rad}(\mathbf{M\Delta}) = \mathrm{rad}(\mathbf{DM\Delta D}^{-1}) \leq \|\mathbf{DM\Delta D}^{-1}\| < 1,$$

we conclude that $\mathbf{I} - \mathbf{M\Delta}$ is an invertible operator on L_2 for every $\mathbf{\Delta} \in \mathcal{UP}(\varepsilon_o)$ with ε_o as chosen above. Finally, by Lemma 5.11, we know there exists $\nu > 0$ such that $\mathbf{I} - \mathbf{M\Delta}$ is invertible for every $\mathbf{\Delta} \in \mathcal{UL}_{LTI}(\nu)$.

In order to complete our proof, we need only show that the family of maps $(\mathbf{I} - \mathbf{M\Delta})^{-1}$ is uniformly bounded for all $\mathbf{\Delta} \in \mathcal{UL}_{LTI}(\nu)$. From an earlier part of the proof, we already know that by our choice of $\mathbf{D}$, $\|\mathbf{DM\Delta D}^{-1}\| =: \beta < 1$ for every $\mathbf{\Delta} \in \mathcal{UL}_{LTI}(\nu)$. Using a Neumann series expansion (see, for instance, [19]), we can conclude that $\|\mathbf{D}^{-1}(\mathbf{I} - \mathbf{M\Delta})^{-1}\mathbf{D}\| = \|(\mathbf{I} - \mathbf{DM\Delta D}^{-1})^{-1}\| \leq \frac{1}{1-\beta}$ for every $\mathbf{\Delta} \in \mathcal{UL}_{LTI}(\nu)$. By appealing to the submultiplicative inequality, we then find that

$$\begin{aligned}\|(\mathbf{I} - \mathbf{M\Delta})^{-1}\| &= \|\mathbf{DD}^{-1}(\mathbf{I} - \mathbf{M\Delta})^{-1}\mathbf{DD}^{-1}\| \\ &\leq \|\mathbf{D}\| \cdot \|\mathbf{D}^{-1}(\mathbf{I} - \mathbf{M\Delta})^{-1}\mathbf{D}\| \cdot \|\mathbf{D}^{-1}\| \\ &\leq \frac{1}{1-\beta} \cdot \|\mathbf{D}\| \cdot \|\mathbf{D}^{-1}\|\end{aligned}$$

for every $\mathbf{\Delta} \in \mathcal{UL}_{LTI}(\nu)$. Finally, since both $\|\mathbf{D}\|$ and $\|\mathbf{D}^{-1}\|$ are finite, our proof is complete. □

6. Conclusions and future considerations. This paper establishes the theoretical framework for the analysis of quasi-LTI uncertainty in sampled-data systems, and the main contribution of this paper was to provide an exact characterization of uniform robust performance against the set of quasi-LTI perturbations. Having completed this analysis, we now make a few concluding remarks.

Computation of the stability radius, from the conditions presented here, of a given sampled-data system subject to quasi-LTI uncertainty is an important related problem for applying the methods of this paper. Theorem 5.2 is used as the starting point in [9], where we develop a framework for obtaining upper and lower bounds to the stability radius using only convex matrix calculations.

Although the emphasis was placed on sampled-data systems in this paper, we should mention once again that many of the results we present apply to a larger class of periodic continuous time systems. Specifically, suppose an h-periodic (closed-loop) system is specified by an exponentially stable realization $(A(t), B(t), C(t), D(t) = 0)$ with all these functions being bounded and h-periodic. Then, via lifting, it is possible to explicitly write an equivalent frequency response representation $M(e^{j\omega})$ for the system, which has the same continuity properties required in this paper. Further, at each frequency, $M(e^{j\omega})$ is the sum of a finite rank operator and an integral operator; this integral operator is necessarily compact, and thus the analysis of this paper immediately generalizes.

As a comment, we conjecture that uniform robust performance against quasi-LTI perturbations is equivalent to robust performance in the context of the sampled-data systems studied here.

It should also be noted that the procedure outlined in Appendix B can serve as a guideline for constructing appropriate D-scaling operators in a process such as D-K iteration, providing a robust synthesis heuristic similar to that for standard LTI systems.

Finally, the results of this paper, together with previously published work, provide a complete set of tests for *all* of the standard types of dynamic uncertainty sets encountered in robust control analysis. We believe that the tools and techniques developed in this body of research will allow for a straightforward generalization of the *integral quadratic constraint* (IQC) framework to sampled-data and multirate systems.

Appendix A. State space formulae. In this appendix, we provide state space realizations for several representations of the operator $\mathbf{M}$. In particular, we make explicit the operators $\tilde{M}$, $\tilde{M}(e^{j\omega})$, and $M(e^{j\omega_\circ})$. The details concerning the derivation of the formulae can be found in [12] or in [8].

Recall from section 3 that our plant $\mathbf{G}$ has been conformably partitioned to satisfy equations of the form

$$\begin{bmatrix} z \\ y \end{bmatrix} = \begin{bmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} \\ \mathbf{G}_{21} & \mathbf{G}_{22} \end{bmatrix} \begin{bmatrix} w \\ u \end{bmatrix}.$$

We also provided a state space realization for $\mathbf{G}$,

$$\hat{G}(s) = \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & 0 & D_{12} \\ C_2 & 0 & 0 \end{array}\right],$$

while the matrices $(A_{K_d}, B_{K_d}, C_{K_d}, D_{K_d})$ constituted a minimal realization for our discrete time controller $\mathbf{K}_d$. Using the above information and referring back to Figure 3.1, we see that the equations

$$\begin{aligned} z &= \mathbf{M}w, \\ u &= \mathbf{H}\mathbf{K}_d\mathbf{S}y \end{aligned}$$

represent the behavior of our sampled-data system.

The procedure for obtaining $\tilde{M}$ now proceeds as follows. First, provide a state space realization for the last system of equations and integrate over the interval $[kh, kh+\tau)$. Next, use the procedure described in section 4.1 to lift the input and output signals to the operator $\mathbf{M}$. Set $\tilde{w} = \mathbf{W}w$ and $\tilde{z} = \mathbf{W}z$, where $\mathbf{W}$ is the operator defined by (4.1). Recall that we are also using the shorthand $\tilde{w}[k]$, $\tilde{z}[k]$ to represent the functions $(\tilde{w}[k])(\tau) = w(\tau + kh)$ and $(\tilde{z}[k])(\tau) = z(\tau + kh)$, respectively, for $\tau \in [0, h)$. These definitions in hand, it can be shown that

$$\begin{aligned} \begin{bmatrix} x_G((k+1)h) \\ x_{K_d}[k+1] \end{bmatrix} &= A_d \begin{bmatrix} x_G(kh) \\ x_{K_d}[k] \end{bmatrix} + \check{B}\tilde{w}[k], \qquad \begin{bmatrix} x_G(0) \\ x_{K_d}[0] \end{bmatrix} = 0, \\ \tilde{z}[k] &= \check{C} \begin{bmatrix} x_G(kh) \\ x_{K_d}[k] \end{bmatrix} + \check{D}\tilde{w}[k], \end{aligned} \tag{A.1}$$

where the operators $A_d \in \mathbb{C}^{\tilde{n}\times\tilde{n}}$, $\check{B} : \mathcal{K}_2 \to \mathbb{C}^{\tilde{n}}$, $\check{C} : \mathbb{C}^{\tilde{n}} \to \mathcal{K}_2$, and $\check{D} : \mathcal{K}_2 \to \mathcal{K}_2$ are

defined as follows:

$$A_d := \begin{bmatrix} e^{Ah} + \int_0^h e^{A(h-\eta)}\, d\eta B_2 D_{K_d} C_2 & \int_0^h e^{A(h-\eta)}\, d\eta B_2 C_{K_d} \\ B_{K_d} C_2 & A_{K_d} \end{bmatrix},$$

$$\check{B}\psi := \begin{bmatrix} \int_0^h e^{A(h-\eta)} B_1 \psi(\eta)\, d\eta \\ 0 \end{bmatrix},$$

$$(\check{C}\xi)(\tau) := \begin{bmatrix} C_1 e^{A\tau} & C_1 \int_0^h e^{A(\tau-\eta)}\, d\eta B_2 + D_{12} \end{bmatrix} \begin{bmatrix} I & 0 \\ D_{K_d} & C_{K_d} \end{bmatrix} \xi,$$

$$(\check{D}\psi)(\tau) := C_1 \int_0^h e^{A(\tau-\eta)} B_1 \psi(\eta)\, d\eta.$$

The lifted system $\tilde{M} = \mathbf{W}\mathbf{M}\mathbf{W}^{-1}$ is then given by (A.1), whereas the operator $\check{M}(e^{j\omega}) = \check{C} e^{j\omega}(I - e^{j\omega} A_d)^{-1} \check{B} + \check{D}$.

Finally, we expand the operators $\check{B}$, $\check{C}$, and $\check{D}$ with respect to the basis $\{\psi_k\}$ defined in (4.2) in order to obtain the matrices $(\tilde{B})_k := \check{B} J^*_{\omega_o}$, $(\tilde{C})_l := J_{\omega_o} \check{C}$, and $(\tilde{D})_{lk} := J_{\omega_o} \check{D} J^*_{\omega_o}$, which make up the frequency response operator $M(e^{j\omega_o})$. They are

$$(\tilde{B})_k = h^{-1/2} \begin{bmatrix} I \\ 0 \end{bmatrix} e^{Ah} \int_0^h e^{(j\theta_k I - A)\tau}\, d\tau B_1,$$

$$(\tilde{C})_l = h^{-1/2} \begin{bmatrix} C_1 & D_{12} \end{bmatrix} \int_0^h \exp\left(\begin{bmatrix} A - j\theta_l I & B_2 \\ 0 & -j\theta_l I \end{bmatrix} \tau \right) d\tau \begin{bmatrix} I & 0 \\ D_{K_d} C_2 & C_{K_d} \end{bmatrix},$$

$$(\tilde{D})_{lk} = h^{-1} C_1 \int_0^h e^{(A - j\theta_l I)\tau} \int_0^h e^{(j\theta_k I - A)\eta}\, d\eta\, d\tau B_1,$$

where the frequency $\theta_k = \frac{2\pi\nu_k - \omega_o}{h}$ and ν_k is the sequence $\{0, 1, -1, 2, -2, \dots\}$.

Appendix B. Proof of Theorem 5.2. We now concern ourselves with the proof of Theorem 5.2. Of course this means that we will be dealing with the class of unstructured operators only. The proof presented here can be changed to deal with the class of structured operators with modest technical difficulty. We begin by restating the result for convenience.

THEOREM 5.2. *The following equality holds:*

$$\sup_{\omega \in (-\pi, \pi]} \inf_{D \in \mathfrak{D}^u} \|DM(e^{j\omega})D^{-1}\| = \inf_{\mathbf{D} \in \mathfrak{D}^u_{LTI}} \|\mathbf{D}\mathbf{M}\mathbf{D}^{-1}\|.$$

An outline of the proof of the result spans this entire appendix. Notice that a number of lemmas are pursued within the proof. Their purpose is to help divide the presentation into more manageable pieces. We refer the interested reader to [8] for additional details concerning the omitted proofs.

Proof. Since the frequency response $D(e^{j\omega})$ of a D-scaling operator $\mathbf{D} \in \mathfrak{D}^u_{LTI}$ takes its values in the set $\mathfrak{D}^u$, it readily follows that

$$\begin{aligned} \inf_{\mathbf{D} \in \mathfrak{D}^u_{LTI}} \|\mathbf{D}\mathbf{M}\mathbf{D}^{-1}\| &= \inf_{D \in \mathfrak{D}^u_{LTI}} \sup_{\omega \in (-\pi, \pi]} \|D(e^{j\omega})M(e^{j\omega})D^{-1}(e^{j\omega})\| \\ &\geq \sup_{\omega \in (-\pi, \pi]} \inf_{\mathbf{D} \in \mathfrak{D}^u_{LTI}} \|D(e^{j\omega})M(e^{j\omega})D^{-1}(e^{j\omega})\| \\ &\geq \sup_{\omega \in (-\pi, \pi]} \inf_{D \in \mathfrak{D}^u} \|DM(e^{j\omega})D^{-1}\|. \end{aligned}$$

The reverse inequality is significantly more challenging to prove. The approach we adopt is based on work found in [6] and the sampled-data framework and tools introduced so far. We begin by constructing a smooth operator-valued function mapping the interval $(-\pi, \pi]$ to $\mathfrak{D}^u$, which satisfies our norm condition. We then use this smooth D-scaling operator in order to define a smooth function on the boundary of the closed right half-plane, whose behavior we approximate by a proper real rational function which lies in the half-plane algebra $\mathcal{A}_{\mathbb{C}^+}$. The D-scaling operator we are ultimately interested in is simply that which has this real rational function as its transfer function representation in the half-plane algebra.

Now, in order to show the inequality, we will prove that given any $\alpha > 0$,

$$\text{if} \sup_{\omega\in(-\pi,\pi]} \inf_{D\in\mathfrak{D}^u} \|DM(e^{j\omega})D^{-1}\| < \alpha, \text{ then } \inf_{\mathbf{D}\in\mathfrak{D}^u_{LTI}} \|\mathbf{DMD}^{-1}\| < \alpha \text{ also.}$$

The first step in validating this assertion is to prove that we can choose a smooth function that satisfies our norm constraint. The construction proceeds in two steps. First, we show that we can partition the interval $(-\pi, \pi]$ into a finite number of subintervals on which the D-scaling function can be chosen constant. In addition, we can choose the D-scales from the set

$$\overline{\mathfrak{D}}_k := \left\{ D = \begin{bmatrix} \Pi_k Y(\Pi_k)^* & 0 \\ 0 & I \end{bmatrix} : Y \in \mathfrak{D}^u \right\}$$

for some choice of $k \in \mathbb{N}_{\mathrm{o}}$, where the truncation operator $\Pi_k : L_2 \to \mathbb{R}_k$ is defined by

$$\Pi_k(a_0, a_1, \dots) := (a_0, \dots, a_k).$$

Namely, the set $\overline{\mathfrak{D}}_k$ is the subset of $\mathfrak{D}^u$ which consists of D-scales in which only the first $k+1$ blocks are not prespecified to be the identity. This process is the subject of the next lemma. Finally, in a second lemma, we smoothly join our constant D-scaling operators to form an infinitely differentiable D-scaling operator on the interval $(-\pi, \pi]$.

LEMMA B.1. *Suppose that* $\sup_{\omega\in(-\pi,\pi]} \inf_{D\in\mathfrak{D}^u} \|DM(e^{j\omega})D^{-1}\| < \alpha$. *Then, for some pair of integers* n *and* k, *there exists a partition*

$$-\pi < a_1 < b_1 < a_2 < \cdots < a_n < b_n < \pi$$

of the interval $(-\pi, \pi]$ *along with a corresponding set of invertible* D*-scaling operators* $D_0, D_0^{-1}, \dots, D_n, D_n^{-1} \in \overline{\mathfrak{D}}_k$ *such that*

$$\|D_l M(e^{j\omega})D_l^{-1}\| < \alpha \text{ is satisfied for } \begin{cases} \omega \in (-\pi, a_1], & l = 0, \\ \omega \in [b_l, a_{l+1}], & l = 1, \dots, n-1, \\ \omega \in [b_l, \pi], & l = n. \end{cases}$$

Moreover, we can choose these frequencies and the D*-scales so that for a fixed* $l = 1, \dots, n$ *the inequalities*

$$\|D_l M(e^{j\omega})D_l^{-1}\| < \alpha \text{ and } \|D_{l-1} M(e^{j\omega})D_{l-1}^{-1}\| < \alpha$$

are simultaneously satisfied for every $\omega \in (a_l, b_l)$.

The construction described in the statement of the lemma is shown pictorially in Figure B.1 below.

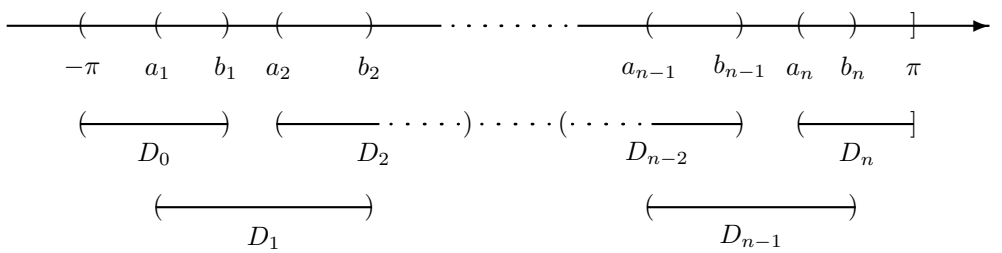


FIG. B.1. *Finite partition of* $(-\pi, \pi]$.

The proof of this first result makes use of the continuity of the sampled-data frequency response operator $M(\cdot)$ and Lemma 4.3. When combined, these two properties allow us to show that the domain of the frequency response operator, the interval $(-\pi, \pi]$, exhibits many properties shared by compact intervals on the real line, which in turn enables the above construction.

The next step in our proof is to smoothly join our newly defined D-scales. We begin by appealing to the function

$$\lambda(t) = \begin{cases} 0, & t \leq 0, \\ e^{-\frac{1}{t}}, & t > 0. \end{cases}$$

It is easy to verify that $\lambda(t)$ is smooth on $\mathbb{R}$. We now use this first definition in order to introduce the "bump" function $\phi_{(a,b)}$ via

$$\phi_{(a,b)}(t) := \frac{\lambda(t-a)}{\lambda(t-a) + \lambda(b-t)}.$$

It is routine to verify that $\phi_{(a,b)}(t) = 0$ for $t \leq a$, $\phi_{(a,b)}(t) = 1$ for $t \geq b$, and that $\phi_{(a,b)}(t)$ is smooth for all $t \in \mathbb{R}$.

We are now in a position to define our smooth D-scaling function. Let $N(e^{j\omega})$ be given by

$$N^2(e^{j\omega}) := \begin{cases} D_0^* D_0, & \omega \in (-\pi, a_1], \\ D_l^* D_l, & \omega \in [b_l, a_{l+1}],\ l = 1, 2, \ldots, n-1, \\ D_n^* D_n, & \omega \in [b_n, \pi], \\ D_{l-1}^* D_{l-1} + (D_l^* D_l - D_{l-1}^* D_{l-1})\phi_{(a_l, b_l)}(\omega), & \\ & \omega \in (a_l, b_l),\ l = 1, 2, \ldots, n, \end{cases} \tag{B.1}$$

where D_l denotes the D-scaling operator defined in the previous lemma for each $l = 0, \ldots, n$. In defining N, we have made use of the bump function in the intervals (a_l, b_l), where two constant D-scales overlap. The function allows us smoothly interpolate the constant D-scales. The process (in the scalar case) is illustrated in Figure B.2.

It is easy to see that such an N is positive and is also a smooth function of ω. Hence we need only demonstrate that our norm constraint still holds using this new function in order to complete our construction. This is done in the following lemma whose proof follows by the definition of N and repeated application of Proposition 4.1.

LEMMA B.2. *Given the function* $N : (-\pi, \pi] \to \mathfrak{D}^u$ *defined in* (B.1)*, the following inequality is satisfied:*

$$\sup_{\omega \in (-\pi, \pi]} \|N(e^{j\omega}) M(e^{j\omega}) N^{-1}(e^{j\omega})\| < \alpha.$$

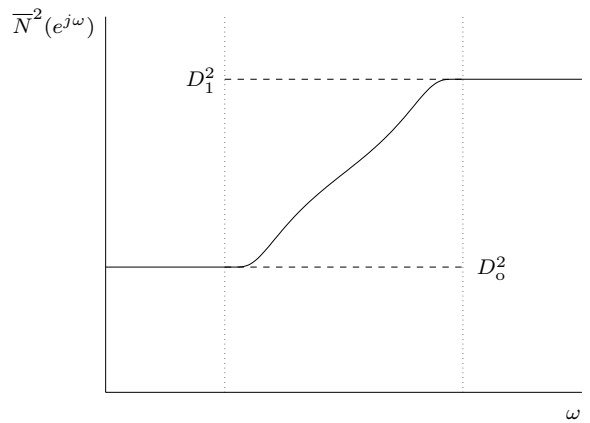


FIG. B.2. *Creating a smooth D-scale.*

The goal now is to show that we can use the function N to help us define a D-scaling operator $\mathbf{D}_o \in \mathfrak{D}^u_{LTI}$ which satisfies the inequality $\|\mathbf{D}_o\mathbf{M}\mathbf{D}_o\|_{L_2\to L_2} < \alpha$. Begin by defining the scalar function $\hat{n} : j\mathbb{R} \to (0,\infty)$ via

$$N(e^{j\omega}) = \operatorname{diag}(\hat{n}(j\omega_0)I_m, \hat{n}(j\omega_1)I_m, \dots), \tag{B.2}$$

where $\omega_k = \frac{2\pi\nu_k - \omega}{h}$ for $\omega \in (-\pi,\pi]$. Using the properties of N, we can easily see that $\hat{n}(j\theta)$ is a positive and bounded function on $j\mathbb{R}$ and that for θ sufficiently large, $\hat{n}(j\theta) = 1$. Moreover, it follows that $\|\hat{n}\|_\infty = \sup_{\omega\in(-\pi,\pi]} \|N(e^{j\omega})\|_{\ell_2\to\ell_2}$. Hence we can conclude that $\hat{n} \in \mathcal{L}_\infty(j\mathbb{R})$, the set of complex-valued functions bounded on the imaginary axis.

Let us now discuss the continuity of $\hat{n}$. Since $N(e^{j\omega})$ is smooth on $(-\pi,\pi]$, we see that $\hat{n}(j\theta)$ is smooth on $[\frac{2\pi\nu_l-\pi}{h}, \frac{2\pi\nu_l+\pi}{h})$ for every $l \in \mathbb{N}_o$. Now, from the definition of N and Proposition 4.3, we have that

$$\lim_{\omega\to-\pi^+} N(e^{j\omega}) = X^*N(-1)X = X^*\begin{bmatrix} \hat{n}(j\frac{2\pi\nu_0-\pi}{h})I_m & & 0 \\ & \hat{n}(j\frac{2\pi\nu_1-\pi}{h})I_m & \\ 0 & & \ddots \end{bmatrix} X,$$

from which it follows that $\hat{n}(j\theta)$ (and hence $N(e^{j\omega})$ also) is smooth on $\mathbb{R}$.

Now, given any function $\mathbf{D} \in \mathfrak{D}^u_{LTI}$, we similarly have [12] that

$$D(e^{j\omega}) = \begin{bmatrix} \hat{d}(j\omega_0)I_m & & 0 \\ & \hat{d}(j\omega_1)I_m & \\ 0 & & \ddots \end{bmatrix},$$

where $D(e^{j\omega})$ is the frequency response of $\mathbf{D}$ and $\hat{d}(s)I_m$ is its transfer function representation in $\mathcal{A}_{\mathbb{C}^+}$. The following lemma says that if $\hat{d}$ closely approximates $\hat{n}$, then $\mathbf{D}$ is the D-scaling operator we require in order to complete the proof.

LEMMA B.3. *Suppose* $\sup_{\omega\in(-\pi,\pi]} \|N(e^{j\omega})M(e^{j\omega})N^{-1}(e^{j\omega})\| < \alpha$, *where* N *is the operator defined in* (B.1). *Then there exists an* $\varepsilon > 0$ *such that if* $\mathbf{D} \in \mathfrak{D}^u_{LTI}$ *and satisfies*

$$\sup_{\theta\in\mathbb{R}} |\hat{d}(j\theta) - \hat{n}(j\theta)| < \varepsilon,$$

we have

$$\sup_{\omega\in(-\pi,\pi]} \|D(e^{j\omega})M(e^{j\omega})D^{-1}(e^{j\omega})\| < \alpha.$$

The proof, which is also omitted, relies on nothing more than the triangle and submultiplicative inequalities along with the fact that N and $\hat{n}$, and D and $\hat{d}$, are isomorphic pairs of operators.

The following lemma is the final result that we present and is easily proved using two important properties of the function space $\mathcal{RL}_\infty(j\mathbb{R})$. This space, consisting of all proper real rational functions with no poles on the imaginary axis, is dense in $\mathcal{L}_\infty(j\mathbb{R})$. Moreover, for any function $\hat{q} \in \mathcal{RL}_\infty(j\mathbb{R})$, there exists a spectral factorization $\hat{q} = \hat{d}^*\hat{d}$, where $\hat{d}$ and $\hat{d}^{-1}$ lie in the space $\mathcal{RH}_\infty$. Using these facts, we can interpolate our function $\hat{n}$ to any desired accuracy using a proper real rational function in the half-plane algebra.

LEMMA B.4. *Given the smooth function $\hat{n}$ defined through* (B.2). *For every $\varepsilon > 0$, there exists a function $\hat{d} \in \mathcal{RH}_\infty$ with $\hat{d}^{-1} \in \mathcal{RH}_\infty$ such that*

$$\left\| |\hat{d}| - \hat{n} \right\|_\infty < \varepsilon.$$

Now, for every $\varepsilon > 0$, there exists an operator $\mathbf{D} \in \mathfrak{D}^u_{LTI}$ such that $\left\| |\hat{d}_\mathrm{o}| - \hat{n} \right\|_\infty < \varepsilon$ by Lemma B.4. Lemma B.3 then says that we can choose the above $\varepsilon > 0$ small enough so that

$$\sup_{\omega\in(-\pi,\pi]} \|D(e^{j\omega})M(e^{j\omega})D^{-1}(e^{j\omega})\| < \alpha,$$

where the operator-valued function

$$D(e^{j\omega}) := \begin{bmatrix} |\hat{d}_\mathrm{o}(j\omega_0)|I_m & & 0 \\ & |\hat{d}_\mathrm{o}(j\omega_1)|I_m & \\ 0 & & \ddots \end{bmatrix}.$$

Let $\mathbf{D}_\mathrm{o}$ be the operator in the space $\mathfrak{D}^u_{LTI}$ whose transfer function representation in the half-plane algebra $\mathcal{A}_{\mathbb{C}^+}$ is $\hat{D}_\mathrm{o} := \hat{d}_\mathrm{o} I_m$. Its frequency response $D_\mathrm{o}(e^{j\omega})$ then satisfies

$$D_\mathrm{o}(e^{j\omega}) = \begin{bmatrix} \hat{d}_\mathrm{o}(j\omega_0)I_m & & 0 \\ & \hat{d}_\mathrm{o}(j\omega_1)I_m & \\ 0 & & \ddots \end{bmatrix} = D(e^{j\omega})\Xi(e^{j\omega}),$$

where $\Xi(e^{j\omega_\mathrm{o}})$ is a unitary operator on ℓ_2 at each frequency $\omega_\mathrm{o} \in (-\pi, \pi]$. It then follows that

$$\|D_\mathrm{o}(e^{j\omega})M(e^{j\omega})D_\mathrm{o}^{-1}(e^{j\omega})\| = \|D(e^{j\omega})M(e^{j\omega})D^{-1}(e^{j\omega})\|$$

for each $\omega \in (-\pi, \pi]$. Finally, using the above definitions, we can see that

$$\begin{aligned} \inf_{\mathbf{D}\in\mathfrak{D}^u_{LTI}} \|\mathbf{DMD}^{-1}\|_{L_2\to L_2} &\le \|\mathbf{D}_\mathrm{o}\mathbf{M}\mathbf{D}_\mathrm{o}^{-1}\|_{L_2\to L_2} \\ &= \sup_{\omega\in(-\pi,\pi]} \|D_\mathrm{o}(e^{j\omega})M(e^{j\omega})D_\mathrm{o}^{-1}(e^{j\omega})\|_{\ell_2\to\ell_2} \\ &= \sup_{\omega\in(-\pi,\pi]} \|D(e^{j\omega})M(e^{j\omega})D^{-1}(e^{j\omega})\|_{\ell_2\to\ell_2} \\ &< \alpha. \end{aligned}$$

Hence our proof is now complete. □

REFERENCES

[1] M. Araki and Y. Ito, *Frequency response of sampled-data systems* I: *Open-loop considerations*, in Proceedings of the IFAC World Congress, Sydney, Australia, 1993, pp. 259–262.

[2] M. Araki and Y. Ito, *Frequency response of sampled-data systems* II: *Closed-loop considerations*, in Proceedings of the IFAC World Congress, Sydney, Australia, 1993, pp. 263–266.

[3] G. Balas, J. Doyle, K. Glover, A. Packard, and R. Smith, *The μ Analysis and Synthesis Toolbox*, Mathworks and MUSYN, Minneapolis, MN, 1991.

[4] B. Bamieh and J. Pearson, *A general framework for linear periodic systems with application to $\mathcal{H}_\infty$ sampled-data control*, IEEE Trans. Automat. Control, 37 (1992), pp. 418–435.

[5] B. Bamieh, J. Pearson, B. Francis, and A. Tannenbaum, *A lifting technique for linear periodic systems with applications to sampled-data control*, Systems Control Lett., 16 (1991), pp. 399–409.

[6] H. Bercovici, C. Foias, and A. Tannenbaum, *Structured interpolation theory*, in Extension and Interpolation of Linear Operators and Matrix Functions, Oper. Theory Adv. Appl. 47, Birkhäuser-Verlag, Basel, 1990, pp. 195–220.

[7] B. Bollobàs, *Linear Analysis*, Cambridge University Press, Cambridge, UK, 1990.

[8] S. Bourdon, *Analysis of Linear Quasi-Time-Invariant Uncertainty in Sampled-Data Systems*, Master's thesis, Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario, Canada, 1997.

[9] S. Bourdon and G. Dullerud, *Computing quasi-LTI robustness margins in sampled-data systems*, IEEE Trans. Automat. Control, 46 (2001), pp. 607–613.

[10] T. Chen, *Control of Sampled-Data Systems*, Ph.D. thesis, Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario, Canada, 1991.

[11] J. Doyle, *Analysis of feedback systems with structured uncertainty*, in Proc. IEE-D, 129 (1982), pp. 242–250.

[12] G. Dullerud, *Control of Uncertain Sampled-Data Systems*, Birkhäuser, Boston, 1996.

[13] G. Dullerud and K. Glover, *Robust stabilization of sampled-data systems to structured LTI perturbations*, IEEE Trans. Automat. Control, 38 (1993), pp. 1497–1508.

[14] G. Dullerud and K. Glover, *Robust performance of periodic systems*, IEEE Trans. Automat. Control, 41 (1996), pp. 1146–1159.

[15] P. Halmos, *Introduction to Hilbert Space*, Chelsea, New York, 1957.

[16] S. Hara, M. Nakajima, and P. Kabamba, *Robust stabilization in digital control systems*, Trans. Soc. Instrument Control Engineers, 128 (1992), pp. 10–19.

[17] E. Kreyszig, *Introductory Functional Analysis with Applications*, John Wiley and Sons, New York, 1989.

[18] A. Megretski and S. Treil, *Power distribution inequalities in optimization and robustness of uncertain systems*, J. Math. Systems Estimation Control, 3 (1993), pp. 301–319.

[19] A. Naylor and G. Sell, *Linear Operator Theory in Engineering and Science*, Springer-Verlag, New York, 1982.

[20] A. Packard, *What's New With μ: Structured Uncertainty in Multivariable Control*, Ph.D. thesis, Department of Mechanical Engineering, University of California, Berkeley, CA, 1988.

[21] A. Packard and J. Doyle, *The complex structured singular value*, Automatica J. IFAC, 29 (1993), pp. 71–109.

[22] F. Paganini and J. Doyle, *Analysis of implicitly defined systems*, in Proceedings of the IEEE Conference on Decision and Control, Lake Buena Vista, FL, IEEE Control Systems Society, Piscataway, NJ, 1994, pp. 3673–3678.

[23] K. Poolla and A. Tikku, *Robust performance against slowly-varying structured perturbations*, in Proceedings of the IEEE Conference on Decision and Control, San Antonio, TX, IEEE Control Systems Society, Piscataway, NJ, 1993, pp. 990–995.

[24] W. Rudin, *Boundary values of continuous analytic functions*, Proc. Amer. Math. Soc., 7 (1956), pp. 808–811.

[25] M. Safonov, *Tight bounds on the response of multivariable systems with component uncertainty*, in the Sixteenth Annual Allerton Conference on Communication, Control, and Computing, University of Illinois at Urbana-Champaign, Monticello, IL, 1978, pp. 451–460.

[26] N. Sivashankar and P. Khargonekar, *Robust stability and performance analysis of sampled-data systems*, IEEE Trans. Automat. Control, 38 (1993), pp. 58–69.

[27] P. Thompson, R. Daily, and J. Doyle, *New conic sectors for sampled-data system feedback systems*, Systems Control Lett., 7 (1986), pp. 395–404.

[28] P. Thompson, G. Stein, and M. Athans, *Conic sectors for sampled-data feedback systems*, Systems Control Lett., 3 (1983), pp. 77–82.

[29] H. Toivonen, *Sampled-Data Control of Continuous-Time Systems with an $\mathcal{H}_\infty$ Optimality Criterion*, Tech. report 90-1, Department of Chemical Engineering, Abo Akademi, Turku, Finland, 1990.

[30] Y. Yamamoto, *A new approach to sampled-data control systems—a function space viewpoint with applications to tracking problems*, in Proceedings of the IEEE Conference on Decision and Control, Honolulu, HI, IEEE Control Systems Society, Piscataway, NJ, 1990, pp. 1882–1887.

[31] Y. Yamamoto and P. Khargonekar, *Frequency response of sampled-data systems*, IEEE Trans. Automat. Control, 41 (1996), pp. 166–176.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 1, pp. 328–332

COMMENT ON AN EXISTENCE RESULT FOR A NONCOERCIVE NONCONVEX VARIATIONAL PROBLEM*

ERIK J. BALDER†

Abstract. An essential improvement is given of a recent existence result of Crasta [G. Crasta, *SIAM J. Control Optim.*, 38 (1999), pp. 237–253] for a nonconvex, noncoercive variational problem whose integrand does not depend on the state variable. This is shown to follow by the methods based on Fatou's lemma in several dimensions of [E. J. Balder, *J. Math. Anal. Appl.*, 101 (1984), pp. 527–539]. The associated Euler–Lagrange inclusions follow from well-known optimality conditions for such problems [V. M. Alekseev, V. M. Tichimirov, and S. V. Fomin, *Optimal Control*, Consultants Bureau, New York, 1987; V. I. Arkin and V. L. Levin, *Uspekhi Mat. Nauk*, 27 (1972), pp. 21–77].

Key words. nonconvex variational problems, existence, Fatou's lemma in several dimensions, measurable selections

AMS subject classifications. 49J15, 49J45

PII. S0363012900371848

1. An existence result.

In [15] Crasta considers the following variational problem:

$$\text{(P)} \qquad \min_{v\in AC([0,R])^d)} \left\{ \int_0^R g(t,v'(t))dt : v(R)=0 \right\}.$$

In Theorem 3.2, the main result of [15], he proves the existence of an optimal solution for problem (P) by means of a complicated proof with several intermediate steps, such as an Euler–Lagrange inclusion for a convexified version of the above problem and an L_∞-truncation procedure (see [15, pp. 245–250]). In [15] this result is subsequently applied to radially symmetric variational problems. To be precise, that main result is as follows. Let $[0,R]$ be equipped with the Lebesgue σ-algebra and Lebesgue measure, and let $\mathbf{R}^2$ have the Borel σ-algebra. Let $g : [0,R]\times\mathbf{R}^d \to (-\infty,+\infty]$ be given as above. Let $g^{**}(t,\cdot)$ $[g^*(t,\cdot)]$ be the Fenchel-biconjugate (Fenchel-conjugate) of $g(t,\cdot)$, and let $\partial g^*(t,p)$ be the subdifferential of $g^*(t,\cdot)$ at the point $p\in\mathbf{R}^d$. The following conditions are needed in [15, Theorem 3.2].

- (G1′) g is product-measurable and such that $g(t,\cdot)$ is lower semicontinuous on $\mathbf{R}^d$ for almost every (a.e.) t in $[0,R]$.
- (G1″) $g^{**}(t,x) < +\infty$ for a.e. t in $[0,R]$ and for all $x\in\mathbf{R}^d$.
- (G2) $g^{**}(\cdot,0)$ is integrable on $[0,R]$.
- (G3) There exist $m>0$ and $b\in\mathcal{L}_1([0,R])$ such that $g^{**}(t,x)\geq m|x|-b(t)$ for a.e. t in $[0,R]$ and for all $x\in\mathbf{R}^d$.
- (G4) For every $r_1>0$ there exists $r_2>0$ such that for a.e. t in $[0,R]$ and every $p\in\mathbf{R}^d$ the following implication holds:

$$\partial g^*(t,p)\cap\{x\in\mathbf{R}^d : |x|\leq r_1\}\neq\emptyset \Rightarrow \partial g^*(t,p)\subset\{x\in\mathbf{R}^d : |x|<r_2\}.$$

- (G5) $g(\cdot,\xi)$ is integrable on $[0,R]$ for every $\xi\in\mathbf{R}^d$.

*Received by the editors May 3, 2000; accepted for publication (in revised form) December 30, 2000; published electronically June 26, 2001.
http://www.siam.org/journals/sicon/40-1/37184.html
†Mathematical Institute, University of Utrecht, P.O. Box 80.010, Budapestlaan 6, 3508 TA Utrecht, The Netherlands (balder@math.uu.nl).

The following theorem improves Theorem 3.2 of [15]. It shows that (G1″) and the technical polar condition (G4) are redundant and that much less is needed in (G5), bringing it exactly in line with (G2). Furthermore, we work with a general dimension d, whereas Crasta's approach really uses $d = 1$. Our approach will also show that problem (P) can be handled directly without the intervention of (P**). The latter is a relaxation of problem (P), obtained by replacing g with g^{**}.

THEOREM 1.1. (i) *Under* (G1′), (G2), *and* (G3) *the problem* (P**) *has an optimal solution* $\bar{v} \in AC([0,R])^d$, $\bar{v}(R) = 0$, *that satisfies the Euler–Lagrange inclusion*

$$0 \in \partial g^{**}(t, \bar{v}'(t)) \text{ for a.e. } t \text{ in } [0,R].$$

(ii) *Moreover, already if* (G5) *just holds for* $\xi = 0$, *then* $\bar{v}$ *satisfies* $g(t,\bar{v}'(t)) = g^{**}(t,\bar{v}'(t))$ *a.e. and is also an optimal solution to* (P).

By the substitution $v(t) = -\int_t^R u(s)ds$, problem (P) is evidently equivalent to the following unconstrained optimal control problem:

$$(\text{CP}) \qquad \min_{u \in \mathcal{L}_1([0,R])^d} \int_0^R g(t,u(t))dt,$$

and a similar reformulation holds for (P**); the latter optimal control problem is indicated by (CP**). Problems (CP) and (CP**) are particular, unconstrained instances of a *Lyapunov-type* optimization problem. Such problems have been studied extensively [1, 3, 5, 6, 7, 9, 11], both for their existence aspects and optimality conditions. Seen from that body of knowledge, Theorem 1.1 is entirely standard. For instance, already by introducing the singular component $\bar{y}_u(t) := \int_0^t g(s,u(s))ds$, the existence part of Theorem 1.1 follows directly from Corollary 2.9 in [7][1] (see also [13] for related results). However, since we wish to stress the general background of problem (P), we have chosen two quite general tools. The first of these is Fatou's lemma in several dimensions; see Theorem 2.1. The second general result is the reduction Theorem 2.2, which is a well-known measurable selection result about "switching infima and integral signs."

2. Proof of Theorem 1. Let $(T,\mathcal{T},\mu)$ be a finite and complete measure space. By $\mathcal{L}_1(T)^m$ we denote the set of all functions from T into $\mathbf{R}^m$ with μ-integrable component functions. The unifying Fatou lemma in several dimensions of [6] is as follows.

THEOREM 2.1 (see [6]). *Suppose* $(f_k)_k \subset \mathcal{L}_1(T)^m$ *is such that* $(\max(0,-f_k^i))_k$ *is uniformly* μ*-integrable,* $i =, \ldots, m$, *and such that* $a := \lim_k \int f_k \, d\mu$ *exists in* $\mathbf{R}^m$. *Then there exists* $f_* \in \mathcal{L}_1(T)^m$ *such that* $\int f_* \, d\mu \le a$ *(coordinatewise) and*

$$f_*(t) \text{ is a limit point of } (f_k(t))_k \text{ for } \mu\text{-a.e. } t \text{ in } T.$$

This result extends similar results in [5, 11, 4, 22]. The following version of the reduction theorem comes from [8, Appendix B], which essentially mimics [12, VII]. Similar results can also be found in [5, 19, 21].

THEOREM 2.2 (see [8, Theorem B.1]). *Let* X *be a Suslin space. For every* $\mathcal{T} \times \mathcal{B}(X)$*-measurable function* $f : T \times X \to [-\infty, +\infty]$ *and every decomposable*[2] *set*

[1] Namely, set $a, \bar{a} \equiv 0$, $b \equiv 0$, $\bar{b} \equiv 1$, $\bar{c} := g$, and $J(u) := \bar{y}_u(R)$ in [7].

[2] That is, for every $A \in \mathcal{T}$, $u \in \mathcal{U}$ and every $(\mathcal{T},\mathcal{B}(X))$-measurable $b : T \to X$ with $b(T) \subset X$ relatively compact, the concatenation $1_A b + 1_{T \setminus A} u$ belongs to $\mathcal{U}$.

$\mathcal{U}$ of $(\mathcal{T}, \mathcal{B}(X))$-measurable functions

$$\inf_{u\in\mathcal{U}} I_f(u) = \int_T \left(\inf_{x\in X} f(t,x)\right)\mu(dt),$$

provided that the left-hand side does not equal $+\infty$.

Above the integrals $I_f(u) := \int_T f(t,u(t))\mu(dt)$ must be interpreted with the following *convention*: $I_f(u) := \int_T \max(f(t,u(t)),0)\mu(dt) - \int_T \max(-f(t,u(t)),0)\mu(dt)$, where the "tie" $(+\infty)-(+\infty)$ has to be read as $+\infty$. (This coincides precisely with convention VII.7 in [12].) A similar convention also applies to the integral on the right, whose integrand is already $\mathcal{T}$-measurable by the measurable projection theorem. (Apply [12, III.39].)

Proof of Theorem 1.1. Let T be the set $[0,R]\backslash N$, where N is the union of the exceptional null sets in (G1′) and (G3). We equip T with its Lebesgue σ-algebra $\mathcal{T}$ and the Lebesgue measure μ.

(i) Define $\iota := \inf(\mathrm{CP}^{**}) \geq -\int_T b$. (Use (G3).) By (G2), (CP**) is feasible, so ι is a finite number. Let $(u_k)_k \subset \mathcal{L}_1(T)^d$ be a minimizing sequence for (CP**). Then by (G3) the sequence $(\int_T |u_k|d\mu)_k$ is bounded, so without loss of generality we can suppose that $\ell := \lim_k \int_T |u_k|\, d\mu$ exists in $\mathbf{R}$. Form $(f_k)_k \subset \mathcal{L}_1(T)^2$ by setting $f_k(t) := (g^{**}(t,u_k(t)), |u_k(t)|)$. Then the above yields $\int_T f_k\, d\mu \to (\iota,\ell)$. Also, observe that $\max(0,-f_k^i)$ is uniformly integrable for each i: for $i=1$ this follows by $\max(0,-f_k^1(t)) \leq \max(0,-b(t))$, because of (G3), and for $i=2$ it is trivial. By Theorem 2.1 there exists $f_* := (f_*^1, f_*^2) \in \mathcal{L}^1(T)^2$ such that (a) $\int_T f_*\, d\mu \leq (\iota,\ell)$ and (b) $f_*(t)$ is a limit point of $(g^{**}(t,u_k(t),u_k(t)))_k$ for μ-a.e. t in T. According to (b), for a.e. t there is a subsequence (k_t) of (k) such that $(g^{**}(t,u_{k_t}(t)) \to f_*^1(t)$ and $|u_{k_t}(t)| \to f_*^2(t)$. The latter implies that a further subsequence of $(u_{k_t})_{k_t}$ converges in $\mathbf{R}^d$ to some limit point $u_{*,t}$. By lower semicontinuity of $g^{**}(t,\cdot)$, the preceding implies $f_*^1(t) \geq g^{**}(t,u_{*,t})$ for a.e. t. By continuity of $|\cdot|$, the same also implies $f_*^2(t) = |u_{*,t}|$. By the implicit measurable function result in [12, III.38][3] there exists a measurable function $\bar{u} : T \to \mathbf{R}^d$ such that $f_*^1(t) \geq g^{**}(t,\bar{u}(t))$ and $f_*^2(t) = |\bar{u}(t)|$ for a.e. t. By $\int_T f_*^2 d\mu \leq \ell < +\infty$ (see (a)) we obtain $\bar{u} \in \mathcal{L}_1(T)^d$, and by $\int_T f_*^1 d\mu \leq \iota$ the optimality of $\bar{u}$ for (CP**) follows. Next, recall from [12] that $\mathcal{L}_1([0,R])^d$ is decomposable. Again using (G2), Theorem 2.2 gives $I_{g^{**}}(\bar{u}) = \inf(CP^{**}) = \int_T \inf_{x\in\mathbf{R}^d} g^{**}(t,x)dt$, which is clearly equivalent to $g^{**}(t,\bar{u}(t)) = \inf_{x\in\mathbf{R}^d} g^{**}(t,x)$ a.e. This amounts to the stated Euler–Lagrange equation, because $\bar{v}(t) := -\int_t^R \bar{u}$ defines an optimal solution of (P**), for which $\bar{v}' = \bar{u}$ a.e.

(ii) The proof of existence of an optimal solution u^* of (CP) by means of Theorem 2.1 is precisely the same as the one given above. By Theorem 2.2 and (G5) (for $\xi = 0$) we have $+\infty > I_g(u^*) = \inf_{u\in\mathcal{L}_1([0,R])^d} I_g(u) = \int_T \phi(t)dt$ with $\phi(t) := \inf_{x\in\mathbf{R}} g(t,x)dt$. Because of this and (G2), ϕ belongs to $\mathcal{L}^1(T)$. So the identity $\int_T [g(t,u^*(t)) - \phi(t)]dt = 0$, where the integrand is clearly nonnegative, gives $g(t,u^*(t)) = \phi(t)$ a.e. By elementary properties of Fenchel conjugation, (G3) implies that $-\infty < \inf_{x\in\mathbf{R}^d} g(t,x) = \inf_{x\in\mathbf{R}^d} g^{**}(t,x)$ for all (or a.e.) t. So u^* satisfies $g^{**}(t,u^*(t)) \leq g(t,u^*(t)) = \inf_{x\in\mathbf{R}} g^{**}(t,x)$ a.e. Evidently, this implies that u^* is also optimal for (CP). The corresponding optimal solution of (P) and (P**) is given jointly by $v^*(t) := -\int_t^R u^*$. □

[3] On p. 85 of [12] one should substitute $\Sigma \equiv \mathbf{R}^d$ and $\Theta(t) := \{x \in \mathbf{R}^d : g^{**}(t,x) \leq f_*^1(t), |x| = f_*^2(t)\}$.

3. Epilogue. Let us present some observations and remarks about Theorem 1.1. Observe that our proof does not use the convexity of $g^{**}(t,\cdot)$, which is unlike [15]. This also explains why we could prove part (ii) independently from part (i). Moreover, in contrast to [15], Theorem 1.1 and its proof extend immediately to the situation where the place of $\mathbf{R}^d$ is taken by a separable reflexive Banach space X. (Replace $|\cdot|$ by the norm of X, but equip X with the weak topology.) Further, for the variational problem (P) in its equivalent form (CP), we can work with a finite measure space instead of $[0,R]$. Also, by using well-known Kuhn–Tucker-type optimality characterizations for (Q), which involve Theorem 2.2 in connection with scalar multipliers, we could have dealt with constrained versions of problem (P). This is because (P) is of so-called *Lyapunov-type*, as studied in [2, 18] and in [1, 4.3.1]. See [10] for some additional comments on how to handle the situation where the measure space may have atoms.

Finally, we wish to observe that the proof of Theorem 3.2 in [15] seems to have been inspired by the approach in the earlier paper [14], which, in turn, has been inspired by work by Olech [20]. However, [14, Theorem 3.1] deals with a more difficult existence problem, which is for the following variant of (P):

$$(\mathrm{P}') \qquad \min_{v\in AC([0,R])^d}\left\{\int_0^R g(t,v'(t))dt : v(0)=v_0, v(R)=v_1\right\}.$$

The reformulation of this as an optimal control problem is

$$(\mathrm{CP}') \qquad \min_{u\in\mathcal{L}_1([0,R])^d}\left\{\int_0^R g(t,u(t))dt : \int_0^R u(t)dt = v_1-v_0\right\}.$$

This problem cannot be treated by the general apparatus for Lyapunov-type optimization problems, because the uniform integrability condition for negative parts in Theorem 2.1 fails to hold under the conditions of [14]. In terms of the proof of Theorem 1.1(i), this can be seen by observing that the adapted proof would have to use $f_k(t) := (g^{**}(t,u_k(t)), |u_k(t)|, u_k(t))$ or something similar. This fails because $(\max(0,-u_k^i))_k$ is not uniformly integrable, given Crasta's slow growth conditions for g^{**}. In other words, we conclude that the method of proof followed in [14] is quite appropriate for the problem (P') but that the related proof in [15] for the much simpler problem (P) falls short.

After the submission of this note it was pointed out to the present author that in [16, Theorem 3.10] Crasta has obtained an existence result for a more general Bolza problem that, when specialized to the setting of the present note, gives precisely the existence result of Theorem 1.1. Also, in [17] much more general results have been given for the problem (P^{**}) by exploiting the associated Euler–Lagrange equations in connection with fixed point results.

Acknowledgments. The author is indebted to two anonymous referees for helpful references and suggestions about the presentation of this note.

REFERENCES

[1] V. M. Alekseev, V. M. Tichimirov, and S. V. Fomin, *Optimal Control*, Consultants Bureau, New York, 1987.

[2] V. I. Arkin and V. L. Levin, *Convexity of values of vector integrals, theorems on measurable choice and variational problems*, Uspekhi Mat. Nauk, 27 (1972), pp. 21–77.

[3] Z. Artstein, *On a variational problem*, J. Math. Anal. Appl., 45 (1974), pp. 404–415.

[4] Z. Artstein, *A note on Fatou's lemma in several dimensions*, J. Math. Econom., 6 (1979), pp. 277–282.
[5] R. J. Aumann and M. Perles, *A variational problem arising in economics*, J. Math. Anal. Appl., 11 (1965), pp. 488–503.
[6] E. J. Balder, *A unifying note on Fatou's lemma in several dimensions*, Math. Oper. Res., 9 (1984), pp. 267–275.
[7] E. J. Balder, *Existence results without convexity conditions for general problems of optimal control with singular components*, J. Math. Anal. Appl., 101 (1984), pp. 527–539.
[8] E. J. Balder, *On seminormality of integral functionals and their integrands*, SIAM J. Control Optim., 24 (1986), pp. 95–121.
[9] E. J. Balder, *New existence results for optimal controls in the absence of convexity: The importance of extremality*, SIAM J. Control Optim., 32 (1994), pp. 890–916.
[10] E. J. Balder and M. R. Pistorius, *On an optimal consumption problem for p-integrable consumption plans*, Econom. Theory, 17 (2001), pp. 721–737.
[11] H. Berliocchi and J. M. Lasry, *Intégrandes normales et mesures paramétrées en calcul des variations*, Bull. Soc. Math. France, 101 (1972), pp. 129–184.
[12] C. Castaing and M. Valadier, *Convex Analysis and Measurable Multifunctions*, Lecture Notes in Math. 580, Springer-Verlag, Berlin, 1977.
[13] L. Cesari, *An existence theorem without convexity conditions*, SIAM J. Control, 12 (1974), pp. 319–331.
[14] G. Crasta, *Existence of minimizers for nonconvex variational problems with slow growth*, J. Optim. Theory Appl., 99 (1998), pp. 381–401.
[15] G. Crasta, *On the minimum problem for a class of noncoercive nonconvex functionals*, SIAM J. Control Optim., 38 (1999), pp. 237–253.
[16] G. Crasta, *On a Class of Non-Convex Coercive Bolza Problems with Constraints on the Derivatives*, preprint, University of Modena, Modena, Italy, 1999.
[17] G. Crasta and A. Malusa, *Euler-Lagrange inclusions and existence of minimizers for a class of non-coercive variational problems*, J. Convex Anal., 7 (2000), pp. 167–182.
[18] E. Giner, *Minima sous contrainte, de fonctionelles intégrales*, C. R. Acad. Sci. Paris Sér. I Math., 321 (1995), pp. 429–431.
[19] A. D. Ioffe and V. M. Tichomirov, *Duality of convex functions and extremum problems*, Uspekhi Mat. Nauk, 23 (1968), pp. 51–116.
[20] C. Olech, *Existence theory in optimal control—the underlying ideas*, in Proceedings of the International Conference on Differential Equations, H. A. Antosiewicz, ed., Academic Press, New York, 1975, pp. 612–635.
[21] R. T. Rockafellar, *Integral functionals, normal integrands and measurable selections*, in Nonlinear Operators and the Calculus of Variations, Lecture Notes in Math. 543, J. P. Gossez et al., eds., Springer-Verlag, Berlin, 1976, pp. 157–207.
[22] D. Schmeidler, *Fatou's lemma in several dimensions*, Proc. Amer. Math. Soc., 24 (1970), pp. 300–306.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 2, pp. 333–347

CHARACTERIZATION OF EXTREMAL CONTROLS FOR INFINITE DIMENSIONAL TIME-VARYING SYSTEMS*

DIOMEDES BARCENAS† AND HUGO LEIVA†

Abstract. In this paper we prove some properties of attainable sets for time-varying infinite dimensional linear control systems with time-varying constrained controls and target sets. We also characterize the extremal controls and give necessary and sufficient conditions for the normality of the system. Moreover, we prove an existence theorem for the time optimal control, establishing two maximum principles.

Key words. attainable sets, extremal control, time optimal control, normal system

AMS subject classifications. 34C35, 34D05, 34G10

PII. S0363012997323631

1. Introduction. The time optimal control problem has been studied by many authors; while Lee–Markus [24] and Hermes–Lassalle [16] conform a good reference for the finite dimensional cases, for infinite dimensional linear systems it is worth mentioning the contributions of authors such as Ahmed [1], Ahmed–Teo [2], Curtain–Pritchard [8], Fattorini [12], [13], [14], Friedman [15], Hoppe [19], Raymond–Zidani [32], Li–Yong [25], and Papageorgiou [29] among others. All of these references, except [29], have the particularity of working with a constant target and hypotheses like reflexivity to obtain the corresponding time optimal control and maximum principle.

One of the purposes of this paper is to remove these hypotheses and the continuity of the adjoint evolution operator generated by the system as we indicate below.

Here we consider the general system

$$\begin{aligned} &\dot{x}(t) = A(t)x(t) + B(t)u(t), \quad t > 0, \\ &x(0) = x_0 \in X, \quad u(t) \in \mathcal{U}(t) \subset U, \quad x(t^*) \in G(t^*), \end{aligned} \tag{1.1}$$

for some $t^* > 0$ minimum, where the state $x(t) \in X$ and X and U are nonnecessarily separable Banach spaces. The family of unbounded operators $A(t)$ generates a strongly continuous evolution operator $S(t,s)$ according to [30] and [8], such that for each $t > 0$ the mapping $s \to S^*(t,s)$ is strongly continuous on $[0,t)$ and $B \in L^\infty_{loc}(0,+\infty; L(U,X))$, where $L(U,X)$ is the space of linear and bounded operators $T : U \to X$, and the controls values $\mathcal{U}(t)$ and the target set $G(t) \subset X$ are time-varying. For this class of systems, we get the results announced in the abstract.

As we have pointed out at the beginning of this introduction, the problem for infinite dimensional linear systems has been studied by several authors. For example, Friedman [15] established the existence of time optimal control for a class of time-invariant linear systems for which the target set is a singleton; and Ahmed–Teo [1] established the same where the target set is just $\{0\}$ and the system is time-varying. A generalization of Friedman's work can be found in Fattorini [13], where some hypotheses are relaxed.

*Received by the editors June 30, 1997; accepted for publication (in revised form) December 22, 2000; published electronically July 19, 2001. This research was partially supported by CDCHT-ULA under project C-840-97.

http://www.siam.org/journals/sicon/40-2/32363.html

†Department of Mathematics, Universidad de los Andes, Merida 5101, Venezuela (barcenas@ciens.ula.ve, hleiva@ciens.ula.ve).

More recently, Papageorgiou [29] extends part of these mentioned works for infinite dimensional time-varying systems, removing some restrictive hypotheses like invertibility of the evolution operator in [1], [12], [15], and target sets time-varying with a nonempty interior, which makes a different treatment to the problem since it does not include the case when the target set is a singleton.

The key hypothesis in Papageorgiou's work is the strong continuity of the adjoint evolution operator, which, incidentally, also generalizes some works of Peichl–Schappacher [31] and Barcenas–Leiva [5], in the sense that the space X does not need to be reflexive as in [5] and [31].

In this paper, inspired by Papageorgiou [29] and Barcenas–Diestel [3], using only the uniform integrability of the Bochner integral, we remove the continuity of the adjoint evolution operator at $s = t$, incorporating into the subject an important class of partial differential equations such as diffusion processes in nonreflexive Banach spaces; our target sets are time-varying and upper semicontinuous (USC) instead of continuous in the Hausdorff metric as in [29].

The paper is organized as follows. In section 3, a time optimal control theorem is obtained, where, as indicated, our target set is not continuous as in [29], and neither is the state space X reflexive as in [25] and [32]; this last reference deals with second order elliptic operators in the reflexive Banach space L^s $(s > 2)$, and, consequently, the adjoint of the associated strongly continuous semigroup is strongly continuous. As we can see in Example 6.1, our existence theorem (Theorem 3.1) applies to a one dimensional heat equation in nonreflexive Banach spaces, which is a case out of range of previous works since, as it is shown in this example, the adjoint evolution operator $S^*(t,s) = T^*_{t-s}$ is not strongly continuous at $t = s$; while this important example illustrates a time-invariant system, Example 6.2 exhibits a case of a time-varying system where Theorem 3.1 can be applied.

In section 4, we obtain a maximum principle (Theorem 4.1) which is different from that in Papageorgiou's paper; in fact, since the hypothesis in [29] requires $\operatorname{int} G(t) \neq \emptyset$, we observe that its maximum principle does not include the important case of a singleton $G(t) = \{x_0\}$. We also have that Theorem 4.1 applies only in the special case in which X is a reflexive Banach space and Theorem 4.2 is a generalization of Papageorgiou's maximum principle (Theorem 4.1 of [29]); to this end, we would like to remark that Theorem 4.1 is different from Theorem 4.2 since the latter requires the sets of attainable points to have nonempty interior. For this reason we prove Theorem 4.1 rather than Theorem 4.2.

We also would like to remark on the difference of our two maximum principles from those found in Li–Yong [25]. Several maximum principles are found in [25, Chapters V and VII] without any allusion to reflexivity in the respective hypotheses; however, we can see that the corresponding maximum principles stated on pages 170, 188, 203, and 212 of [25] ensure the associated solution in a suitable reflexive space (actually a Sobolev space), while the maximum principle given on page 292 (Chapter VII) of the same reference requires the existence of the optimal time t^* as a hypothesis, and such an existence is proved under the reflexivity assumption (see Theorem 5.9 of [25]). So the presence of reflexivity means that these results are comparable only with Theorem 4.1 of this paper; here we notice that in the case of time invariant systems the Li–Yong treatment is more general than this one since the only requirement is $\operatorname{Codim}(K(t) - G(t)) < \infty$. However, our conclusion allows us to characterize the extremal controls.

We also notice that the Raymond–Zidani maximum principle is stated in reflexive

Banach spaces.

In section 5 we extend the notion of normality from finite dimensional systems [24] to infinite dimensional systems, and we get the strict convexity of the attainable sets, a characterization of normality, and a bang-bang principle.

In the last section of this work we provide some examples where our results can be applied in several familiar situations.

Finally, we want to make the following remark. When we study a time-varying system

$$\dot{x}(t) = A(t)x(t), \quad x \in X, \quad t \geq 0, \tag{1.2}$$

we can try to transform the system (1.2) into a time-independent system by increasing the dimension of the phase space in the following way:

$$\begin{cases} \dot{x}(t) & = A(t)x, \quad t \in \mathbb{R}_+, \\ \dot{t} & = 1. \end{cases} \tag{1.3}$$

The system (1.3) generates a semigroup (semiflow) $\{\pi_t\}_{t\geq 0}$ on the space $X \times \mathbb{R}$ given by $\pi_t(x,s) = (T(t+s,s,)x, s+t)$, $t \geq 0$, where $T(t,s)$ is the evolution operator associated with (1.2). Although this approach is good for some time dependent dynamical systems, we do not know anything about the adjoint evolution operator. So the techniques illustrated here may not be applied.

2. Preliminaries. Even though in this paper we will work with the Lebesgue measure on the real line, we start our preliminaries in a general form. Let X and U be arbitrary Banach spaces, and let (Ω, Σ, μ) be a nonnegative, complete, finite measure space. We will use the following notations.

$$P_{f(c)}(U) =: \{A \subset U : \quad A \text{ closed (convex)}, \quad A \neq \emptyset\},$$

$$P_{wkc}(X) =: \{A \subset X : \quad A \neq \emptyset; \quad \text{weakly compact convex}\}.$$

The following definition comes from [4].

DEFINITION 2.1. *A multifunction $F : \Omega \to P_f(U)$ is called μ-measurable if there is a sequence $f_n : \Omega \to U$ of measurable functions and $N \in \Sigma$ with $\mu(N) = 0$ so that*

$$F(\omega) = \overline{\{f_n(\omega) : n \geq 1\}} \quad \forall \omega \in \Omega \setminus N.$$

As it is shown in [4], this definition coincides with the classical one when U is a separable Banach space.

DEFINITION 2.2. *A measurable selector of F is a μ-measurable function $f : \Omega \to U$ such that*

$$f(\omega) \in F(\omega) \quad \mu \text{ almost everywhere (a.e.).}$$

We denote

$$S_F^1 = \{f : f \in L_U^1(\mu); f \text{ is a measurable selector of } F\}.$$

DEFINITION 2.3. *A μ-measurable multifunction $F : \Omega \to U$ is called integrably bounded if there is $g \in L^1(\mu)$ such that*

$$\int_\Omega \|f(\omega)\| d\omega \leq \int_\Omega g d\mu$$

for every μ-measurable selector f of F.

DEFINITION 2.4. *On $P_f(U)$ the Hausdorff metric is defined by setting*

$$h(A,B) = \max\left\{\sup_{a\in A} d(a,B), \sup_{b\in B} d(A,b)\right\}.$$

The metric space $(P_f(U), h)$ is complete, and $P_{f(c)}(U)$ is a closed subspace of it.

DEFINITION 2.5. *If V is a Hausdorff topological space, the multifunction $F : V \to P_f(U)$ is continuous if it is a continuous function with the Hausdorff metric.*

F is called USC if for each nonempty open subset $\mathcal{A}$ of U, the set

$$F^{-1}(\mathcal{A}) = \{v \in V : F(v) \subset \mathcal{A}\}$$

is open in V.

3. Time optimal control problem. In this section we shall study the time optimal control problem associated with the linear system (1.1) which can be written in the following way using the foregoing notation:

$$\begin{aligned} &\dot{x}(t) = A(t)x(t) + B(t)u(t), \quad t > 0, \\ &x(0) = x_0, \quad u \in S^1_{\mathcal{U}}, \quad x(t^*) \in G(t^*), \end{aligned} \tag{3.1}$$

where $\mathcal{U} : [0,\infty) \to P_{wkc}(U)$ is an integrably bounded multifunction.

A mild solution of (3.1) is a function $x_u(\cdot) : [0,\infty) \to X$ defined by

$$x_u(t) = S(t,0)x_0 + \int_0^t S(t,\alpha)B(\alpha)u(\alpha)d\alpha, \quad t \geq 0, \tag{3.2}$$

where $u \in S^1_{\mathcal{U}}$.

DEFINITION 3.1. *For $t_1 > 0$ the set of admissible controls on $[0,t_1]$ is defined by*

$$C(t_1) = \{u \in L^1(0,t_1;U) : u(t) \in \mathcal{U}(t) \ \text{ a.e. in } \ [0,t_1]\} = S^1_{\mathcal{U}_{t_1}},$$

and the corresponding set of attainable points is defined by

$$K(t_1) = \{x_u(t_1) : x_u(\cdot) \ \text{ is a mild solution of } \ (3.1), \ u \in C(t_1)\},$$

where $\mathcal{U}_{t_1}$ is the restriction of $\mathcal{U}$ over $[0,t_1]$.

The following definition is a generalization of the similar one given in [24, p. 73].

DEFINITION 3.2. *A control $u \in C(t_1)$ is called an* extremal control *if the corresponding solution x_u of (3.1) satisfies $x_u(t_1) \in \partial K(t_1)$.*

DEFINITION 3.3. *For each $t \geq 0$, consider a target set $G(t) \subset X$. Suppose $t^* > 0$ and $u^* \in C(t^*)$ such that $x^*(t^*) \in G(t^*)$. Then u^* is called an optimal control if*

$$t^* = \inf\{t \in [0,\infty) : K(t) \cap G(t) \neq \emptyset\}.$$

The goal of this section is to provide the existence of optimal control for (3.1) under the hypothesis of upper semicontinuity on the target set. The importance of this result (Theorem 3.1) is that, even with constant target, it applies to linear partial differential equations in nonreflexive Banach spaces. In order to do that, we need to prove some propositions.

PROPOSITION 3.1. *$K(t_1)$ is convex and weakly compact in X.*

Proof. Define the multifunction

$$\Gamma : [0, t_1] \to P(X) \quad \text{by} \quad \Gamma(s) = S(t_1, s)B(s)\mathcal{U}(s).$$

The multifunction Γ is measurable, convex, a.e. weakly compact valued, and integrably bounded. By Theorem 3.2 of [4], S^1_Γ is weakly compact in $L^1(0, t_1; X)$.

Since $K(t_1)$ is merely a translation of $\int_0^{t_1} S^1_\Gamma$ and the integration is a bounded linear operator, the conclusion follows. □

PROPOSITION 3.2. *$K(t)$ is continuous in $t \in [0, T]$ with respect to the Hausdorff metric if $\bigcap_{t\in(0,T]} \mathcal{U}(t) \neq \emptyset$.*

Proof. Let $t_1 > 0$ be fixed and $\epsilon > 0$. We must find $\delta > 0$ such that

$$\text{if} \quad |t_1 - t_2| < \delta, \quad \text{then} \quad h(K(t_1), \quad K(t_2)) < \epsilon.$$

Let $t_2 \in (0, T)$ be with $0 < t_2 - t_1 < t_1$. If $x \in K(t_1)$, there exists $u \in C(t_1)$ such that

$$x = S(t_1, 0)x_0 + \int_0^{t_1} S(t_1, \alpha)\, B(\alpha)u(\alpha)d\alpha.$$

Define

$$\bar{u}(t) = \begin{cases} u(t) & \text{if } 0 \le t \le t_1, \\ v & \text{if } t > t_1, \end{cases}$$

where $v \in \bigcap_{t\in(0,T]} \mathcal{U}(t)$. Then

$$y = S(t_2, 0)x_0 + \int_0^{t_2} S(t_2, \alpha)B(\alpha)\bar{u}(\alpha)d\alpha \in K(t_2)$$

by the strong continuity of the evolution operator $S(t, s)$, the absolute continuity of the Bochner integral, and the Lebesgue dominated convergence theorem; given $\epsilon > 0$, there exists $\delta > 0$ such that, if $t_2 - t_1 < \delta$, then

$$\begin{aligned} \|x - y\| &\le \|S(t_2, 0)x_0 - S(t_1, 0)x_0\| + \left\| \int_{t_1}^{t_2} S(t_2, \alpha)B(\alpha)\bar{u}(\alpha)d\alpha \right\| \\ &\quad + \int_0^{t_1} \|S(t_2, \alpha)B(\alpha)u(\alpha) - S(t_1, \alpha)B(\alpha)u(\alpha)\|d\alpha \\ &< \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon. \end{aligned}$$

If $t_2 < t_1$, the conclusion is immediate. □

Even though the following corollary is an infinite dimensional version of the one found in Lee–Markus [24], our proof is very simple.

COROLLARY 3.1. *If $p \in \text{int}K(t_1)$ and $\bigcap_{t>0} \mathcal{U}(t) \neq \emptyset$, then there is a neighborhood N of p and $\delta > 0$ such that $N \subset K(t_2)$ for $|t_2 - t_1| < \delta$.*

Proof. Since $p \in \text{int}K(t_1)$, there is $r > 0$ such that $\overline{B(p, r)} \subset K(t_1)$ and

$$\alpha = \inf\{\|x - y\| : x \in \partial B(p, r), \quad y \in \partial K(t_1)\} > 0.$$

On the other hand, from Proposition 3.2, there exists $\delta > 0$ such that

$$|t - t_1| < \delta \Longrightarrow \text{h}(K(t_1), K(t)) < \frac{\alpha}{2}. \tag{3.3}$$

Now suppose that there exists $x_0 \in B(p,r) \setminus K(t)$ for some $t \in (t_1-\delta, t_1+\delta)$. Then for $y \in \partial K(t_1)$ and $x \in K(t)$ we have

$$\alpha < \mathrm{d}(x_0, y) \leq \mathrm{d}(x, y) + \mathrm{d}(x, x_0).$$

Hence

$$\begin{aligned}
\alpha &< \inf_{x\in K(t)} \mathrm{d}(x,y) + \inf_{x\in K(t)} \mathrm{d}(x_0,x) \\
&= \mathrm{d}(K(t),y) + \mathrm{d}(K(t),x_0) \\
&\leq \sup_{y\in K(t_1)} \mathrm{d}(K(t),y) + \frac{\alpha}{2} \\
&< \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha,
\end{aligned}$$

which is in contradiction to (3.3). Therefore, $B(p,r) \subset K(t)$ for $|t-t_1| < \delta$. □

THEOREM 3.1 (existence of the time optimal control). *Suppose the target $G(t)$ is convex, closed, and USC with respect to the Hausdorff metric. If there is a control $u \in C(t_1)$ such that $x_u(t_1) \in G(t_1)$, then there exists a time optimal control $u^* \in C(t^*)$.*

Proof. Put

$$H = \{t \in [0, t_1] : K(t) \cap G(t) \neq \emptyset\}$$

and $t^* = \inf H$. Then there is a decreasing sequence $\{t_n\} \subseteq H$ such that $\lim_{n\to\infty} t_n = t^*$, which implies the existence of a sequence $\{u_n\} \subset S^1_{\mathcal{U}_{t_1}}$ such that

$$S(t_n, 0)x_0 + \int_0^{t_n} S(t_n, s)B(s)u_n(s)ds \in G(t_n).$$

Since for each $n \in \mathbb{N}$

$$\int_0^{t_n} S(t_n,s)B(s)u_n(s)ds = \int_0^{t^*} S(t_n,s)B(s)u_n(s)ds + \int_{t^*}^{t_n} S(t_n,s)B(s)u_n(s)ds,$$

$\mathcal{U}$ is integrably bounded, and $t_n \to t^*$, then by the absolute continuity of the Bochner integral, given $\epsilon > 0$, there is $n_0 \in \mathbb{N}$ such that, if $n \geq n_0$, we have

$$\left\| \int_0^{t_n} S(t_n,s)B(s)u_n(s)ds - \int_0^{t^*} S(t_n,s)B(s)u_n(s)ds \right\| < \epsilon.$$

Since $S^1_{\mathcal{U}_{t^*}}$ is weakly compact in $L^1(0,t^*;U)$, then we can choose $\{u_n\}_{n\geq 1}$ converging weakly in $L^1(0,t^*;U)$ to some $u \in S^1_{\mathcal{U}_{t^*}}$.

On the other hand, if we define the multifunction

$$F : [0, t^*] \to P_{wkc}(X)$$

by the formula

$$F(s) = S(t^*, s)B(s)\mathcal{U}(s),$$

then S^1_F is weakly compact in $L^1(0,t^*;X)$. Hence there is $f \in S^1_F$ such that $S(t^*,\cdot)B(\cdot)u_n(\cdot)$ converges in the weak topology of $L^1(0,t^*;X)$ to f.

Suppose $f \neq S(t^*,\cdot)B(\cdot)u(\cdot)$ a.e.. Then by Diestel–Uhl [10, Corollary II.2.5], there is $E \in \Sigma$ such that

$$\int_E f d\mu \neq \int_E S(t^*,\cdot)B(\cdot)u(\cdot)ds.$$

Now, given $\eta = \int_E \|f - S(t^*,\cdot)B(\cdot)u(\cdot)\|ds$, there is $\delta > 0$ such that

$$t^* - t^{**} < \delta \Rightarrow \int_{t^{**}}^{t^*} \|S(t^*,s)B(s)u(s)\|ds < \frac{\eta}{4} \text{ uniformly in } u \in S^1_{\mathcal{U}}$$

and $\int_{t^{**}}^{t^*} \|f(s)\|ds < \frac{\eta}{4}$. So, if $x^* \in X^*$ and $\|x^*\| = 1$, then

$$\begin{aligned}
&\left|\left\langle x^*, \int_0^{t^*} (S(t^*,s)B(s)u(s) - f(s))ds \right\rangle\right| \\
&\quad \leq \left|\left\langle x^*, \int_0^{t^{**}} (S(t^*,s)B(s)u(s) - f(s))ds \right\rangle\right| + \frac{\eta}{2} \\
&= \lim_{n\to\infty} \left| \int_0^{t^{**}} \langle x^*, S(t^*,s)B(s)u(s) - S(t^*,s)B(s)u_n(s)\rangle \, ds \right| + \frac{\eta}{2} \\
&= \lim_{n\to\infty} \left| \int_0^{t^{**}} \langle B^*(s)S^*(t^*,s)x^*, u(s) - u_n(s)\rangle \, ds \right| + \frac{\eta}{2}.
\end{aligned}$$

Since $S^*(t^*,\cdot)x^*$ is continuous in $[0,t^{**}]$, we conclude that

$$B^*(\cdot)S^*(t^*,\cdot)x^* \in L^\infty(0,t^{**};U^*) \subset \left(L^1(0,t^{**};U)\right)^*,$$

and since u_n weakly converges to u in $L^1(0,t^{**};U)$, the above limit is equal to zero. So, we get that $S(t^*,\cdot)B(\cdot)u_n(\cdot)$ weakly converges to $S(t^*,\cdot)B(\cdot)u(\cdot)$ in $L^1(0,t^*;X)$.

Since

$$\begin{aligned}
\int_0^{t_n} S(t_n,s)B(s)u_n(s)ds &= \int_0^{t^*} S(t_n,s)B(s)u_n(s)ds \\
&\quad + \int_{t^*}^{t_n} S(t_n,s)B(s)u_n(s)ds
\end{aligned}$$

and

$$\int_{t^*}^{t_n} S(t_n,s)B(s)u_n(s)ds \to 0, \quad \text{as} \quad n \to \infty,$$

we shall concentrate our attention on the first right-hand side of the foregoing equality. Since $0 \leq s \leq t^* \leq t_n$, we get

$$\int_0^{t^*} S(t_n,s)B(s)u_n(s)ds = S(t_n,t^*) \int_0^{t^*} S(t^*,s)B(s)u_n(s)ds.$$

Therefore,

$$\lim_{n\to\infty} \left\langle x^*, \int_0^{t^*} S(t_n,s)B(s)u_n(s)ds \right\rangle = \left\langle x^*, \int_0^{t^*} S(t^*,s)B(s)u(s)ds \right\rangle.$$

Hence

$$\begin{aligned}\text{w-}\lim_{n\to\infty} x(t_n) &= \text{w-}\lim_{n\to\infty}\left[S(t_n,0)x_0 + \int_0^{t_n} S(t_n,s)B(s)u_n(s)ds\right]\\ &= S(t^*,0)x_0 + \int_0^{t^*} S(t^*,s)B(s)u(s)ds\\ &= x(t^*).\end{aligned}$$

Since $x(t_n)\in G(t_n)$ and G is USC, then Theorem 2.4 of [21] implies that $x(t^*)\in G(t^*)$ and the control $u\in S^1_{\mathcal{U}_{t_1}}$ is a required optimal control. □

The following corollary contains Lemma 2.1 of Fattorini [13] since this reference uses the hypothesis $L^\infty(X) = L^1(X^*)^*$, which implies reflexivity.

COROLLARY 3.2. *If $A = A(t)$ is the infinitesimal generator of a strongly continuous semigroup, if $L^\infty(0,T;X^*) = L^1(0,T;X)^*$ for $T>0$, if the target $G(t)$ is USC, and if $K(t)\cap G(t)\neq\emptyset$ for some $t\in(0,T]$, then the system* (3.1) *has an optimal solution.*

Proof. $L^\infty(0,T;X^*) = L^1(0,T;X)^*$ if and only if X^* has the Radon–Nikodym property [10]. In this case A generates a strongly continuous semigroup $\{T_t\}_{t\geq 0}$ whose adjoint semigroup is strongly continuous on $(0,+\infty)$ [3], [27]. If we put $S(t,s) = T_{t-s}$, then the conclusion follows by applying the foregoing theorem. □

4. The maximum principle. Two maximum principles are given in this section; the first one holds in reflexive Banach spaces since $\text{int}K(t)\neq\emptyset$. It is so because in this case, 0 has a relative weakly compact neighborhood, which implies that X is reflexive (see [11, p. 425]).

THEOREM 4.1. *Suppose* $\text{int}K(t)\neq\emptyset$ *for $t>0$. If $u^*\in C(t^*)$ is an optimal control and the target $G(t)$ is convex and continuous in the Hausdorff metric, then there is $x^*\neq 0$ such that*

$$m(s) = \max_{u(s)\in\mathcal{U}(s)}\langle x^*, S(t^*,s)Bu(s)\rangle = \langle x^*, S(t^*,s)Bu^*(t)\rangle$$

a.e. on $[0,t^*]$ whenever $\bigcap_{t\in(0,t^*+\epsilon]}\mathcal{U}(t)\neq\emptyset$.

Proof. $G(t^*)\cap\text{int}K(t^*) = \emptyset$. In fact, for the purpose of contradiction, let us suppose that there exists $p\in G(t^*)\cap\text{int}K(t^*)$. Then by Corollary 3.1 there is an open subset N containing p and $\delta>0$ such that

$$t^*-\delta < t < t^* \Rightarrow N\subset K(t)\Rightarrow G(t)\subset N^c.$$

On the other hand,

$$0<d = \text{dist}(p,N^c)\leq \inf_{x\in G(t)}\|p-x\|,\quad t\in(t^*-\delta,t).$$

Hence $\text{h}(G(t^*),G(t))\geq d$, which contradicts the continuity of $G(t)$ with respect to the Hausdorff metric. So the statement is proved.

Applying the Hahn Banach theorem, we find $x^*\in X^*,\ x^*\neq 0$, such that

$$\sup x^*(K(t^*))\leq\inf x^*(G(t^*)).$$

Then

$$\sup_{u\in S^1_{\mathcal{U}_{t^*}}}\left\langle x^*, S(t^*,0)x_0 + \int_0^{t^*} S(t^*,s)B(s)u(s)ds\right\rangle$$

$$\leq \left\langle x^*, S(t^*,0)x_0 + \int_0^{t^*} S(t^*,s)B(s)u^*(s)ds \right\rangle,$$

where u^* is an optimal control as in Theorem 3.1.

Therefore,

$$\sup_{u\in S^1_{\mathcal{U}_{t^*}}} \int_0^{t^*} x^*S(t^*,s)B(s)u(s)ds = \int_0^{t^*} x^*S(t^*,s)B(s)u^*(s)ds.$$

From Theorem 2.2 of Hiai–Umegaki [17] we get

$$\begin{aligned}\sup_{u\in S^1_{\mathcal{U}_{t^*}}} \int_0^{t^*} x^*S(t^*,s)B(s)u(s)ds &= \int_0^{t^*} \sup_{u(s)\in\mathcal{U}(s)} \langle x^*, S(t^*,s)B(s)u(s)\rangle\, ds \\ &= \int_0^{t^*} x^*S(t^*,s)B(s)u^*(s)ds.\end{aligned}$$

Hence

$$\sup_{u(s)\in\mathcal{U}(s)} \langle x^*, S(t^*,s)B(s)u(s)\rangle = \langle x^*, S(t^*,s)B(s)u^*(s)\rangle.$$

Since $S(t^*,s)B(s)\mathcal{U}(s)$ is weakly compact for each $s\in[0,t^*]$, then James's weak compactness theorem [20] implies that

$$\max_{u(s)\in\mathcal{U}(s)} \langle x^*, S(t^*,s)B(s)u(s)\rangle = \langle x^*, S(t^*,s)B(s)u^*(s)\rangle \quad \text{a.e.}$$

This completes the proof. □

In the proof of Theorem 4.1 we used some ideas from Theorem 4.1 of [29]. Here we shall state a similar result to that one without assuming the strong continuity at $s=t^*$ of the adjoint evolution operator $S^*(t^*,s)$. Since the proof of it is essentially the same, we will omit it.

THEOREM 4.2. *If $G(t)$ is convex and continuous with respect to the Hausdorff metric and* $\operatorname{int}G(t)\neq\emptyset\ \forall t\in[0,t^*]$, *and if u^* is an optimal control and X is separable, then there is $x^*\in X^*\setminus\{0\}$ so that*

$$m(s) = \max_{u(s)\in\mathcal{U}(s)} \langle x^*, S(t^*,s)B(s)u(s)\rangle = \langle x^*, S(t^*,s)B(s)u^*(s)\rangle \ \ a.e.$$

Remark 4.1. Theorems 4.1 and 4.2 admit comparison only when X is reflexive and separable. In fact, suppose the hypotheses of Theorem 4.1 hold. Since $K(t)$ is weakly compact and $\operatorname{int}K(t)\neq\emptyset$, then the Banach space X must be reflexive. So apparently, Theorem 4.2 is more general than Theorem 4.1. However, the hypothesis $\operatorname{int}G(t)\neq\emptyset$ is too restrictive and does not include the simple case $G(t)=\{x_1\}$, even in finite dimensional systems. We recall that the hypothesis of "$S^*(t^*,s)$" having been strongly continuous on $0\leq s<t^*$ allows us to incorporate into the subject the study of the heat equation in nonreflexive Banach spaces, a case which is not considered in [29].

THEOREM 4.3. *Suppose that the hypothesis of Theorem* 4.1 *holds. Then a control $u^*\in C(t^*)$ is extremal if and only if there is $x^*\in X^*\setminus\{0\}$ so that*

$$\max_{u(s)\in\mathcal{U}(s)} \langle x^*, S(t^*,s)B(s)u(s)\rangle = \langle x^*, S(t^*,s)B(s)u^*(s)\rangle \ \ a.e.$$

Proof. If u^* is an extremal control, then the corresponding solution $x(\cdot)$ of (3.1) satisfies $x(t^*) \in \partial K(t^*)$. Since $K(t^*)$ is convex and weakly compact, and since $\text{int}K(t^*) \neq \emptyset$, there exists $x^* \in X^* \setminus \{0\}$ such that

$$\sup_{x \in K(t^*)} \langle x^*, x \rangle = \langle x^*, x(t^*) \rangle .$$

So,

$$\sup_{u \in S^1_{\mathcal{U}_{t^*}}} \left\langle x^*, \int_0^{t^*} S(t^*, s)B(s)u(s)ds \right\rangle = \left\langle x^*, \int_0^{t^*} S(t^*, s)B(s)u^*(s)ds \right\rangle,$$

and the proof follows as in Theorem 4.1.

For the converse, if $u^* \in C(t^*)$, $x^* \in X^* \setminus \{0\}$,

$$\langle x^*, S(t^*, s)B(s)u^*(s) \rangle = \max_{u(s) \in \mathcal{U}(s)} \langle x^*, S(t^*, s)B(s)u(s) \rangle$$

a.e. on $[0, t^*]$ and $x(\cdot)$ is the solution of (1.1) corresponding to u^*, then for each $\hat{x}(t^*) \in K(t^*)$ there exists $\hat{u} \in C(t^*)$ such that

$$\hat{x}(t^*) = S(t^*, 0)x_0 + \int_0^{t^*} S(t^*, s)B(s)\hat{u}(s)ds.$$

Hence

$$\langle x^*, \hat{x}(t^*) - x(t^*) \rangle = \int_0^{t^*} \langle x^*, S(t^*, s)B(s)(\hat{u}(s) - u^*(s)) \rangle ds \leq 0,$$

which implies that $x(t^*) \in \partial K(t^*)$. □

Remark 4.2. The hypothesis $\text{int}K(t) \neq 0$ in the converse portion of the Theorem 4.3 is not necessary.

5. Normal systems. The following definition is a generalization of Lee–Markus [24, p. 79].

DEFINITION 5.1. *The control system* (1.1) *is called* normal *if the following implication holds. If* $u_1, u_2 \in C(t_1)$ *transfer* x_0 *to the same* $p \in \partial K(t_1)$*, then* $u_1 = u_2$ *a.e. on* $[0, t_1]$.

We recall that a convex set K in a Banach space X is strictly convex if each support hyperplane meets K in at most one point. This notion was introduced by Clarkson [7], and a nice study of it can be found in Diestel [9].

THEOREM 5.1. *Under the hypotheses of Theorems* 4.1 *and* 4.2*, if the control system* (1.1) *is normal, then* $K(t_1)$ *is strictly convex.*

Proof. Suppose that Π_{t_1} is a support hyperplane for $K(t_1)$ such that $p_a, p_b \in \Pi_{t_1} \cap K(t_1)$ with $p_a \neq p_b$ and $u_a, u_b \in C(t_1)$ as their corresponding controls.

We now consider the Banach space $Y = X \times X$ with the norm

$$\|y\|_Y = \| \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \| = \|x_1\|_X + \|x_2\|_X$$

and the function

$$f(t) = \begin{pmatrix} S(t_1, s)B(s)u_a(s) \\ S(t_1, s)B(s)u_b(s) \end{pmatrix}, \quad s \in [0, t_1],$$

with values in Y. Clearly $f \in L^1(0,t_1,Y)$. By Lyapunov's convexity theorem [33] the set

$$F = \left\{ w(D) = \int_D f(t)dt : D \subset [0,t_1] \ \text{ is measurable} \right\}$$

has convex closure. Thus

$$\frac{1}{2}w([0,t_1]) = \frac{1}{2}w([0,t_1]) + \frac{1}{2}w(\emptyset) \in \bar{F}.$$

Therefore, there exists a sequence $\{w(D_n)\}$ contained in F such that

$$\lim_{n\to\infty} w(D_n) = \frac{1}{2}w([0,t_1]), \quad \lim_{n\to\infty} w([0,t_1] \setminus D_n) = \frac{1}{2}w([0,t_1]).$$

Consider the controls

$$u_n^{(1)}(s) = \begin{cases} u_a(s), & s \in D_n, \\ u_b(t), & s \in [0,t_1] \setminus D_n, \end{cases}$$

$$u_n^{(2)}(s) = \begin{cases} u_a(s), & s \in [0,t_1] \setminus D_n, \\ u_b(t), & s \in D_n, \end{cases}$$

with corresponding solutions $x_n^{(1)}(\cdot)$ and $x_n^{(2)}(\cdot)$. It is easy to see that

$$\lim_{n\to\infty} x_n^{(1)}(t_1) = \lim_{n\to\infty} x_n^{(2)}(t_1) = \frac{1}{2}p_a + \frac{1}{2}p_b.$$

Since $C(t_1)$ is weakly compact in $L^1(0,t_1;U)$, we can suppose that the sequences $\{u_n^{(1)}\}$ and $\{u_n^{(2)}\}$ weakly converge to the controls $u_1, u_2 \in C(t_1)$, respectively.

Since $S^*(t_1,s)$ is strongly continuous in $s \in [0,t_1)$, then for each $x^* \in X^*$ and $\alpha \in (0,t_1)$ we have that

$$B^*(\cdot)S^*(t_1,\cdot)x^*u_n^{(i)}(\cdot) \ \text{ weakly converges to } \ B^*(\cdot)S^*(t_1,\cdot)x^*u_i(\cdot), \quad i = 1,2,$$

in $L^1(0,\alpha;\mathbb{R})$. If we take $\alpha_m \to t_1$, we see that

$$B^*(\cdot)S^*(t_1,\cdot)x^*u_n^{(i)}(\cdot)\mathcal{X}_{[0,\alpha_m]} \ \text{ converges to } \ B^*(\cdot)S^*(t_1,\cdot)x^*u^i(\cdot)\mathcal{X}_{[0,t_1]}, \quad i = 1,2,$$

on $[0,t_1]$. Since $\mathcal{U}$ is integrably bounded we get

$$\|B^*(\cdot)S^*(t_1,\cdot)x^*u^i\|_{L^1} \le \sup_{s\in[0,t_1)} \|S(t_1,s)\|\|x^*\|\|B\|_\infty \|g\|_{L^1},$$

where $g \in L^1[0,t_1]$ and $\|u(s)\| \le g(s)$, a.e. $\forall u \in S^1_{\mathcal{U}_{t_1}}$.

Thus, by applying the Lebesgue dominated convergence theorem, we get that

$$B^*(\cdot)S^*(t_1,\cdot)x^*u_n^{(i)}(\cdot) \ \text{ converges to } \ B^*(\cdot)S^*(t_1,\cdot)x^*u^{(i)}(\cdot), \quad i = 1,2,$$

in the weak topology of $L^1(0,t_1;U)$. Hence

$$\begin{aligned} \lim_{n\to\infty} \left\langle x^*, x_n^1(t_1) \right\rangle &= \langle x^*, S(t_1,0)x_0 \rangle + \left\langle x^*, \int_0^{t_1} S(t_1,s)B(s)u^1(s)ds \right\rangle \\ &= \left\langle x^*, \frac{1}{2}(p_a + p_b) \right\rangle = \left\langle x^*, S(t_1,0)x_0 + \int_0^{t_1} S(t_1,s)B(s)u^2(s)ds \right\rangle \\ &= \lim_{n\to\infty} \left\langle x^*, x_n^2(t_1) \right\rangle. \end{aligned}$$

Since this happens for each $x^* \in X^*$, we conclude that

$$S(t_1,0)x_0 + \int_0^{t_1} S(t_1,s)B(s)u^1(s)ds = S(t_1,0)x_0 + \int_0^{t_1} S(t_1,s)B(s)u^2(s)ds,$$

and by the normality of the system (3.1), we obtain

$$u_1(t) = u_2(t) \ \text{ a.e on } \ [0,t_1]. \tag{5.1}$$

Since $C(t_1) \subset L^1(0,t_1;U)$ is weakly compact, from the equality (5.1) we get

$$\lim_{n\to\infty} \left\langle u^*, u_n^{(1)} - u_n^{(2)} \right\rangle = 0 \ \text{ for each } \ u^* \in \left(L^1(0,t_1;U)\right)^*.$$

Therefore, by the definition of u_n^1 and u_n^2 we get $\langle u^*, u_a - u_b\rangle = 0$. Thus $u_a(t) = u_b(t)$ a.e. on $J = [0,t_1]$, and, consequently, $p_a = p_b$, which is a contradiction. This concludes the proof. □

THEOREM 5.2. *Suppose the hypothesis in Theorem* 4.1 *holds. Then the control system* (3.1) *is normal if and only if for each* $x^* \in X^* \setminus \{0\}$ *and a pair of controls* $u_1, u_2 \in C(t_1)$*, such that*

$$\begin{aligned} &\langle\ x^*, S(t_1,s)B(s)u_1(s)\rangle = \langle x^*, S(t_1,s)B(s)u_2(s)\rangle \\ &= \max_{u(s)\in\mathcal{U}(s)} \langle x^*, S(t_1,s)B(s)u(s)\rangle \ \ \textit{a.e on } \ [0,t_1] \end{aligned} \tag{5.2}$$

implies $u_1 = u_2$ *a.e. on* $[0,t_1]$.

Proof. Suppose the system (4.1) is normal; consider $x^* \in X^* \setminus \{0\}$, and let $u_1(\cdot), u_2(\cdot)$ be controls in $C(t_1)$ with the corresponding solutions $x_1(\cdot), x_2(\cdot)$, such that

$$\begin{aligned} &\langle\ x^*, S(t_1,s)B(s)u_1(s)\rangle = \langle x^*, S(t_1,s)B(s)u_2(s)\rangle \\ &= \max_{u(s)\in\mathcal{U}(s)} \langle x^*, S(t_1,s)B(s)u(s)\rangle \ \text{ a.e. on } \ [0,t_1]. \end{aligned}$$

Let Π be the hyperplane defined by

$$\Pi = \{x \in X : \langle x^*, x - x_1(t_1)\rangle = 0\}.$$

Π supports $K(t_1)$ at $x_1(t_1)$ and $x_2(t_1)$, and by the foregoing theorem $x_1(t_1) = x_2(t_1)$. Since (4.1) is normal, $u_1(t) = u_2(t)$ a.e. on $[0,t_1]$.

Conversely, let u_1, u_2 be controls belonging $C(t_1)$ which transfer x_0 to p. Then, by Theorem 4.3,

$$\begin{aligned} &\langle x^*, S(t_1,s)B(s)u_1(s)\rangle = \langle x^*, S(t_1,s)B(s)u_2(s)\rangle \\ &= \max_{u(s)\in\mathcal{U}(s)} \langle x^*, S(t_1,s)B(s)u(s)\rangle \ \text{ a.e. on } \ [0,t_1]. \end{aligned}$$

Therefore, by (5.2) we obtain $u_1(t) = u_2(t)$ a.e. on $[0,t_1]$. Hence the system (1.1) is normal. □

THEOREM 5.3. *If for almost all* s, $\mathcal{U}(s)$ *is strictly convex and* $\mathrm{Ker}B^*(s)S^*(t^*,s) = \{0\}$*, then the system* (1.1) *is normal.*

Proof. Suppose $x^* \in X^* \setminus \{0\}$. Then $x_s^* = B^*(s)S^*(t^*,s)x^* \neq 0$. Since $\mathcal{U}(s)$ is weakly compact and strictly convex, x_s^* attains its maximum at a unique point in $\mathcal{U}(s)$.

So, if $u_1,\ u_2 \in C(t^*)$ with

$$\begin{aligned}&\langle x^*, S(t^*,s)B(s)u_1(s)\rangle = \langle x^*, S(t^*,s)B(s)u_2(s)\rangle\\ &= \max_{u(s)\in\mathcal{U}(s)} \langle x^*, S(t^*,s)B(s)u(s)\rangle \ \text{ a.e. on } \ [0,t_1],\end{aligned}$$

then $u_1 = u_2$ a.e. and the system (1.1) is normal by Theorem 5.2. □

As a consequence we get the following corollary.

COROLLARY 5.1 (bang-bang principle). *Under the hypothesis of the former theorem, the optimal control u^* is unique and $u^*(t) \in \partial\mathcal{U}(t)$ a.e.*

6. Examples. In this section we shall show how some of our results can be applied in many evolution processes.

Example 6.1. Consider the one dimensional heat equation

$$\text{(6.1)}\qquad \begin{cases} x_t(t,\xi) = x_{\xi\xi}(t,\xi) + bu(t,\xi), \quad 0<\xi<1, \quad t>0,\\ x(t,0) = x(t,1), \quad x_\xi(t,0) = x_\xi(t,1), \quad t\geq 0,\\ x(0,\xi) = x_0(x), \end{cases}$$

as shown in Pazy [30, section 8.2]; if we associate with (6.1) the operator $A\phi = \phi_{\xi\xi}$, then A generates a strongly continuous semigroup $\{T_t\}_{t\geq 0}$ of compact operators on the nonreflexive Banach spaces $X = U = C_p[0,1]$ of all continuous and periodic functions with period 1 and the supremum norm.

Since T_t is compact, it is continuous in the uniform topology of $L(X)$ for $t>0$, and therefore the adjoint semigroup T_t^* is strongly continuous away from 0, and Theorem 3.1 can be applied with any target satisfying the hypothesis of that theorem.

We will show that T_t^* fails to be strongly continuous at $t=0$, and, consequently, this example does not fit in the theory developed in the cited references by simply putting $S(t,s) = T_{t-s}$.

In fact, since T_t is compact for $t>0$, then X is a sun-reflexive Banach space under the action of T_t (see [27], [28]). Therefore, X^* is a weakly compact generated Banach space [28]. Now a theorem from Kuo [23] shows that X^* has the Radon–Nikodym property, which is lacked by X^* in our present case.

We recall that a Banach space X has the Radon–Nikodym property when every X-valued, countably additive, and bounded variation vector measure has a Bochner integrable Radon–Nikodym derivative.

Example 6.2. If A is the generator of a strongly continuous semigroup T_t which is uniformly continuous for $t>0$ and $F \in L^\infty(0,T;L(X))$, then the evolution operator $S(t,s)$ generated by $A+F(s)$ is uniformly continuous in s for $0\leq s<t$ and is given by

$$S(t,s)x = T_{t-s}x + \int_s^t T_{t-r}F(r)S(t,r)dr.$$

The uniform continuity of $S(t,s)$ comes from the uniform integrability of

$$\{T_{t-(\cdot)}F(\cdot)S(t,\cdot)x : \quad \|x\|\leq 1\};$$

this implies that the adjoint evolution operator is strongly (actually uniformly) continuous in $0\leq s<t$.

Actually, if T_t is compact for $t>0$, and X^* is not weakly compactly generated, then as in Example 6.1 we can show that T_t^* is not strongly continuous at $t=0$. Hence, if $S(t,s) = T_{t-s}$, then $S^*(t,s)$ is not strongly continuous at $t=s$.

These kinds of examples appear in optimal control periodic problems and are studied in Li–Yong [25, pp. 160–164] in the particular case when T_t is compact for $t > 0$. However, neither time optimal control nor maximum principle is considered for this particular case.

Example 6.3. If we consider the Schrodinger operator $A\phi = i\Delta\phi$ in the Hilbert space $X = L^2(\mathbb{R}^n)$, then it is proved in Pazy [30, Chapter VII] that the operator A generates a strongly continuous group in X. If $B \in L^\infty(0,T;X)$ such that $B(t)$ is invertible a.e., then

$$\mathrm{Ker}B^*(s)S^*(t-s) = \{0\} \quad \mu. \text{ a.e.},$$

and defining the multifunction

$$\begin{aligned} \mathcal{U} : [0,T] &\to X, \\ s &\to sB_X, \end{aligned}$$

where $B_X = \{x \in X : \|x\| \le 1\}$, then $\mathcal{U}(t)$ is strictly convex for every $s \in [0,T]$ and so the system (1.1) is normal by Theorem 5.3.

Example 6.4. Consider the following evolution system:

$$x_t(t,\xi) = k\Delta x(t,\xi) + b(t,\xi)u(t,\xi), \quad t \ge 0, \quad \xi \in \Omega. \tag{6.2}$$

If $\Omega = \mathbb{R}^n$ (no boundary conditions), then we work in the space $X = L^1(\mathbb{R}^n)$. In this case the operator $A\phi = \Delta\phi$ also generates a strongly continuous semigroup $\{T_t\}_{t\ge 0}$ given by

$$T_t\phi(\xi) = \frac{1}{2^n(\pi kt)^{n/2}} \int_{\mathbb{R}^n} \exp\left(\frac{-|\xi-\eta|^2}{4kt}\right)\phi(\eta)d\eta, \quad t > 0.$$

As is shown in [13], the adjoint semigroup is strongly continuous away from zero. If this semigroup were continuous at $t = 0$, then it would be strongly continuous on $L^\infty(\mathbb{R}^n)$, and so by Lotz [26], T_t^* has a bounded generator which implies that T_t has a bounded generator also, which is a plain contradiction.

Acknowledgments. We want to thank the referees for their comments which helped us to improve the presentation of this paper.

REFERENCES

[1] N. V. Ahmed, *Finite time null controllability for a class of linear evolution equations on a Banach space with constraints*, J. Optim. Theory Appl., 44 (1985), pp. 129–158.
[2] N. V. Ahmed and K. L. Teo, *Optimal Control of Distributed Parameter Systems*, North-Holland, Amsterdam, The Netherlands, 1981.
[3] D. Barcenas and J. Diestel, *Constrained controllability in non-reflexive Banach spaces*, Quaest. Math., 18 (1995), pp. 185–198.
[4] D. Barcenas and W. Urbina, *Measurable multifunctions in nonseparable Banach spaces*, SIAM J. Math. Anal., 28 (1997), pp. 1212–1226.
[5] D. Barcenas and H. Leiva, *Controlabilidad con restricciones en espacios de Banach*, Acta Cient. Venezolana, 40 (1989), pp. 181–185.
[6] H. Brezis, *Analyse Fonctionnelle*, Masson Editeur, Paris, 1983.
[7] J. A. Clarkson, *Uniformly convex spaces*, Trans. Amer. Math. Soc., 40 (1936), pp. 396–414.
[8] R. F. Curtain and A. J. Pritchard, *Infinite Dimensional Linear Systems Theory*, Lecture Notes in Control and Inform. Sci. 8, Springer-Verlag, Berlin, New York, 1978.
[9] J. Diestel, *Geometry of Banach Spaces—Selected Topics*, Lecture Notes in Math. 485, Springer-Verlag, Berlin, New York, 1975.

[10] J. Diestel and J. J. Uhl, *Vector Measures*, Math. Surveys Monogr. 15, AMS, Providence, RI, 1977.
[11] N. Dunford and J. T. Schwartz, *Linear Operators. Part* I, Interscience, New York, 1958.
[12] H. O. Fattorini, *Time-optimal control of solutions of operational differential equations*, SIAM J. Control, 2 (1964), pp. 54–59.
[13] H. O. Fattorini, *The time optimal control problem in Banach spaces*, Appl. Math. Optim., 1 (1974), pp. 163–188.
[14] H. O. Fattorini, *Existence theory and the maximum principle for relaxed infinite-dimensional optimal control problems*, SIAM J. Control Optim., 32 (1994), pp. 311–331.
[15] A. Friedman, *Optimal control in Banach spaces*, J. Math. Anal. Appl., 19 (1967), pp. 35–55.
[16] H. Hermes and J. P. Lassalle, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.
[17] F. Hiai and H. Umegaki, *Integrals, conditional expectations, and martingales of multivalued functions*, J. Multivariate Anal., 7 (1977), pp. 149–182.
[18] E. Hille and R. S. Phillips, *Functional Analysis and Semi-groups*, Amer. Math. Soc. Colloq. Publ. 31, AMS, Providence, RI, 1957.
[19] R. H. W. Hoppe, *On the approximate solution of time-optimal control problems*, Appl. Math. Optim., 9 (1983), pp. 263–290.
[20] R. C. James, *Weakly compact sets*, Trans. Amer. Math. Soc., 113 (1964), pp. 129–140.
[21] M. Kisielewicz, *Differential Inclusions and Optimal Control*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1991.
[22] V. I. Korobov and N. K. Son, *Controllability of linear systems in Banach space in the presence of constraints on the control* I, Diff. Uravn., 16 (1980), pp. 806–817 (in Russian).
[23] T. Kuo, *On conjugate Banach spaces with the Radon–Nikodym property*, Pacific J. Math., 59 (1975), pp. 497–503.
[24] E. B. Lee and L. Markus, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
[25] X. Li and J. Yong, *Optimal Control Theory for Infinite Dimensional Systems*, Birkhäuser, Boston, 1995.
[26] H. P. Lotz, *Uniform convergence of operators on L^∞ and similar spaces*, Math. Z., 190 (1985), pp. 207–220.
[27] J. Van Neerven, *Reflexivity, the dual Radon–Nikodym property and continuity of the adjoint semigroups*, Indag. Math. (N.S.), 1 (1990), pp. 365–379.
[28] J. Van Neerven, *The Adjoint of a Semigroup of Linear Operators*, Lecture Notes in Math. 1529, Springer-Verlag, Berlin, 1992.
[29] N. S. Papageorgiou, *Time optimal control for infinite dimensional linear systems*, Indian J. Pure Appl. Math., 24 (1993), pp. 155–198.
[30] A. Pazy, *Semigroups of Linear Operators with Application to Partial Differential Equations*, Springer-Verlag, Berlin, New York, 1983.
[31] G. Peichl and W. Schappecher, *Constrained controllability in Banach spaces*, SIAM J. Control Optim., 24 (1986), pp. 1261–1275.
[32] J. P. Raymond and H. Zidani, *Pontryagin's principle for time-optimal problems*, J. Optim. Theory Appl., 101 (1999), pp. 375–402.
[33] J. J. Uhl, *The range of a vector measure*, Proc. Amer. Math. Soc., 23 (1969), pp. 158–163.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 2, pp. 348–369

LIMITING DISCOUNTED-COST CONTROL OF PARTIALLY OBSERVABLE STOCHASTIC SYSTEMS*

ONÉSIMO HERNÁNDEZ-LERMA† AND ROSARIO ROMERA‡

Abstract. This paper presents two main results on partially observable (PO) stochastic systems. In the first one, we consider a general PO system

$$x_{t+1} = F(x_t, a_t, \xi_t), \quad y_t = G(x_t, \eta_t) \quad (t = 0, 1, \dots) \qquad (*)$$

on Borel spaces, with possibly unbounded cost-per-stage functions, and we give conditions for the existence of α-discount optimal control policies ($0 < \alpha < 1$). In the second result we specialize $(*)$ to additive-noise systems

$$x_{t+1} = F_n(x_t, a_t) + \xi_t, \quad y_t = G_n(x_t) + \eta_t \quad (t = 0, 1, \dots)$$

in Euclidean spaces with $F_n(x, a)$ and $G_n(x)$ converging pointwise to functions $F_\infty(x, a)$ and $G_\infty(x)$, respectively, and we give conditions for the limiting PO model

$$x_{t+1} = F_\infty(x_t, a_t) + \xi_t, \quad y_t = G_\infty(x_t) + \eta_t$$

to have an α-discount optimal policy.

Key words. partially observable control systems, partially observable Markov control processes, hidden Markov models, discounted cost criterion

AMS subject classifications. 93E20, 90C40

PII. S0363012999365194

1. Introduction.

In this paper we consider a nonlinear, time-varying, partially observable (PO) stochastic control system with state process $\{x_t\}$ evolving according to the equation

$$x_{t+1} = F_t(x_t, a_t) + \xi_t\ , \quad t \in \mathbb{N}, \qquad (1.1)$$

where $\mathbb{N} := \{0, 1, \dots\}$, and observations $\{y_t\}$ are of the form

$$y_t = G_t(x_t) + \eta_t, \quad t \in \mathbb{N}. \qquad (1.2)$$

Assuming that the functions F_t and G_t converge pointwise to functions F_∞ and G_∞, that is, as $t \to \infty$

$$F_t(x, a) \to F_\infty(x, a) \quad \text{and} \quad G_t(x) \to G_\infty(x) \qquad (1.3)$$

for all (x, a) and x, respectively, we investigate the existence of optimal control policies for the *limiting PO system*

$$x_{t+1} = F_\infty(x_t, a_t) + \xi_t, \quad y_t = G_\infty(x_t) + \eta_t, \qquad (1.4)$$

*Received by the editors December 1, 1999; accepted for publication (in revised form) February 19, 2001; published electronically July 19, 2001. This research was partially supported by CONACYT (México) grant 32299-E for OHL and by DGES (Spain) grant PB96-0111 for RR.
http://www.siam.org/journals/sicon/40-2/36519.html

†Departamento de Matemáticas, CINVESTAV-IPN, A. Postal 14-740, México D.F. 07000, México (ohernand@math.cinvestav.mx). On sabbatical leave at Departamento de Sistemas, UAM-Azcapotzalco, Ave. San Pablo No. 180, México D.F. 02200, México.

‡Universidad Carlos III de Madrid, Departamento de Estadística y Econometría, C/Madrid 126-128, 28903 Getafe (Madrid), España (mrromera@est-econ.uc3m.es).

when the optimality criterion is the α-*discounted cost* ($0 < \alpha < 1$).

In fact, we present two main results. In the first one, we consider a *general PO system*

$$x_{t+1} = F(x_t, a_t, \xi_t), \quad y_t = G(x_t, \eta_t), \tag{1.5}$$

in which the *state space* X and the *observation set* Y are *Borel spaces* (that is, Borel subsets of complete and separable metric spaces). Similarly, the state and observation disturbances ξ_t and η_t take values in Borel spaces S and S', respectively, whereas the control actions a_t are taken from a compact metric space A. In this setting, we give conditions for the existence of α-discount optimal policies, allowing the cost-per-stage to be possibly *unbounded*. (See Theorem 2.5.)

In the second main result (Theorem 3.4), we consider the additive-noise case (1.1), (1.2) and the limiting system (1.4) on the spaces $\mathrm{X} = S = \mathbb{R}^{d_1}$ and $\mathrm{Y} = S' = \mathbb{R}^{d_2}$. Assuming (1.3), we give conditions ensuring the existence of an optimal control policy for (1.4).

To prove these results we begin by writing (1.5) as a PO Markov control (or decision) process (MCP), also known as a controlled "hidden Markov model" [6]. In other words, we work with a *general state transition law* and a *general observation kernel*, as in (2.10) and (2.11), respectively, which can be specialized in the obvious manner to (1.5), say. (See (2.12) and (2.13).) The formulation (2.10), (2.11) has, of course, technical advantages, but what is even more important is that it includes a class of models larger than (1.5). Namely, there are many applications in control of queues, fisheries, learning processes, and others (see [4, 6, 7, 13, 18, 20, 21]) described by "stochastic kernels" as in (2.10) and (2.11), on possibly finite or countable spaces, rather than by a "difference equation" model such as (1.5).

Our original motivation to study the limiting control problem, which naturally led us to consider the general system (1.5), was our interest in some biotechnological processes and other *time-varying* systems, as in (1.1), (1.2), but for which it is known that their "coefficients" tend to stabilize in the sense of (1.3); see [1, 16]. Alternatively, if the disturbances ξ_t and η_t have zero means, then (1.3) is equivalent to the convergence of the expected values

$$F_t(x, a) = E(x_{t+1} | x_t = x, a_t = a) \text{ and } G_t(x) = E(y_t | x_t = x). \tag{1.6}$$

Thus our Theorem 3.4 can also be interpreted as a result on the adaptive control of (1.4) when the terms $F_\infty(x, a)$ and $G_\infty(x)$ are *unknown* but they are being estimated by the conditional expectations in (1.6). Similarly, using (2.10) and (2.11), Theorem 3.4 is easily related to results on either the approximation or the adaptive control of PO systems with unknown state transition law and observation kernel [4, 6, 7, 9, 13, 14, 25]. This interpretation of Theorem 3.4 is valid, of course, for the *completely observable* (CO) case which results when $y_t = x_t$ for all time index t; see [5, 9, 15, 19, 22]. Similarly, in the *noncontrolled* case (namely, when the control space A is a one-point set, say), our results on (1.1)–(1.4) can be seen as stating the convergence of filtering models—see Lemma 5.1.

Our approach is somewhat related to the CO case considered in [15], but the technical requirements are quite different. This is due to the fact that the analysis of (1.5) requires us to introduce an equivalent CO system with values in a set of probability measures (see (2.5)–(2.7)). Thus, for instance, some "pointwise" statements in [15] in our present setting turn out to be statements on the convergence of measures in some suitable sense.

The remainder of the paper is organized as follows. In section 2 we state our assumptions and main result (Theorem 2.5) on the general PO system (1.5). Section 3 consists of two parts. In the first we consider the additive-noise system (3.1) and show that the assumptions in section 2 can be replaced with conditions on (3.1) itself. This is important to keep in mind because one of those assumptions (Assumption 2.4) is imposed on a "transformed" PO system, whereas the conditions in section 3 (Hypotheses A to D) are all on the original PO system (3.1), and so—at least in principle—they are easier to verify. In the second part of section 3 we state our result (Theorem 3.4) on the limiting PO system (1.4). Sections 4 and 5 contain the proofs of Theorem 2.5 and 3.4, respectively, and, finally, we conclude in section 6 with some general comments.

2. The general PO system. We begin with the following remark on the terminology and notation we shall use and then proceed to state the optimal control problem we are concerned with.

REMARK 2.1. (a) *Given a Borel space* X*, we denote by* $\mathcal{B}(\mathrm{X})$ *its Borel* σ*-algebra, and by* $\mathbb{P}(\mathrm{X})$ *the family of probability measures on* X*, endowed with the usual weak topology* $\sigma(\mathbb{P}(\mathrm{X}), C_b(\mathrm{X}))$*, where* $C_b(\mathrm{X})$ *stands for the Banach space of continuous bounded functions* u *on* X *with the sup norm* $\|u\| := \sup_x |u(x)|$*. Thus a sequence* $\{\mu_k\}$ *in* $\mathbb{P}(\mathrm{X})$ *is said to converge weakly to* μ *if*

$$\int_{\mathrm{X}} u d\mu_k \to \int_{\mathrm{X}} u d\mu \quad \forall u \in C_b(\mathrm{X}). \tag{2.1}$$

As X *is a Borel space, so is* $\mathbb{P}(\mathrm{X})$*. (See* [2, 3, 23]*, for instance.)*

(b) *Let* X *and* Y *be Borel spaces. A measurable function* $q : \mathrm{Y} \to \mathbb{P}(\mathrm{X})$ *is called a* stochastic kernel *on* X *given* Y*, and we denote by* $\mathbb{P}(\mathrm{X}|\mathrm{Y})$ *the family of all those stochastic kernels. Equivalently,* $q(dx|y)$ *is in* $\mathbb{P}(\mathrm{X}|\mathrm{Y})$ *if* $q(\cdot\,|y)$ *is a probability measure on* X *for each fixed* $y \in \mathrm{Y}$*, and* $q(B|\cdot\,)$ *is a measurable function on* Y *for each fixed* $B \in \mathcal{B}(\mathrm{X})$*. If* $\mathrm{X} = \mathrm{Y}$*, then a stochastic kernel is called a* Markov transition probability.

Throughout the paper we suppose the following.

ASSUMPTION 2.2. *All the stochastic processes considered below are defined on an underlying probability space* $(\Omega, \mathcal{F}, P)$*. In addition, the following hold.*

(a) *The state space* X*, the observation set* Y*, and the disturbance spaces* S *and* S' *are all Borel spaces.*

(b) *The control (or action) set* A *is a compact metric space.*

(c) *The state and observation disturbances* ξ_t *and* $\eta_t, t \in \mathbb{N}$*, form independent sequences of independent and identically distributed (i.i.d.) random variables with values in* S *and* S'*, respectively. These sequences are also independent of the initial state* x_0*. We denote by* $\mu \in \mathbb{P}(S)$ *and* $\nu \in \mathbb{P}(S')$ *the common distributions of* ξ_t *and* η_t*, respectively.*

(d) *The functions* $F(x, a, s)$ *and* $G(x, s')$ *in* (1.5) *are continuous.*

(e) *The cost-per-stage function* $c : \mathrm{X} \times A \to \mathbb{R}$ *is* (e_1) *nonnegative and lower semicontinuous (l.s.c.), and* (e_2) $c(x, a)$ *is continuous in* $x \in \mathrm{X}$ *uniformly on* A*.*

(f) *There exists a constant* C *and a continuous function* $w \geq 1$ *on* X *such that* $c(x, a) \leq Cw(x)$ *for all* $x \in \mathrm{X}$ *and* $a \in A$*.*

For examples of cost functions $c(x, a)$ that satisfy Assumptions 2.2(e) and (f), see, for instance, [11, 17, 20]. In particular, both assumptions trivially hold if A is a finite set (as in [7, 18, 21, 24]), whereas (e_2) holds if $c(x, a) = c_1(x) + c_2(a)$, where c_1 and c_2 are nonnegative functions with $c_1(x)$ continuous and $c_2(a)$ l.s.c.

The PO control problem. Let $\mathcal{Y}_t := \sigma(y_0, \dots, y_t)$ be the σ-algebra generated by the observations up to time t. By an *admissible control policy* (or simply a *policy*) we mean a sequence $\pi = \{a_t\}$ of A-valued random variables such that a_t is $\mathcal{Y}_t$-measurable for each $t \in \mathbb{N}$. We shall denote by Π the set of all such policies.

Let $\alpha \in (0,1)$ be a fixed "discount factor." For each policy $\pi \in \Pi$ and initial distribution $\varphi \in \mathbb{P}(\mathrm{X})$ (that is, φ is the a priori distribution of x_0), the corresponding α-discounted cost is defined as

$$V(\pi, \varphi) := \sum_{t=0}^{\infty} \alpha^t E_\varphi^\pi \left[c(x_t, a_t) \right], \tag{2.2}$$

where E_φ^π denotes the expectation operator with respect to the probability measure P_φ^π induced by π and φ. Let

$$V^*(\varphi) := \inf_\pi V(\pi, \varphi) \quad \text{for} \quad \varphi \in \mathbb{P}(\mathrm{X}) \tag{2.3}$$

be the *optimal* α-discounted cost. The *PO optimal control problem* is then to find an optimal policy π^*, that is, a policy such that

$$V(\pi^*, \varphi) = V^*(\varphi) \quad \forall \varphi \in \mathbb{P}(\mathrm{X}). \tag{2.4}$$

The CO control problem. To study the PO control problem we shall follow the standard procedure in which the PO problem is transformed into a CO problem using the *filtering process* $\{\varphi_t\}$ in $\mathbb{P}(\mathrm{X})$ defined as follows. For each policy $\pi \in \Pi$ and initial distribution $\varphi \in \mathbb{P}(\mathrm{X})$,

$$\varphi_0(B) := P_\varphi^\pi(x_0 \in B) = \varphi(B), \tag{2.5}$$

$$\varphi_t(B) := P_\varphi^\pi(x_t \in B | \mathcal{Y}_t) \quad \text{for} \quad t \geq 1, \tag{2.6}$$

which are defined for all B in $\mathcal{B}(\mathrm{X})$. The filtering process depends, of course, on the policy π and the initial distribution φ, and so, strictly speaking, we should write φ_t as $\varphi_{t,\varphi}^\pi$, for instance. However, we shall use the simpler notation in (2.5) and (2.6) unless we need to remark which π and φ are being used.

To continue with the description of the PO problem, we use the well-known fact (see, for instance, [2, 6, 25, 27, 28] and (3.5), (3.6) below) that there exists a measurable function $H : \mathbb{P}(\mathrm{X}) \times A \times \mathrm{Y} \to \mathbb{P}(\mathrm{X})$ such that (2.6) can be written as

$$\varphi_{t+1} = H(\varphi_t, a_t, y_{t+1}) \quad \forall t \in \mathbb{N} \tag{2.7}$$

with initial condition (2.5). (Note that, by Remark 2.1(b), H is a stochastic kernel on $\mathbb{P}(\mathrm{X})$ given $\mathbb{P}(\mathrm{X}) \times A \times \mathrm{Y}$.) Moreover, using the notation

$$\widehat{c}(\varphi, a) := \int_{\mathrm{X}} c(x, a) \varphi(dx) \quad \text{for} \quad \varphi \in \mathbb{P}(\mathrm{X}),\ a \in A, \tag{2.8}$$

we can rewrite the α-discounted cost in (2.2) as

$$V(\pi, \varphi) = \sum_{t=0}^{\infty} \alpha^t E_\varphi^\pi \left[\widehat{c}(\varphi_t, a_t) \right]. \tag{2.9}$$

Finally, the *CO problem* is to minimize (2.9) over all $\pi \in \Pi$, subject to (2.5) and (2.6), and this problem is equivalent to the original PO problem in the sense that an optimal policy for CO is optimal for PO.

Solution of the CO problem. To state our first main result we need some notation. Let $P \in \mathbb{P}(\mathrm{X}|\mathrm{X} \times A)$ and $Q \in \mathbb{P}(\mathrm{Y}|\mathrm{X})$ be the *state transition law* and the *observation kernel* corresponding to (1.5), that is,

$$P(B|x,a) := \mathrm{Prob}(x_{t+1} \in B | x_t = x, a_t = a) \tag{2.10}$$

and

$$Q(C|x) := \mathrm{Prob}(y_t \in C | x_t = x) \tag{2.11}$$

for each $B \in \mathcal{B}(\mathrm{X}), C \in \mathcal{B}(\mathrm{Y}), x \in \mathrm{X}, a \in A$, and $t \in \mathbb{N}$. More explicitly, in view of (1.5) and Assumption 2.2(c), we have that

$$P(B|x,a) = \int_S I_B\left[F(x,a,s)\right]\mu(ds), \tag{2.12}$$

where I_B denotes the indicator function of a set B, and, similarly,

$$Q(C|x) = \int_{S'} I_C\left[G(x,s')\right]\nu(ds'). \tag{2.13}$$

From (2.12) and (2.13), together with the bounded convergence theorem, it follows that P and Q are both *weakly continuous*; that is, if $x^n \to x$ and $a^n \to a$, then

$$\int_{\mathrm{X}} u(x')P(dx'|x^n,a^n) \to \int_{\mathrm{X}} u(x')P(dx'|x,a) \quad \forall u \in C_b(\mathrm{X}), \tag{2.14}$$

and

$$\int_{\mathrm{Y}} v(y)Q(dy|x^n) \to \int_{\mathrm{Y}} v(y)Q(dy|x) \quad \forall v \in C_b(\mathrm{Y}). \tag{2.15}$$

We also require the following conditions on the state transition law P.

ASSUMPTION 2.3. (a) *If $a^n \to a$, then $P(\cdot|x,a^n) \to P(\cdot|x,a)$ weakly, uniformly in $x \in \mathrm{X}$.*

(b) *If $(\varphi^n, a^n) \to (\varphi, a)$, then there exists an integer $N = N_{\varphi,a}$ and a finite measure $\zeta = \zeta_{\varphi,a}$ such that*

$$\int_{\mathrm{X}} P(\cdot\,|x,a^n)\varphi^n(dx) \le \zeta(\cdot) \quad \forall n \ge N.$$

The existence of a "majorant" measure ζ as in Assumption 2.3(b), which we will use in conjunction with Lemma 4.5, below, is discussed at the end of this section; see Remark 2.6 and Example 2.7. On the other hand, to verify Assumption 2.3(a) one may try to use one of the several metrics that metrize the weak convergence, such as the Dudley metric [29, Corollary 4.3.6]. For instance, suppose that $\mathrm{X} = \mathbb{R}^d$ and that $F(x,a,s)$ is of the form $F(x,a,s) = F_1(x) + F_2(a,s)$, where F_2 is a continuous bounded function on $A \times S$. Now choose an arbitrary Lipschitz bounded function $u : \mathrm{X} \to \mathbb{R}$ with Lipschitz constant $\overline{u}$, that is, $|u(x) - u(x')| \le \overline{u}|x - x'|$. Then, by (2.12), if $a^n \to a$, we obtain

$$\begin{aligned}
&\Big|\int_X u(x')P(dx'|x,a^n) - \int_X u(x')P(dx'|x,a)\Big| \\
&\quad \le \int_S |u[F(x,a^n,s)] - u[F(x,a,s)]|\mu(ds) \\
&\quad \le \overline{u}\int_S |F_2(a^n,s) - F_2(a,s)|\mu(ds) \\
&\quad \to 0 \quad \text{uniformly in} \quad x \in \mathrm{X}.
\end{aligned}$$

Hence, as the choice of u was arbitrary, $P(\cdot|x,a^n) \to P(\cdot|x,a)$ converges in the Dudley metric [29], uniformly in $x \in \mathrm{X}$, and so Assumption 2.3(a) follows. A similar conclusion holds, of course, if F is of the form $F(x,a,s) = F_1(x,s) + F_2(a)$, where F_2 is a continuous function on A. (For the additive-noise case, see (3.10).)

Now, for each $C \in \mathcal{B}(\mathrm{Y}), \varphi \in \mathbb{P}(\mathrm{X})$, and $a \in A$, consider the stochastic kernel

$$\widehat{q}(C|\varphi,a) := \mathrm{Prob}(y_{t+1} \in C|\varphi_t = \varphi, a_t = a), \tag{2.16}$$

which, using (2.10)–(2.13), can be written as

$$\begin{aligned}\widehat{q}(C|\varphi,a) &= \int_{\mathrm{X}}\int_{\mathrm{X}} Q(C|x')P(dx'|x,a)\varphi(dx) \\ &= \int_{\mathrm{X}}\int_{S}\int_{S'} I_C\left[G(F(x,a,s),s')\right]\nu(ds')\mu(ds)\varphi(dx).\end{aligned} \tag{2.17}$$

Finally, for each $D \in \mathcal{B}(\mathbb{P}(\mathrm{X})), \varphi \in \mathbb{P}(\mathrm{X}), a \in A,$ and $t \in \mathbb{N}$, let

$$\widehat{P}(D|\varphi,a) := \mathrm{Prob}(\varphi_{t+1} \in D|\varphi_t = \varphi, a_t = a)$$

be the *transition law of the filtering process* (2.7), which we can also write as

$$\widehat{P}(D|\varphi,a) = \int_{\mathrm{Y}} I_D\left[H(\varphi,a,y)\right]\widehat{q}(dy|\varphi,a). \tag{2.18}$$

ASSUMPTION 2.4. *Let H and $w \geq 1$ be as in* (2.7) *and Assumption* 2.2(f)*, respectively, and define $\widehat{w} : \mathbb{P}(\mathrm{X}) \to \mathbb{R}$ as $\widehat{w}(\varphi) := \int_{\mathrm{X}} w(x)\varphi(dx)$.*

(a) *H is continuous.*

(b) *There is a number $1 \leq \beta < 1/\alpha$ such that*

$$\int_{\mathbb{P}(\mathrm{X})} \widehat{w}(\varphi')\widehat{P}(d\varphi'|\varphi,a) \leq \beta\widehat{w}(\varphi) \quad \forall\varphi \in \mathbb{P}(\mathrm{X}), \quad a \in A. \tag{2.19}$$

Observe that the property "$w \geq 1$" of w is inherited by $\widehat{w}$ because

$$\widehat{w}(\varphi) := \int_{\mathrm{X}} w d\varphi \geq \varphi(\mathrm{X}) = 1 \quad \forall\varphi \in \mathbb{P}(\mathrm{X}).$$

We shall denote by $\mathbb{B}_w(\mathbb{P}(\mathrm{X}))$ the (vector) space of all real-valued measurable functions u on $\mathbb{P}(\mathrm{X})$ such that

$$\|u\|_w := \sup_{\varphi} |u(\varphi)|/\widehat{w}(\varphi) < \infty.$$

We can now state our first optimality result as follows.

THEOREM 2.5. *If Assumptions* 2.2, 2.3, *and* 2.4 *are satisfied, then the following hold.*

(a) *The optimal cost function $V^*(\varphi) := \inf_\pi V(\pi,\varphi)$, with $V(\pi,\varphi)$ as in* (2.9)*, is the unique solution in $\mathbb{B}_w(\mathbb{P}(\mathrm{X}))$ of the Bellman (or dynamic programming) equation*

$$V^*(\varphi) = \min_{a\in A}\left[\widehat{c}(\varphi,a) + \alpha\int_{\mathbb{P}(\mathrm{X})} V^*(\varphi')\widehat{P}(d\varphi'|\varphi,a)\right] \tag{2.20}$$

for all $\varphi \in \mathbb{P}(\mathrm{X})$.

(b) *Moreover, V^* is l.s.c.*

(c) *There exists a measurable function $f^* : \mathbb{P}(\mathrm{X}) \to A$ that attains the minimum in* (2.20)*, i.e., for all $\varphi \in \mathbb{P}(\mathrm{X})$*

$$V^*(\varphi) = \widehat{c}(\varphi, f^*(\varphi)) + \alpha \int_{\mathbb{P}(\mathrm{X})} V^*(\varphi')\widehat{P}(d\varphi'|\varphi, f^*(\varphi)), \tag{2.21}$$

and f^* determines an optimal control policy $\pi^* = \{a_t^*\}$ given by

$$a_t^* := f^*(\varphi_t) \quad \forall t \in \mathbb{N},$$

where $\{\varphi_t\}$ is the filtering process.

Theorem 2.5, which is proved in section 4, is essentially standard except for the fact that we are allowing a *general* PO system (1.5) and a possibly *unbounded* cost-per-stage $c(x, a)$, as in Assumption 2.2(e), (f). To the best of our knowledge, the only case studied in the literature in which $c(x, a)$ is unbounded is for the so-called *linear-quadratic-Gaussian* (*LQG*) PO system. Furthermore, the existence of the "filtering function" H in (2.7) depends only on the state transition law and the observation kernel in (2.10) and (2.11), not on the particular PO model (1.5). This means, in other words, that Theorem 2.5 is valid for general PO systems on Borel spaces, and so, in particular, it includes systems on *countable* spaces, which are very common in applications; see [4, 6, 7, 13, 18, 20, 21, 24, 25].

On the other hand, it goes without saying that in Theorem 2.5 the most restrictive hypothesis is Assumption 2.4 because it is stated in terms of the components H and $\widehat{P}$ of the CO problem—in contrast to Assumptions 2.2 and 2.3 that are given on the *original PO system*, and so, in principle, they are "easier" to verify in particular PO models. In the following section we show, among other things, that Assumptions 2.3 and 2.4 hold for additive-noise models under reasonably mild conditions.

We conclude this section with some comments on Assumption 2.3(b) which are used in the following sections.

REMARK 2.6. (a) *Let* X *be an arbitrary Borel space, and let* Γ *be an arbitrary subfamily of* $\mathbb{P}(\mathrm{X})$*. A measure* γ^m *on* X *is said to be a* majorant *of* Γ *if* $\gamma^m(\cdot) \geq \gamma(\cdot)$ *for all* $\gamma \in \Gamma$*. If* Γ *has a* finite *majorant, we then say that* Γ *is* order-bounded from above [30, 31].

(b) *A family* $\Gamma \subset \mathbb{P}(\mathrm{X})$ *always has an* upper envelope*, that is, a majorant* γ^u *such that* $\gamma^u(\cdot) \leq \gamma^m(\cdot)$ *for any majorant* γ^m *of* Γ*. The construction of* γ^u*, which is used in the proof of Lemma* 5.1*, is as follows. Let* ρ^u *be the set function defined as*

$$\rho^u(B) := \sup\{\gamma(B)|\gamma \in \Gamma\} \quad \forall\, B \in \mathcal{B}(\mathrm{X}). \tag{2.22}$$

Then, as in Theorem 2.2 *in* [8]*, for instance, the upper envelope of* Γ *is the measure* γ^u *on* $\mathcal{B}(\mathrm{X})$ *given by*

$$\gamma^u(B) := \sup\left\{\sum_{k=1}^{\infty} \rho^u(B_k)|\{B_k\} \subset \mathcal{B}(\mathrm{X}) \quad \textit{is a partition of } B\right\}. \tag{2.23}$$

(There are more "explicit" ways of constructing γ^u *if* Γ *is a countable family* [8*, Remark* 2.4]*.) Clearly,* Γ *is order-bounded from above if and only if* γ^u *is a finite measure, i.e.,*

$$\gamma^u(\mathrm{X}) < \infty. \tag{2.24}$$

(c) *Let* $\| \lambda \|_{TV} := |\lambda|(\mathrm{X})$ *be the total variation norm of a finite signed measure* λ *on* X*, where* $|\lambda| = \lambda^+ + \lambda^-$ *denotes the variation measure. Let* $\{\gamma_n\}$ *be a sequence in* $\mathbb{P}(\mathrm{X})$ *such that* $\| \gamma_n - \gamma \|_{TV} \to 0$*. Then (by the definition of order-convergence—see* [31, Definition 2, p. 366]*) there exists a nonincreasing sequence of finite measures* $\widehat{\gamma}_n$ *on* X *such that* $\widehat{\gamma}_n(\mathrm{X}) \downarrow 0$ *and*

$$|\gamma_n - \gamma| \le \widehat{\gamma}_n \quad \forall\, n = 1, 2, \dots . \tag{2.25}$$

Now choose an arbitrary integer N *(for instance, such that* $\widehat{\gamma}_N(\mathrm{X}) \le \varepsilon$ *for some* $\varepsilon > 0$*.) It follows from* (2.25) *that*

$$\gamma_n(\cdot) \le \gamma(\cdot) + \widehat{\gamma}_N(\cdot) \quad \forall\, n \ge N. \tag{2.26}$$

In other words, if $\lambda_n \to \lambda$ *in the total variation norm, then, for any* $N > 0$*, the sequence* $\{\gamma_n, n \ge N\}$ *has the finite majorant* $\gamma(\cdot) + \widehat{\gamma}_N(\cdot)$*.*

For additional criteria for order-boundedness of measures, see, e.g., [30]. On the other hand, an obvious sufficient condition for Assumption 2.3(b) is that the whole family $\{P(\cdot|x,a) : x \in \mathrm{X}, a \in A\}$ is order-bounded from above, i.e.,

$$P(\cdot|x,a) \le \zeta(\cdot) \quad \forall\, x \in \mathrm{X},\ a \in A, \tag{2.27}$$

for some finite measure $\zeta(\cdot)$. This is the case in the following well-known example.

Example 2.7 (see, e.g., [32, 33, 34]). *Consider the additive-noise system*

$$x_{t+1} = F(x_t, a_t) + \xi_t \tag{2.28}$$

with state space $\mathrm{X} = \mathbb{R}^{d_1}$*, say. Suppose that* $F : \mathrm{X} \times A \to \mathrm{X}$ *is continuous and bounded and that the i.i.d. disturbances* ξ_t *have a continuous and bounded density* g_ξ *with respect to the Lebesgue measure* λ_1 *on* X*. Then, by* (2.10) *and* (2.12)*,*

$$P(B|x,a) = \int_B g_\xi(s - F(x,a))\lambda_1(ds). \tag{2.29}$$

Then, as the closure of the set $\{F(x,a)|x \in \mathrm{X}, a \in A\}$ *is compact in* X *and the function* $w \to g_\xi(s-w)$ *is continuous for each* s*, there is a bounded,* λ_1*-integrable function* $\widehat{g}_\xi$ *on* X *such that* $g_\xi(s - F(x,a)) \le \widehat{g}_\xi(s)$ *for all* s, x, a*. Hence* (2.27) *holds with*

$$\zeta(B) := \int_B \widehat{g}_\xi(s)\lambda_1(ds).$$

Remark 2.8. (a) *An argument as in Example* 2.7 *shows that* (2.27) *holds, for instance, for the nonadditive-noise models in Examples* 8.6.2 *and* 8.6.4 *in* [11].

(b) *Consider an additive-noise* observation process

$$y_t = G(x_t) + \eta_t \tag{2.30}$$

on $Y := \mathbb{R}^{d_2}$ *with i.i.d. disturbances* η_t*, which have a density* g_η *with respect to the Lebesgue measure* λ_2 *on* $\mathbb{R}^{d_2}$*. Then, by* (2.11) *and* (2.13)*,*

$$Q(C|x) = \int_C g_\eta(s' - G(x))\lambda_2(ds'), \tag{2.31}$$

and, as in Example 2.7*, it follows that if* g_η *and* G *are continuous bounded functions (on their corresponding domains), then there exists a finite measure* γ *on* Y *such that*

$$Q(\cdot|x) \le \gamma(\cdot) \quad \forall\, x \in \mathrm{X}.$$

(c) *The additive-noise systems in* (b) *and Example* 2.7 *with continuous bounded "drifts"* $G(x)$ *and* $F(x,a)$*, respectively, are also* order-bounded from below [32, 33, 34]*; that is, there exist nontrivial substochastic measures* γ_1^l *on* Y *and* γ_2^l *and* X *such that*

$$Q(\cdot|x) \ge \gamma_1^l(\cdot) \quad \text{and} \quad P(\cdot|x,a) \ge \gamma_2^l(\cdot) \quad \forall\, x, a. \tag{2.32}$$

For additional conditions ensuring order-boundedness from below, see [9, 11, 30, 34, 35]*, for instance.*

3. Additive-noise models and the limiting PO system. The main objective in this section is to study the limiting system (1.4). With this in mind, we first study the general additive-noise system (3.1), below, which serves several purposes. It illustrates the concepts introduced in section 2; it gives conditions ensuring that Assumptions 2.3 and 2.4 are satisfied; and it is an introduction to our main result (Theorem 3.4) on (1.4).

Additive-noise models. Consider the PO additive-noise system

$$x_{t+1} = F(x_t, a_t) + \xi_t, \quad y_t = G(x_t) + \eta_t, \quad t \in \mathbb{N}, \tag{3.1}$$

with $\mathrm{X} = S = \mathbb{R}^{d_1}, \mathrm{Y} = S' = \mathbb{R}^{d_2}$, and A a compact metric; see Assumptions 2.2(a), (b). In addition, the disturbances $\{\xi_t\}$ and $\{\eta_t\}$ are as in Assumption 2.2(c), except that now we also suppose the following.

Hypothesis A. *The noise distributions* μ *and* ν *are absolutely continuous, say,*

$$\mu(ds) = g_\xi(s)\lambda_1(ds) \quad \textit{and} \quad \nu(ds') = g_\eta(s')\lambda_2(ds'), \tag{3.2}$$

where λ_i $(i = 1, 2)$ *denotes the Lebesgue measure on* $\mathbb{R}^{d_i}$*, and, moreover,* g_ξ *and* g_η *are* continuous bounded *density functions.*

Then, as in (2.29) and (2.31), the state transition law and the observation kernel are given by

$$P(B|x,a) = \int_B g_\xi(s - F(x,a))\lambda_1(ds) \tag{3.3}$$

and

$$Q(C|x) = \int_C g_\eta(s' - G(x))\lambda_2(ds'), \tag{3.4}$$

respectively. On the other hand, as is well known [4, 9, 14, 24, 25, 27], the filtering function H in (2.7) is of the form

$$H(\varphi, a, y)(B) = \sigma(\varphi, a, y)(B)/\sigma(\varphi, a, y)(\mathrm{X}) \quad \forall B \in \mathcal{B}(\mathrm{X}) \tag{3.5}$$

with

$$\begin{aligned} \sigma(\varphi, a, y)(B) &= \int_B g_\eta(y - G(x')) \int_{\mathrm{X}} P(dx'|x,a)\varphi(dx) \\ &= \int_{\mathrm{X}} \left[\int_B g_\eta(y - G(x'))P(dx'|x,a) \right] \varphi(dx) \\ &= \int_{\mathrm{X}} \left[\int_B g_\eta(y - G(x'))g_\xi(x' - F(x,a))\lambda_1(dx') \right] \varphi(dx), \end{aligned} \tag{3.6}$$

by (3.3).

On the other hand, Assumption 2.2(d) reduces to the following.

HYPOTHESIS B. *The functions* $F : \mathrm{X} \times A \to \mathrm{X}$ *and* $G : \mathrm{X} \to \mathrm{Y}$ *in* (3.1) *are continuous.*

Let us denote by $\|\cdot\|_{TV}$ the *total variation norm* for measures and suppose that $x^n \to x$ and $a^n \to a$. Then, by (3.3), Hypotheses A and B, and Scheffé's theorem (see, for instance, pp. 223–224 in [3]), we have that

$$\|P(\cdot\,|x^n, a^n) - P(\cdot\,|x, a)\|_{TV} \to 0, \tag{3.7}$$

which is of course a lot stronger than (2.14). Similarly, by (3.4),

$$\|Q(\cdot\,|x^n) - Q(\cdot\,|x)\|_{TV} \to 0. \tag{3.8}$$

On the other hand, in addition to the cases that we have already mentioned in section 2, to obtain Assumption 2.3 we may suppose, for example, the following.

HYPOTHESIS C. (a) *If* $a^n \to a$, *then* $\|P(\cdot\,|x, a^n) - P(\cdot\,|x, a)\|_{TV} \to 0$ *uniformly on X.*

(b) *If* $(\varphi^n, a^n) \to (\varphi, a)$, *then there exist an integer* $N = N_{\varphi,a}$ *and a* λ_1*-integrable function* $\widehat{g} = \widehat{g}_{\varphi,a} \geq 0$ *such that*

$$\int_{\mathrm{X}} g_\xi(s - F(x, a^n))\varphi^n(dx) \leq \widehat{g}(s) \quad \forall s \in \mathrm{X} \text{ and } n \geq N. \tag{3.9}$$

Hypothesis C(a) holds, for instance, if $F(x, a)$ is "separable" in x and a, say,

$$F(x, a) = F_1(x) + F_2(a), \tag{3.10}$$

where F_1 and F_2 are continuous functions. This follows from (3.3) and using the change of variable $y := s - F_1(x)$. In fact, (3.10) is similar to the "separable" case in the paragraph after Assumption 2.3, and it covers many control models. For example, (3.10) appears in the *cash-balance model* of Hordijk and Yushkevich [17, section 6], in which the state process, the "cash balance," follows the scalar linear equation

$$x_{t+1} = x_t + a_t + \xi_t. \tag{3.11}$$

In (3.11), the disturbances ξ_t are i.i.d. standard Gaussian variables, and the control action $a_t = a$ corresponds to a withdrawal of money of size $-a$ if $a < 0$, or to a supply a if $a > 0$. The control set is a given compact interval, say, $A = [-M, M]$. In this setting, (3.9) also holds if the probability measure φ^n satisfies that, for some constant k, $\int \exp(x^2)\varphi^n(dx) \leq k$ for all n sufficiently large (see [17, p. 445]). Observe that (3.11) does not satisfy the boundedness of $F(x, a)$ used in Example 2.7.

PROPOSITION 3.1. *Hypothesis C implies Assumption* 2.3.

Proof. It is evident that Hypothesis C(a) implies Assumption 2.3(a). Now, to obtain Assumption 2.3(b), note that (3.3) and (3.9) yield

$$\begin{aligned}\int_{\mathrm{X}} P(B|x, a^n)\varphi^n(dx) &= \int_B \int_{\mathrm{X}} g_\xi(s - F(x, a^n))\varphi^n(dx)\lambda_1(ds) \\ &\leq \int_B \widehat{g}(s)\lambda_1(ds)\end{aligned}$$

for all $n \geq N$. Thus the measure $\zeta(B) := \int I_B(s)\widehat{g}(s)\lambda_1(ds)$ satisfies that

$$\int_{\mathrm{X}} P(\cdot\,|x, a^n)\varphi^n(dx) \leq \zeta(\cdot) \quad \forall n \geq N, \tag{3.12}$$

and Assumption 2.3(b) follows. □

We next show that Hypothesis C also yields Assumption 2.4(a).

PROPOSITION 3.2. *If Hypothesis C holds, then so does Assumption* 2.4(a).

Proof. Let H be as in (3.5) and suppose that $(\varphi^n, a^n, y^n) \to (\varphi, a, y)$. We wish to prove that

$$H(\varphi^n, a^n, y^n)(\cdot) \to H(\varphi, a, y)(\cdot) \ \text{ weakly.} \tag{3.13}$$

To prove this, let

$$\mu_n(\cdot) := \int_{\mathrm{X}} P(\cdot\,|x, a^n)\varphi^n(dx) \ \text{ and } \ \mu(\cdot) := \int_{\mathrm{X}} P(\cdot\,|x, a)\varphi(dx). \tag{3.14}$$

We will first show that

$$\mu_n(\cdot) \to \mu(\cdot) \ \text{ setwise.} \tag{3.15}$$

By (3.12) and Lemma 4.5 below (in which $\widehat{\mathrm{X}} := \mathbb{P}(\mathrm{X})$), to get (3.15) it suffices to show that $\mu_n \to \mu$ *weakly*. Thus choose an arbitrary function u in $C_b(\mathrm{X})$ and use (3.14) to write

$$\Big|\int_{\mathrm{X}} u d\mu_n - \int_{\mathrm{X}} u d\mu\Big| \leq f(n) + g(n) \tag{3.16}$$

with

$$f(n) := \Big|\int_{\mathrm{X}}\int_{\mathrm{X}} u(x')[P(dx'|x, a^n) - P(dx'|x, a)]\varphi^n(dx)\Big|$$

and

$$g(n) := \Big|\int_{\mathrm{X}}\int_{\mathrm{X}} u(x')P(dx'|x, a)\varphi^n(dx) - \int_{\mathrm{X}}\int_{\mathrm{X}} u(x')P(dx'|x, a)\varphi(dx)\Big|.$$

By (3.3), the integral $\int u(x')P(dx'|x, a)$ is, in particular, continuous in $x \in \mathrm{X}$ for each $a \in A$. Therefore, $g(n) \to 0$. On the other hand, by Hypothesis C(a),

$$f(n) \leq \|u\| \sup_x \|P(\cdot\,|x, a^n) - P(\cdot\,|x, a)\|_{TV} \to 0 \ \text{ as } \ n \to \infty.$$

Hence from (3.16) we conclude that $\mu_n \to \mu$ weakly, which, as was already noted, together with (3.12) and Lemma 4.5, gives (3.15).

Now, going back to (3.13), let u be an arbitrary function in $C_b(\mathrm{X})$, and use (3.14) and (3.6) to write

$$\int_{\mathrm{X}} u(x)\sigma(\varphi^n, a^n, y^n)(dx) = \int_{\mathrm{X}} u(x)g_\eta(y^n - G(x))\mu_n(dx). \tag{3.17}$$

Therefore, as $g_\eta(y^n - G(x)) \to g_\eta(y - G(x))$ for all $x \in \mathrm{X}$, from (3.15) and Lemma 4.4(c) we obtain that

$$\sigma(\varphi^n, a^n, y^n) \to \sigma(\varphi, a, y) \ \text{ weakly.} \tag{3.18}$$

Moreover, taking $u(\cdot) \equiv 1$ in (3.17), we get $\sigma(\varphi^n, a^n, y^n)(\mathrm{X}) \to \sigma(\varphi, a, y)(\mathrm{X})$. This latter fact, combined with (3.18) and (3.5), gives (3.13). □

Finally, we will show that the following hypothesis implies (2.19).

HYPOTHESIS D. *There exist positive constants* $\bar{\alpha}$ *and* $\bar{\sigma}$ *such that*
(a) $g_\eta(y - G(x))/\sigma(\varphi, a, y)(\mathrm{X}) \le \bar{\sigma}$ for all φ, a, y, x,
(b) $\int_{\mathrm{X}} w(x')P(dx'|x,a) \le \bar{\alpha} w(x)$ for all x, a, *and*
(c) $1 \le \bar{\alpha}\bar{\sigma} < 1/\alpha$, *where* α *is the discount factor in* (2.9) (or (2.2)).

For examples satisfying condition (b), see [11 (Chapter 8), 17, 20] and their references. Concerning (a), see Remark 3.6.

PROPOSITION 3.3. *Hypothesis D implies Assumption* 2.4(b) *with* $\beta := \bar{\alpha}\bar{\sigma}$.

Proof. As $\widehat{w}(\varphi) := \int w(x)\varphi(dx)$, writing $\mathbb{P}(\mathrm{X})$ as $\widehat{\mathrm{X}}$, the left-hand side of (2.19) becomes

$$\int_{\widehat{\mathrm{X}}} \widehat{w}(\varphi')\widehat{P}(d\varphi'|\varphi, a) = \int_{\mathrm{X}} w(x) \int_{\widehat{\mathrm{X}}} \varphi'(dx)\widehat{P}(d\varphi'|\varphi, a),$$

and so, using (2.18),

$$\int_{\widehat{\mathrm{X}}} \widehat{w}(\varphi')\widehat{P}(d\varphi'|\varphi, a) = \int_{\mathrm{X}} w(x) \int_{\mathrm{Y}} H(\varphi, a, y)(dx)\widehat{q}(dy|\varphi, a). \tag{3.19}$$

On the other hand, by (3.5), (3.6), and Hypothesis D(a),

$$H(\varphi, a, y)(\cdot\,) \le \bar{\sigma} \int_{\mathrm{X}} P(\cdot\,|x, a)\varphi(dx).$$

This inequality and (3.19) yield

$$\begin{aligned}\int_{\widehat{\mathrm{X}}} \widehat{w}(\varphi')\widehat{P}(d\varphi'|\varphi, a) &\le \bar{\sigma} \int_{\mathrm{X}} \left[\int_{\mathrm{X}} w(x')P(dx'|x, a)\right] \varphi(dx) \\ &\le \bar{\alpha}\bar{\sigma} \int_{\mathrm{X}} w(x)\varphi(dx) \ \text{[by Hypothesis D(b)]} \\ &= \bar{\alpha}\bar{\sigma}\widehat{w}(\varphi).\end{aligned}$$

That is, (2.19) holds with $\beta := \bar{\alpha}\bar{\sigma}$. □

Summarizing, *Hypotheses A to D imply Assumptions* 2.3 *and* 2.4.

The limiting PO system. For each $n \in \mathbb{N}_\infty$, consider the PO control system

$$x_{t+1} = F_n(x_t, a_t) + \xi_t, \quad y_t = G_n(x_t) + \eta_t, \tag{3.20}$$

where $F_n(x, a)$ and $G_n(x)$ are functions that satisfy (1.3). For $n = \infty$, we have the limiting PO system (1.4). We will use a subindex "n" to indicate functions and probabilities corresponding to the model in (3.20). For instance, the α-discounted cost and the optimal cost function in (2.2) and (2.3) become

$$V_n(\pi, \varphi) := \sum_{t=0}^{\infty} \alpha^t E^\pi_{n,\varphi}\left[c(x_t, a_t)\right]$$

and

$$V_n^*(\varphi) := \inf_\pi V_n(\pi, \varphi),$$

respectively.

THEOREM 3.4. *Suppose that for each* finite $n \in \mathbb{N}$, (3.20) *satisfies Assumption* 2.2 *as well as Hypotheses A to D. Moreover, in addition to* (1.3) *we suppose that the*

limiting functions $F_\infty(x,a)$ *and* $G_\infty(x)$ *are continuous. Then Theorem* 2.5 *holds for* $n=\infty$.

Theorem 3.4 is proved in section 5. In the meantime, we may observe that Theorem 3.4 yields an optimal control policy for the limiting system, $n=\infty$, as follows.

For each $n=1,2,\dots$, let f_n^* be an optimal control policy for the nth control model, obtained as in (2.21); that is,

$$V_n^*(\varphi)=\widehat{c}(\varphi,f_n^*(\varphi))+\alpha\int_{\mathbb{P}(\mathrm{X})}V_n^*(\varphi')\widehat{P}_n(d\varphi'|\varphi,f_n^*(\varphi)) \tag{3.21}$$

for all $\varphi\in\mathbb{P}(\mathrm{X})$.

COROLLARY 3.5. *Under the hypotheses of Theorem* 3.4*, there exists a measurable function* $f_\infty^*:\mathbb{P}(\mathrm{X})\to A$ *such that the following hold.*

(a) *For each* $\varphi\in\mathbb{P}(\mathrm{X})$, $f_\infty^*(\varphi)$ *is an accumulation point of the sequence* $\{f_n^*(\varphi)\}$.

(b) f_∞^* *is an optimal control policy for the limiting control model with* $n=\infty$.

Proof. (a) This is a consequence of a result of Schäl [26] (reproduced in [10, Proposition D.7] and also in [11, p. 65]). Part (b) follows from (a) and from an argument as in the proof of Theorem 4.6.5 in [10], for instance. □

Moreover, if for each finite n there is a *unique* policy f_n^* as in (3.21), which is the case for some "convex" control problems, then in Corollary 3.5(a) we obtain that *the whole sequence* $\{f_n^*(\varphi)\}$ converges to $f_\infty^*(\varphi)$ for each φ in $\mathbb{P}(\mathrm{X})$.

We conclude this section with some comments on Hypothesis D(a).

REMARK 3.6. *As* g_η *is bounded, that is,* $g_\eta(y-G(x))\le\|\ g_\eta\ \|$ *for all* x,y*, Hypothesis* D(a) *holds if there is a constant* $\overline{\sigma}$ *such that*

$$\sigma(\varphi,a,y)(\mathrm{X})\ge\|\ g_\eta\ \|/\overline{\sigma}\quad\forall\,\varphi,a,y.$$

On the other hand, if $P(\cdot|x,a)$ *is order-bounded from below, say, as in* (2.32)*, then by* (3.6)

$$\sigma(\varphi,a,y)(\mathrm{X})\ge\int_{\mathrm{X}}g_\eta(y-G(x'))\gamma_2^l(dx')=:\widehat{g}_\eta(y)\quad\forall\,\varphi,a,y. \tag{3.22}$$

Therefore, another sufficient condition for Hypothesis D(a) *is the existence of a constant* $\overline{\sigma}$ *such that*

$$g_\eta(y-G(x))\le\overline{\sigma}\,\widehat{g}_\eta(y)\quad\forall\,x,y. \tag{3.23}$$

Observe that (3.22) *and* (3.23) *are both verifiable for the cases in Example* 2.7 *and Remark* 2.8(b).

4. Proof of Theorem 2.5. To prove Theorem 2.5 let us first write the CO problem (2.7)–(2.9) as an MCP. Thus (as in [2, 9, 10, 11], for instance) consider the control model

$$(\widehat{\mathrm{X}},\widehat{A},\widehat{P},\widehat{c}) \tag{4.1}$$

with *state space* $\widehat{\mathrm{X}}:=\mathbb{P}(\mathrm{X})$ and *control set* $\widehat{A}:=A$. The "state" *transition law* $\widehat{P}$ and the *running cost* $\widehat{c}$ are as in (2.18) and (2.8), respectively. Then Theorem 2.5 will follow from the results in section 8.5 (and section 8.3) of [11] if we show that the MCP in (4.1) satisfies Assumptions 8.5.1 and 8.5.2 in [11], which are reproduced below as Conditions

4.1 and 4.2, respectively. (Under our present Assumption 2.2(e), the running cost or cost-per-stage $c(x,a)$ is *nonnegative,* and, therefore, so is $\widehat{c}(\varphi,a)$. This means that here we do *not* require Assumption 8.5.3 in [11], whereas in Assumption 8.5.2 we need only $\widehat{w}(\varphi)$ to be l.s.c. rather than continuous, where $\widehat{w}(\varphi) := \int w(x)\varphi(dx)$ is the function in Assumption 2.4 above.)

CONDITION 4.1 (see Assumption 8.5.1 in [11]).

(a) $\widehat{\mathrm{X}} := \mathbb{P}(\mathrm{X})$ *is a Borel space, and* $\widehat{A} := A$ *is a compact metric space.*

(b) $\widehat{c}$ *is l.s.c. and nonnegative on* $\mathbb{K} := \widehat{\mathrm{X}} \times \widehat{A}$.

(c) $\widehat{P}$ *is weakly continuous on* $\mathbb{K}$*; that is, if* $(\varphi^n, a^n) \to (\varphi, a)$*, then*

$$\int_{\widehat{\mathrm{X}}} u(\varphi')\widehat{P}(d\varphi'|\varphi^n, a^n) \to \int_{\widehat{\mathrm{X}}} u(\varphi')\widehat{P}(d\varphi'|\varphi, a) \quad \forall u \in C_b(\widehat{\mathrm{X}}).$$

CONDITION 4.2 (see Assumptions 8.5.2 and 8.3.2 in [11]). *There exist nonnegative constants* C *and* β*, with* $1 \le \beta < 1/\alpha$*, and an l.s.c. function* $\widehat{w} \ge 1$ *on* $\widehat{\mathrm{X}}$ *such that for each "state"* $\varphi \in \widehat{\mathrm{X}}$

(a) $\sup_{a\in A} \widehat{c}(\varphi,a) \le C\widehat{w}(\varphi)$*, and*

(b) $\sup_{a\in A} \int_{\mathrm{X}} \widehat{w}(\varphi')\widehat{P}(d\varphi'|\varphi,a) \le \beta\widehat{w}(\varphi)$. *(See* (2.19)*.)*

Comparing these conditions with our Assumptions 2.2, 2.3, and 2.4, we see that to prove Theorem 2.5 it suffices to show that (in order of difficulty)

(i) $\widehat{w}(\varphi) \ge 1$ for all $\varphi \in \mathrm{X}$, and $\widehat{w}$ is l.s.c.,

(ii) $\widehat{c}(\varphi,a)$ satisfies Conditions 4.1(b) and 4.2(a), and

(iii) $\widehat{P}$ satisfies Condition 4.1(c).

Proof of (i). We already noted (after Assumption 2.4) that $\widehat{w} \ge 1$, whereas the lower semicontinuity of $\widehat{w}$ follows from the general well-known fact (see, for instance, statement (12.3.37) on p. 225 of [11]) that *if* φ^n *converges weakly to* φ *and* $v : \mathrm{X} \to \mathbb{R}$ *is l.s.c. and bounded below, then*

$$\liminf_{n\to\infty} \int_{\mathrm{X}} v(x)\varphi^n(dx) \ge \int_{\mathrm{X}} v(x)\varphi(dx). \tag{4.2}$$

Hence, as $w \ge 1$ is continuous (Assumption 2.2(f)), taking $v = w$ in (4.2), we get that $\widehat{w}$ is l.s.c. □

Proof of (ii). Condition 4.2(a) obviously follows from (2.8) and Assumption 2.2(f). Now, to prove that $\widehat{c}(\varphi,a)$ is l.s.c., suppose that $(\varphi^n, a^n) \to (\varphi, a)$. Let

$$\bar{c}_j(x) := \inf_{i\ge j} c(x, a^i) \quad \text{for each } \ j = 1, 2, \dots, \tag{4.3}$$

and observe that $\bar{c}_j(\cdot)$ is continuous (by Assumption 2.2(e_2)) and that

$$c(x, a^n) \ge \bar{c}_j(x) \quad \forall x \in \mathrm{X} \text{ and } \ n \ge j.$$

Therefore, for each $j = 1, 2, \dots$ and $n \ge j$ we have

$$\widehat{c}(\varphi^n, a^n) := \int_{\mathrm{X}} c(x, a^n)\varphi^n(dx) \ge \int_{\mathrm{X}} \bar{c}_j(x)\varphi^n(dx),$$

and taking the limit infimum as $n \to \infty$, from (4.2) we obtain

$$\liminf_{n\to\infty} \widehat{c}(\varphi^n, a^n) \ge \int_{\mathrm{X}} \bar{c}_j(x)\varphi(dx). \tag{4.4}$$

Finally, as $\bar{c_j}(\cdot)$ is nondecreasing and (by Assumption 2.2(e_1))

$$\lim_{j\to\infty} \bar{c_j}(x) = \liminf_{n\to\infty} c(x, a^n) \geq c(x, a),$$

letting $j \to \infty$ in (4.4), we get (by monotone convergence)

$$\liminf_{n\to\infty} \widehat{c}(\varphi^n, a^n) \geq \int_{\mathrm{X}} c(x, a)\varphi(dx) = \widehat{c}(\varphi, a). \tag{4.5}$$

This completes the proof of (ii). □

To prove (iii) we will first state some general preliminary facts in which X stands for an arbitrary Borel space, and $\widehat{\mathrm{X}} := \mathbb{P}(\mathrm{X})$.

LEMMA 4.3. *Let $\{u_n\}$ and $\{\mu_n\}$ be sequences in $C_b(\mathrm{X})$ and $\widehat{\mathrm{X}}$, respectively, such that*

(a) *$u_n \to u$ uniformly on* X*, and*

(b) *$\mu_n \to \mu$ weakly.*

Then

$$\lim_{n\to\infty} \int u_n d\mu_n = \int u d\mu.$$

Proof. By (a), the function u is in $C_b(X)$, and so, by (b),

$$\int u d\mu_n \to \int u d\mu. \tag{4.6}$$

Therefore, as $u_n \leq \| u_n - u \| + u$ we get

$$\limsup_{n\to\infty} \int u_n d\mu_n \leq \int u d\mu. \tag{4.7}$$

Similarly, (4.6) and the inequality $u_n \geq - \| u_n - u \| + u$ yield

$$\liminf_{n\to\infty} \int u_n d\mu_n \geq \int u d\mu.$$

The latter fact and (4.7) give the lemma. □

LEMMA 4.4. *Let u and u_n $(n \in \mathbb{N})$ be measurable functions on* X*, and let $\{\mu_n\}$ be a sequence in $\widehat{\mathrm{X}}$. If μ_n converges setwise to μ, i.e.,*

$$\mu_n(B) \to \mu(B) \quad \forall B \in \mathcal{B}(\mathrm{X}), \tag{4.8}$$

and u and $\{u_n\}$ are bounded below, then

(a) $\liminf \int u d\mu_n \geq \int u d\mu$, *and*

(b) $\liminf \int u_n d\mu_n \geq \int (\liminf u_n) d\mu$. *Therefore,*

(c) *if $\{u_n\}$ is a bounded sequence of measurable functions such that $u_n \to u$, then $\lim \int u_n d\mu_n = \int u d\mu$.*

Proof. (a) As u is bounded below, we have that $u + N \geq 0$ for some constant N. Thus, without loss of generality, we may assume that u is *nonnegative*. Moreover, by (4.8), part (a) holds for indicator functions $u = I_B$ of Borel subsets B of X and also of course for "simple" functions u (that is, finite linear combinations of indicator functions of Borel sets). Now choose an arbitrary measurable function $u \geq 0$, and let

$\{v_k\}$ be a sequence of simple functions such that $v_k(x) \uparrow u(x)$ for all $x \in \mathrm{X}$. Then, for each k,

$$\liminf_{n\to\infty} \int u d\mu_n \geq \liminf_{n\to\infty} \int v_k d\mu_n = \int v_k d\mu,$$

and letting $k \to \infty$, we obtain (a).

(b) This part follows from (a) and an argument as in (4.3)–(4.5). That is, define $\overline{u}_j := \inf_{n\geq j} u_n$ and note that

$$\int u_n d\mu_n \geq \int \overline{u}_j d\mu_n \quad \text{for each} \quad j \quad \text{and} \quad n \geq j.$$

Hence, by (a),

$$\liminf_{n\to\infty} \int u_n d\mu_n \geq \int \overline{u}_j d\mu \quad \text{for each} \quad j,$$

and letting $j \to \infty$, we get (b).

(c) Applying (b) to both u_n and $-u_n$, we get (c). □

The following result is a special case of Proposition 2.3(a) in [12].

LEMMA 4.5. *Let $\{\mu_n\}$ be a sequence in $\widehat{\mathrm{X}}$ and suppose that*

(a) *$\mu_n \to \mu$ weakly, and*

(b) *there is an integer N and a finite measure γ on X such that $\mu_n(\cdot) \leq \gamma(\cdot)$ for all $n \geq N$.*

Then $\mu_n \to \mu$ setwise.

For an example of a sequence $\{\mu_n\}$ that satisfies part (a) in Lemma 4.5 but not part (b), see Remark 3.3 in [8], for instance.

LEMMA 4.6. *Under Assumption* 2.3(a)*, the stochastic kernel $\widehat{q}$ in* (2.17) *is weakly continuous; that is, if $(\varphi^n, a^n) \to (\varphi, a)$, then*

$$\int_{\mathrm{Y}} v(y)\widehat{q}(dy|\varphi^n, a^n) \to \int_{\mathrm{Y}} v(y)\widehat{q}(dy|\varphi, a) \quad \forall v \in C_b(\mathrm{Y}). \tag{4.9}$$

If in addition Assumption 2.3(b) *holds, then*

$$\widehat{q}(\cdot\,|\varphi^n, a^n) \to \widehat{q}(\cdot\,|\varphi, a) \ \textit{ setwise.} \tag{4.10}$$

Proof. Choose an arbitrary function v in $C_b(\mathrm{Y})$, and let

$$v'(x') := \int_{\mathrm{Y}} v(y)Q(dy|x') \quad \forall x' \in \mathrm{X}.$$

By (2.15), $v'(\cdot)$ is in $C_b(\mathrm{X})$, and, therefore, by (2.14), the function

$$v''(x, a) := \int_{\mathrm{X}}\int_{\mathrm{Y}} v(y)Q(dy|x')P(dx'|x, a) \tag{4.11}$$

is in $C_b(\mathrm{X} \times A)$. Suppose now that $(\varphi^n, a^n) \to (\varphi, a)$ and use (4.11) and (2.17) to write

$$\int_{\mathrm{Y}} v(y)\widehat{q}(dy|\varphi^n, a^n) = \int_{\mathrm{X}} v''(x, a^n)\varphi^n(dx). \tag{4.12}$$

Thus, since Assumption 2.3(a) yields that

$$v''(\cdot\,, a^n) \to v''(\cdot\,, a) \ \text{ uniformly on } \ \mathrm{X},$$

from (4.12) and Lemma 4.3 we get (4.9).

To obtain (4.10) observe that (2.17) and Assumption 2.3(b) together give

$$\begin{aligned} \widehat{q}(\cdot\,|\varphi^n, a^n) &= \int_{\mathrm{X}} \int_{\mathrm{X}} Q(\cdot\,|x') P(dx'|x, a^n) \varphi^n(dx) \\ &\le \int_{\mathrm{X}} Q(\cdot\,|x') \beta(dx') =: \gamma(\cdot\,) \end{aligned} \tag{4.13}$$

for all $n \ge N$, and for some integer $N = N_{\varphi,a}$. Hence, as the measure γ in (4.13) is finite, from (4.13), (4.9), and Lemma 4.5 we obtain the setwise convergence in (4.10). □

From these lemmas we can now prove (iii), that is, Condition 4.1(c), as follows.

Proof of (iii). Suppose that $(\varphi^n, a^n) \to (\varphi, a)$ and choose an arbitrary function u in $C_b(\mathrm{X})$. Then, by (2.18),

$$\int_{\widehat{\mathrm{X}}} u(\varphi') \widehat{P}(d\varphi'|\varphi^n, a^n) = \int_{\mathrm{Y}} u\left[H(\varphi^n, a^n, y)\right] \widehat{q}(dy|\varphi^n, a^n), \tag{4.14}$$

and by Assumption 2.4(a)

$$u\left[H(\varphi^n, a^n, y)\right] \to u\left[H(\varphi, a, y)\right] \quad \forall y \in \mathrm{Y}.$$

This fact, combined with (4.14), (4.10), and Lemma 4.4(c), yields Condition 4.1(c). □

As was already mentioned (after Condition 4.2), from (i), (ii), and (iii) we obtain Theorem 2.5.

5. Proof of Theorem 3.4. For each *finite* $n \in \mathbb{N}$, the Bellman equation (2.20) becomes

$$V_n^*(\varphi) = \min_{a \in A} \left[\widehat{c}(\varphi, a) + \alpha \int_{\widehat{\mathrm{X}}} V_n^*(\varphi') \widehat{P}_n(d\varphi'|\varphi, a) \right] \tag{5.1}$$

with $\widehat{\mathrm{X}} := \mathbb{P}(\mathrm{X})$. Thus to prove Theorem 3.4 it suffices to show that V_∞^* satisfies (5.1), i.e.,

$$V_\infty^*(\varphi) = \min_{a \in A} \left[\widehat{c}(\varphi, a) + \alpha \int_{\widehat{\mathrm{X}}} V_\infty^*(\varphi') \widehat{P}_\infty(d\varphi'|\varphi, a) \right], \tag{5.2}$$

because then the desired conclusion follows from the "uniqueness of solutions" in Theorem 2.5(a).

Now, to prove (5.2), let

$$\underline{u}(\varphi) := \liminf_{n \to \infty} V_n^*(\varphi) \quad \text{and} \quad \overline{u}(\varphi) := \limsup_{n \to \infty} V_n^*(\varphi).$$

We wish to show that

$$\underline{u}(\varphi) = \overline{u}(\varphi) = V_\infty^*(\varphi) \quad \forall \varphi \in \widehat{\mathrm{X}}. \tag{5.3}$$

To prove this, let us first note the following.

LEMMA 5.1. *As $n \to \infty$, the following hold.*

(a) $\|\widehat{q}_n(\cdot\,|\varphi,a) - \widehat{q}_\infty(\cdot\,|\varphi,a)\|_{TV} \to 0$ *for each* (φ,a) *in* $\widehat{\mathrm{X}} \times A$, *where* $\|\cdot\|_{TV}$ *denotes the total variation norm.*

(b) $\|H_n(\varphi,a,y)(\cdot\,) - H_\infty(\varphi,a,y)(\cdot\,)\|_{TV} \to 0$ *for all* (φ,a,y) *in* $\widehat{\mathrm{X}} \times A \times \mathrm{Y}$, *where* H_n *is the filtering function in* (3.5), (3.6).

(c) $\widehat{P}_n(\cdot\,|\varphi,a) \to \widehat{P}_\infty(\cdot\,|\varphi,a)$ *weakly for each* (φ,a).

(d) *Furthermore, for each pair* (φ,a) *in* $\widehat{\mathrm{X}} \times A$, *there exists an integer* $N = N_{\varphi,a}$ *and a finite measure* $\gamma = \gamma_{\varphi,a}$ *on* $\mathcal{B}(\widehat{\mathrm{X}})$ *such that* $\widehat{P}_n(\cdot|\varphi,a) \le \gamma(\cdot)$ *for all* $n \ge N$.

(e) *The convergence in* (c) *holds setwise.*

Proof. (a) For each $n \in \mathbb{N}_\infty$, let $P_n(\cdot\,|x,a)$ and $Q_n(\cdot\,|x)$ be as in (3.3) and (3.4); that is,

$$P_n(B|x,a) = \int_B g_\xi(s - F_n(x,a))\lambda_1(ds)$$

and

$$Q_n(C|x) = \int_C g_\eta(s' - G_n(x))\lambda_2(ds').$$

As $g_\xi(s - F_n(x,a)) \to g_\xi(s - F_\infty(x,a))$ for all (x,a,s), it follows from Scheffé's theorem that

$$\|P_n(\cdot\,|x,a) - P_\infty(\cdot\,|x,a)\|_{TV} \to 0 \quad \forall (x,a) \in \mathrm{X} \times A. \tag{5.4}$$

Similarly, as $g_\eta(s' - G_n(x)) \to g_\eta(s' - G_\infty(x))$, we have

$$\|Q_n(\cdot\,|x) - Q_\infty(\cdot\,|x)\|_{TV} \to 0 \quad \forall x \in \mathrm{X}. \tag{5.5}$$

Therefore, by (2.17),

$$\begin{aligned}
\widehat{q}_n(\cdot\,|\varphi,a) &= \int_{\mathrm{X}}\int_{\mathrm{X}} Q_n(\cdot\,|x')P_n(dx'|x,a)\varphi(dx) \\
&= \int_{\mathrm{X}}\int_{\mathrm{X}} [Q_n(\cdot|x') - Q_\infty(\cdot|x')]P_n(dx'|x,a)\varphi(dx) \\
&\quad + \int_{\mathrm{X}}\int_{\mathrm{X}} Q_\infty(\cdot|x')[P_n(dx'|x,a) - P_\infty(dx'|x,a)]\varphi(dx) \\
&\quad + \widehat{q}_\infty(\cdot|\varphi,a),
\end{aligned}$$

and then a straightforward calculation using (5.4) and (5.5) yields (a).

(b) By (3.5) and (3.6), to prove (b) it suffices to show that, for all (φ,a,y),

$$\sigma_n(\varphi,a,y)(B) = \int_{\mathrm{X}} \left[\int_B g_\eta(y - G_n(x'))P_n(dx'|x,a)\right]\varphi(dx)$$

converges to $\sigma_\infty(\varphi,a,y)(B)$ in the total variation norm. To do this, observe that, for all $B \in \mathcal{B}(\mathrm{X})$,

$$\left|\int_B g_\eta(y - G_n(x'))P_n(dx'|x,a) - \int_B g_\eta(y - G_\infty(x'))P_\infty(dx'|x,a)\right|$$

$$\leq \|g_\eta\|\ \|P_n(\cdot\,|x,a) - P_\infty(\cdot\,|x,a)\|_{TV} + \int_{\mathrm{X}} |g_\eta(y - G_n(x')) - g_\eta(y - G_\infty(x'))| P_\infty(dx'|x,a)$$

$$\to 0 \quad \text{as} \quad n \to \infty,$$

and the latter convergence is, of course, uniform in $B \in \mathcal{B}(\mathrm{X})$. This clearly implies

$$\|\sigma_n(\varphi,a,y)(\cdot) - \sigma_\infty(\varphi,a,y)(\cdot)\|_{TV} \to 0 \quad \text{as} \quad n \to \infty,$$

and (b) follows.

(c) Choose an arbitrary function u in $C_b(\widehat{\mathrm{X}})$. Then, by (2.18),

$$\int_{\widehat{\mathrm{X}}} u(\varphi')\widehat{P}_n(d\varphi'|\varphi,a) = \int_{\mathrm{Y}} u\,[H_n(\varphi,a,y)]\,\widehat{q}_n(dy|\varphi,a). \tag{5.6}$$

Now observe that the integrand $u[H_n(\varphi,a,y)]$ is bounded by $\|u\|$ for all n. Therefore, (c) follows from (5.6) together with parts (a) and (b), and Lemma 4.4(c).

(d) Fix an arbitrary pair (φ,a) in $\widehat{\mathrm{X}} \times A$. By part (a) and the Remark 2.6(c), there is an integer $N = N_{\varphi,a}$ and a finite measure $q^* = q^*_{N,\varphi,a}$ on Y such that

$$\widehat{q}_n(\cdot|\varphi,a) \leq q^*(\cdot) \quad \forall\, n \geq N. \tag{5.7}$$

For notational ease, let

$$\widehat{P}_n(\cdot) := \widehat{P}_n(\cdot|\varphi,a), \ \ \widehat{q}_n(\cdot) := \widehat{q}_n(\cdot|\varphi,a), \ \ H_n(y) := H_n(\varphi,a,y).$$

Moreover, for each Borel set $D \subset \widehat{\mathrm{X}}$ let

$$\Delta_n(D) := \{y \in Y | H_n(y) \in D\},$$

so that, replacing u in (5.6) with the indicator function I_D, we get

$$\widehat{P}_n(D) = \int_Y I_D[H_n(y)]\widehat{q}_n(dy) = \widehat{q}_n[\Delta_n(D)]. \tag{5.8}$$

We will now use (2.22) and (2.23) to show that $\Gamma := \{\widehat{P}_n, n \geq n\}$ has a finite upper envelope, which will complete the proof of (d). Let ρ^u be the set function

$$\rho^u(D) := \sup_{n \geq N} \widehat{P}_n(D),$$

and (as in (2.23)) let γ^u be the upper envelope of Γ, i.e.,

$$\gamma^u(D) := \sup\left\{\sum_{k=1}^{\infty} \rho^u(D_k) | \{D_k\} \subset \mathcal{B}(\widehat{\mathrm{X}}) \quad \text{is a partition of } D\right\}$$

for each $D \in \mathcal{B}(\widehat{\mathrm{X}})$. Note that, by (5.8) and (5.7),

$$\rho^u(D) \leq \sup_{n \geq N} q^*[\Delta_n(D)] \leq q^*\left[\bigcup_{n \geq N} \Delta_n(D)\right] \leq q^*(Y) < \infty. \tag{5.9}$$

Moreover, writing $\bigcup_{n\geq N}\Delta_n(D)$ as a union of *disjoint* sets $\Delta'_n(D)$, we have

$$\rho^u(D)\leq q^*\left[\bigcup_{n=N}^{\infty}\Delta_n(D)\right]=\sum_{n=N}^{\infty}q^*[\Delta'_n(D)].$$

Therefore, for any partition $\{D_k\}\subset\mathcal{B}(\widehat{\mathrm{X}})$ of D,

$$\begin{aligned}\sum_{k=1}^{\infty}\rho^u(D_k)&\leq\sum_{k=1}^{\infty}\sum_{n=N}^{\infty}q^*[\Delta'_n(D_k)]\\&=\sum_{n=N}^{\infty}\sum_{k=1}^{\infty}q^*[\Delta'_n(D_k)]\\&=\sum_{n=N}^{\infty}q^*[\Delta'_n(D)]\\&\leq q^*(Y)<\infty\qquad\text{(by (5.9))}.\end{aligned}$$

Hence, as the partition $\{D_k\}$ was arbitrary, it follows that $\gamma^u(D)$ is bounded above by $q^*(Y)<\infty$ for any D in $\mathcal{B}(\widehat{\mathrm{X}})$; thus γ^u is a finite measure, which can be taken as the measure $\gamma=\gamma_{\varphi,a}$ in (d).

(e) This is a consequence of (c), (d), and Lemma 4.5. □

We now go back to the proof of (5.3). First note that using the interchange of infima we get from (5.1) that

$$\inf_{n\geq k}V_n^*(\varphi)=\min_{a\in A}\left[\widehat{c}(\varphi,a)+\alpha\cdot\inf_{n\geq k}\int_{\widehat{\mathrm{X}}}V_n^*(\varphi')\widehat{P}_n(d\varphi'|\varphi,a)\right].$$

Therefore, taking the lim inf in both sides of (5.1) and using Lemmas 5.1(e) and 4.4(b), we obtain

$$\underline{u}(\varphi)\geq\min_{a\in A}\left[\widehat{c}(\varphi,a)+\alpha\int_{\widehat{\mathrm{X}}}\underline{u}(\varphi')\widehat{P}_\infty(d\varphi'|x,a)\right].$$

Therefore, by a standard dynamic programming argument (see, for instance, Lemma 4.2.7 in [10])

$$\underline{u}(\varphi)\geq V_\infty^*(\varphi)\quad\forall\varphi\in\widehat{\mathrm{X}}.\tag{5.10}$$

To complete the proof of (5.3), we next show that

$$\overline{u}(\varphi)\leq V_\infty^*(\varphi)\quad\forall\varphi\in\widehat{\mathrm{X}},\tag{5.11}$$

which, together with (5.10), yields (5.3). To obtain (5.11) we see from (5.1) that

$$V_n^*(\varphi)\leq\widehat{c}(\varphi,a)+\alpha\int_{\widehat{\mathrm{X}}}V_n^*(\varphi')\widehat{P}_n(d\varphi'|\varphi,a)\tag{5.12}$$

for all (φ,a) in $\widehat{\mathrm{X}}\times A$. Furthermore, by Lemma 5.1(e), $\widehat{P}_n(\cdot\,|\varphi,a)$ converges setwise to $\widehat{P}_\infty(\cdot\,|\varphi,a)$, and in addition (by the inequality (8.3.33) in [11], p.52), the sequence $V_n^*(\varphi)$ is uniformly bounded by $C\widehat{w}(\varphi)/(1-\beta)$, where C and $\beta:=\bar{\alpha}\bar{\sigma}$ are the constants in Assumption 2.2(f) and Proposition 3.3, respectively. It follows that the extended

Fatou lemma, Lemma 8.3.7(b) in [11], is applicable to (5.12), so that taking the lim sup as $n \to \infty$, we get

$$\overline{u}(\varphi) \leq \widehat{c}(\varphi, a) + \alpha \int_{\widehat{\mathrm{X}}} \overline{u}(\varphi')\widehat{P}_\infty(d\varphi'|\varphi, a).$$

This implies that

$$\overline{u}(\varphi) \leq \min_{a \in A} \left[\widehat{c}(\varphi, a) + \alpha \int_{\widehat{\mathrm{X}}} \overline{u}(\varphi')\widehat{P}_\infty(d\varphi'|\varphi, a) \right],$$

which in turn, by Lemma 4.2.7 in [10], for instance, yields (5.11).

6. Concluding remarks. As was already mentioned, the results in Theorem 2.5 are essentially well known except for the fact that $c(x, a)$ is allowed to be unbounded and for the generality of the PO system (1.5). However, to our knowledge, the proof itself is new, even for the case of a bounded cost function $c(x, a)$, that is, when $w(\cdot) \equiv 1$ in Assumption 2.2(f). Similarly, parts (a) and (b) in Lemma 5.1, which concern the *total variation norm*, are new.

On the other hand, Theorem 2.5 includes the important case in which the state space X and the observation set Y are *countable*, as occurs in many applications [4, 6, 7, 18, 20, 21, 24, ...]. In such a case, the filtering function H turns out to be similar to (3.5) with

$$\sigma(\varphi, a, y)(x') = Q(y|x') \sum_x P(x'|x, a)\varphi(x)$$

(compare with (3.6)), and so Assumptions 2.2, 2.3, and 2.4 can be simplified in the obvious manner.

REFERENCES

[1] G. Bastin and D. Dochain, *On-Line Estimation and Adaptive Control of Bioreactors*, Elsevier, Amsterdam, 1990.

[2] D. P. Bertsekas and S. E. Shreve, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.

[3] P. Billingsley, *Convergence of Probability Measures*, Wiley, New York, 1968.

[4] T. E. Duncan, B. Pasik-Duncan, and L. Stettner, *Adaptive control of a partially observed discrete time Markov process*, Appl. Math. Optim, 37 (1998), pp. 269–293.

[5] P. K. Dutta, M. K. Majumdar, and R. K. Sundaram, *Parametric continuity in dynamic programming models*, J. Econom. Dynam. Control, 18 (1994), pp. 1069–1092.

[6] R. J. Elliott, L. Aggoun, and J. B. Moore, *Hidden Markov Models: Estimation and Control*, Springer-Verlag, New York, 1994.

[7] E. Fernández-Gaucherand, A. Arapostathis, and S.I. Marcus, *Analysis of an adaptive control scheme for a partially observed Markov chain*, IEEE Trans. Automat. Control, 38 (1993), pp. 987–993.

[8] J. González-Hernández and O. Hernández-Lerma, *Envelopes of sets of measures, tightness, and Markov control processes*, Appl. Math. Optim., 40 (1999), pp. 377–392.

[9] O. Hernández-Lerma, *Adaptive Markov Control Processes*, Springer-Verlag, New York, 1989.

[10] O. Hernández-Lerma and J. B. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, Springer-Verlag, New York, 1996.

[11] O. Hernández-Lerma and J. B. Lasserre, *Further Topics on Discrete-Time Markov Control Processes*, Springer-Verlag, New York, 1999.

[12] O. Hernández–Lerma and J. B. Lasserre, *Fatou's lemma and Lebesgue's convergence theorem for measures*, J. Appl. Math. Stochastic Anal., 13 (2000), pp. 137–146.

[13] O. Hernández-Lerma and S. I. Marcus, *Adaptive control of Markov processes with incomplete state information and unknown parameters*, J. Optim. Theory Appl., 52 (1987), pp. 227–241.

[14] O. Hernández-Lerma and S. I. Marcus, *Nonparametric adaptive control of discrete-time partially observable stochastic systems*, J. Math. Anal. Appl., 137 (1989), pp. 312–334.
[15] N. Hilgert and O. Hernández-Lerma, *Limiting optimal discounted-cost control of a class of time-varying stochastic systems*, Systems Control Lett., 40 (2000), pp. 37–42.
[16] N. Hilgert, R. Senoussi, and J. P. Vila, *Nonparametric estimation of time-varying autoregressive nonlinear processes*, C.R. Acad. Sci. Paris (Sér. 1), 323 (1996), pp. 1085–1090.
[17] A. Hordijk and A. A. Yushkevich, *Blackwell optimality in the class of all policies in Markov decision chains with a Borel state space and unbounded rewards*, Math. Methods Oper. Res., 50 (1999), pp. 421–448.
[18] D. E. Lane, *A partially observable model of decision making by fishermen*, Oper. Res., 37 (1989), pp. 240–254.
[19] H.-J. Langen, *Convergence of dynamic programming models*, Math. Oper. Res., 6 (1981), pp. 493–512.
[20] J. A. Loewe, *Markov Decision Chains with Partial Information*, Ph.D. thesis, Department of Mathematics and Computer Science, Leiden University, Leiden, The Netherlands, 1995.
[21] G. E. Monahan, *A survey of partially observable Markov decision processes: Theory, models, and algorithms*, Manage. Sci., 28 (1982), pp. 1–16.
[22] A. Müller, *How does the value function of a Markov decision process depend on the transition probability?*, Math. Oper. Res., 22 (1997), pp. 872–885.
[23] K. R. Parthasarathy, *Probability Measures on Metric Spaces*, Academic Press, New York, 1971.
[24] U. Rieder, *Structural results for partially observed control models*, Z. Oper. Res., 35 (1991), pp. 473–490.
[25] W. J. Runggaldier and L. Stettner, *Approximations of Discrete Time Partially Observed Control Problems*, Applied Mathematics Monographs CNR 6, Giardini, Pisa, 1994.
[26] M. Schäl, *Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 32 (1975), pp. 179–196.
[27] C. Striebel, *Optimal Control of Discrete Time Stochastic Systems*, Lecture Notes in Econom. and Math. Systems 110, Springer-Verlag, Berlin, 1975.
[28] A. A. Yushkevich, *Reduction of a controlled Markov model with incomplete data to a problem with complete information in the case of Borel state and control spaces*, Theory Probab. Appl., 21 (1976), pp. 153–158.
[29] S. T. Rachev, *Probability Metrics and the Stability of Stochastic Models*, Wiley, Chichester, UK, 1991.
[30] O. Hernández-Lerma and J. B. Lasserre, *Order-Bounded Sequences of Measures*, manuscript, 1996.
[31] K. Yosida, *Functional Analysis*, 6th ed., Springer-Verlag, Berlin, 1980.
[32] R. Cavazos-Cadena and O. Hernández-Lerma, *Recursive adaptive control of Markov decision processes with the average reward criterion*, Appl. Math. Optim., 23 (1991), pp. 193–207.
[33] T. E. Duncan, B. Pasik-Duncan, and L. Stettner, *Adaptive control of discrete time Markov processes by the large deviations method*, Appl. Math. (Warsaw), 27 (2000), pp. 265–285.
[34] M. K. Ghosh and A. Bagchi, *Stochastic games with average payoff criterion*, Appl. Math. Optim., 38 (1998), pp. 283–301.
[35] O. Hernández-Lerma, R. Montes-de-Oca, and R. Cavazos-Cadena, *Recurrence conditions for Markov decision processes with Borel state space: A survey*, Ann. Oper. Res., 28 (1991), pp. 29–46.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 2, pp. 370–392

LIE ALGEBRAIC OBSTRUCTIONS TO Γ-CONVERGENCE OF OPTIMAL CONTROL PROBLEMS*

ARIELA BRIANI† AND FRANCO RAMPAZZO‡

Abstract. We investigate the possibility of describing the "limit problem" of a sequence of optimal control problems $(\mathcal{P})_{(b_n)}$, each of which is characterized by the presence of a time dependent vector valued coefficient $b_n = (b_{n_1}, \dots, b_{n_M})$. The notion of "limit problem" is intended in the sense of Γ-convergence, which, roughly speaking, prescribes the convergence of both the minimizers and the infimum values. Due to the type of growth involved in each problem $(\mathcal{P})_{(b_n)}$ the (weak) limit of the functions $(b^2_{n_1}, \dots, b^2_{n_M})$—beside the limit $(b_1, \dots, b_M)$ of the $(b_{n_1}, \dots, b_{n_M})$—is crucial for the description of the limit problem. Of course, since the b_n are L^2 maps, the limit of the $(b^2_{n_1}, \dots, b^2_{n_M})$ may well be a (vector valued) measure $\mu = (\mu_1, \dots, \mu_M)$. It happens that when the problems $(\mathcal{P})_{(b_n)}$ enjoy a certain commutativity property, then the pair (b, μ) is sufficient to characterize the limit problem.

This is no longer true when the commutativity property is not in force. Indeed, we construct two sequences of problems $(\mathcal{P})_{(b_n)}$ and $(\mathcal{P})_{(\tilde{b}_n)}$ which are equal except for the coefficient $b_n(\cdot)$ and $\tilde{b}_n(\cdot)$, respectively. Moreover, both the sequences (b_n, b_n^2) and $(\tilde{b}_n, \tilde{b}_n^2)$ converge to the same pair (b, μ). However, the infimum values of the problems $(\mathcal{P})_{(b_n)}$ tend to a value which is different from the limit of the infimum values of the $(\mathcal{P})_{(\tilde{b}_n)}$. This means that the mere information contained in the pair (b, μ) is not sufficient to characterize the limit problem. We overcome this drawback by embedding the problems in a more general setting where limit problems can be characterized by triples of functions (B_0, B, γ) with $B_0 \geq 0$.

Key words. Γ-convergence, optimal control problems, Lie brackets

AMS subject classifications. 49J15, 49J45, 93B29

PII. S0363012999363560

1. Introduction.

The general goal in the various theories of variational convergence consists in singling out a notion of limit problem $(\mathcal{P})$ for a sequence of minimum problems $(\mathcal{P}_n)$. Loosely speaking, this means that both the minimizers (provided they exist) of the problems $(\mathcal{P}_n)$ and the corresponding minimum values should converge (in some sense) to the minimizers and the minimum value of $(\mathcal{P})$, respectively.

In this paper we shall deal with the case where the minimum problems $(\mathcal{P}_n)$ have the form of the optimal control problems $(\mathcal{P})_{(b_n)}$ considered below. More precisely, the dependence on n follows by the fact that the dynamics of these problems contain n-dependent time functions b_n. We are motivated to study this particular problem essentially for two reasons. The first one is related to the general problem of homogenization (see, e.g., [BLP78], [LPV85], [SP80]). More specifically, for a control system one could think to the case where the dynamic contains a quite irregular time dependent coefficient. This would motivate the interest in looking for suitable topologies such that the approximation of this coefficient with regular functions would provide a "good" approximation of the given optimal control problem.

The second reason why we are studying the particular class of problems specified below is twofold. On one hand, this class of problems is general enough to display

*Received by the editors October 26, 1999; accepted for publication (in revised form) January 15, 2001; published electronically July 19, 2001.
http://www.siam.org/journals/sicon/40-2/36356.html
†Dipartimento di Matematica, Università di Pisa, Via Buonarroti 2, 56127 Pisa, Italy (briani@dm.unipi.it).
‡Dipartimento di Matematica Pura ed Applicata, Università di Padova, Via Belzoni 7, Padova, Italy (rampazzo@math.unipd.it).

the pathology related to Lie brackets of the involved vector fields (see below). On the other hand, the relatively simple structure of these problems allows one to avoid unessential technicalities which would obscure the nature of the question at issue.

Referring to the appendix for some basic tools of the general issue of variational convergence, let us specify the class of optimal control problems we are going to deal with.

Let $g_0, g_1, \dots, g_M$ be smooth vector fields, and let l, k_i, h_i be given real functions. We shall consider sequences of optimal control problems of the form

$(\mathcal{P})_{(b_n)}$

$$\begin{cases} \dot{x} = g_0(t,x) + \displaystyle\sum_{i=1}^{M} g_i(x) b_{n_i}(t) u_i(t), \quad x(0) = x_0, \\ \min\limits_{u} \left\{ J(x,u) = \displaystyle\int_0^T \left(l(t,x) + \sum_{i=1}^{M} k_i(t,x) b_{n_i}(t) u_i(t) + \sum_{i=1}^{M} h_i^2(x) u_i^2(t) \right) dt \right\}, \end{cases}$$

where $(b_n)_{n\in\mathbb{N}}$ is a sequence of $\mathbb{R}^M$-valued, time dependent coefficients.

We will investigate the Γ-limit (see the appendix) of problems $(\mathcal{P})_{(b_n)}$ when

$$\begin{array}{ll} \lim\limits_{n\to\infty} b_{n_i}(\cdot) = b_i(\cdot) & \text{weakly in } L^2(0,T), \\ \lim\limits_{n\to\infty} b_{n_i}^2(\cdot) = \mu_i(\cdot) & \text{weakly}^* \text{ in } \mathcal{M}([0,T]) \end{array} \tag{1.1}$$

for $i = 1, \dots, M$ (where $L^2(0,T)$ and $\mathcal{M}([0,T])$ denote the space of 2-integrable functions and the space of Borel measures, respectively).

We shall assume the following set of hypotheses on the data.

(Hg0) The function $g_0 : (0,T) \times \mathbb{R}^N \to \mathbb{R}^N$ is continuous. Moreover, for every compact subset $Q \subset \mathbb{R}^N$ there exists a continuous function $\gamma_0(t)$ such that, for every $t \in [0,T]$ and for every $x, y \in Q$, one has

$$|g_0(t,x) - g_0(t,y)| \le \gamma_0(t)|x-y|.$$

(Hg1) For each $i = 1, \dots, M$ the vector fields g_i from $\mathbb{R}^N$ into $\mathbb{R}^N$ are of class C^2, and the trajectories of the equations $\dot{x} = g_i(x)$ exist globally.

(Hb) For each $n \in N$, $b_n(t) = (b_{n_1}(t), \dots, b_{n_M}(t)) \in L^2(0,T;\mathbb{R}^M)$.

(Hu) The controls $u(t) = (u_1(t), \dots, u_M(t))$ belong to $L^2(0,T;\mathbb{R}^M)$.

(Hl) The function $l : [0,T] \times \mathbb{R}^N \to [0,\infty]$ is a Borel function, and for every compact subset $Q \subset \mathbb{R}^N$ there exists an L^1 function $\eta(t)$ such that, for every $t \in [0,T]$ and for every $x, y \in Q$,

$$|l(t,x) - l(t,y)| \le \eta(t)\,|x-y|\,.$$

Moreover, the function $l(t,0)$ belongs to $L^1(0,T)$.

(Hk) For each $i = 1, \dots, M$, $k_i : [0,T] \times \mathbb{R}^N \to [0,\infty]$ is a continuous function. There exists a constant $C > 0$ such that, for each $t \in [0,T]$ and for each $y \in \mathbb{R}^N$

$$|k_i(t,y)| \le C, \qquad i = 1, \dots, M. \tag{1.2}$$

Moreover, there exists a constant $L_k > 0$ such that, for each $t \in [0,T]$ and for each $y, z \in \mathbb{R}^N$,

$$|k_i(t,y) - k_i(t,z)| \le L_k |y-z|, \qquad i = 1, \dots, M. \tag{1.3}$$

(Hh) For each $i = 1, \dots, M$, $h_i : \mathbb{R}^N \to [0, \infty]$ is a Borel function, and for every compact subset $Q \subset \mathbb{R}^N$ there exists a constant L_h such that

$$|h_i(x) - h_i(y)| \leq L_h |x - y|, \quad i = 1, \dots, M$$

for every $x, y \in Q$. Moreover, we assume the following *coercivity* hypothesis. There exists a constant $K > 0$ such that, for every $x \in \mathbb{R}^N$,

$$\sum_{i=1}^{M} h_i^2(x) u_i^2 \geq K |u|^2$$

for every $u \in L^2(0, T)$.

Remark 1.1. Some of these hypotheses can be weakened further. For example, in view of section 5, the constants C and L_k in (1.2), (1.3) may be replaced by two functions in $L^1(0, T)$. Moreover, at the cost of some technical complications in the computation of the Γ-limit in Definition 2.3 below, the maps l, h_i, and k_i may be allowed to depend on n as well.

Let us begin by remarking that some authors (see, e.g., [BC89], [BF93], [Fr98]) studied this problem when the maps $g_1, \dots, g_M$, $h_1, \dots, h_M$ are *constant* and $k_i = 0$, $i = 1, \dots, M$. In particular, in [BF93], [Fr98] one studies the limit of these problems when the L^2 structural parameters $b_n(\cdot) = (b_{n_1}, \dots, b_{n_M})(\cdot)$ converge, say, weakly, to an L^2 map $b(\cdot) = (b_1, \dots, b_M)(\cdot)$. It turns out that in order to single out the limit problem one needs to know the (weak) limit $\mu = (\mu_1, \dots, \mu_M)(\cdot)$ of the maps $b_n^2 = (b_{n_1}^2, \dots, b_{n_M}^2)(\cdot)$ as well. Let us recall that this limit, when it exists, can well be different from $b^2(\cdot)$. (Actually, one has $\mu \geq b^2$.) Moreover, in general, it is not an L^1 function. Actually, it is a measure on $[0, T]$. The main point established in the quoted papers consists in the fact that the pair (b, μ) does single out the limit problem. This result relies upon a crucial assumption, namely, the fact that the g_i and the h_i are independent of x, which, in turn, allows one to regard the limit equation and the limit payoff as relations *in measure*. On the contrary, as soon as the g_i actually depend on x—and a certain commutativity assumption (see below) is not verified—the measure-theoretical approach does not work, as shown by the simple example in section 3.

In this paper we shall study the limit of problem $(\mathcal{P})_{(b_n)}$ when both *the g_i and the h_i can depend on x* and the k_i do not vanish.

Our aim is threefold. To begin with, in section 2 we assume a *commutativity* hypothesis, which generalizes the case where the g_i are constant. Namely, we assume that $[g_i, g_j] = 0$ for all $i, j = 1, \dots, M$ (plus the fact that the k_i and h_i are constant), where $[g_i, g_j]$ denotes the Lie bracket of the fields g_i and g_j. It is remarkable that, under this assumption, one can prove the same result as in the case where the g_i are x-independent. In other words, the limit problem of the $(\mathcal{P})_{(b_n)}$ is still singled out by the limit (b, μ) of the pairs (b_n, b_n^2). This limit is denoted by $\Phi^{-1}(Q_{(b,\mu)})$, for it is the preimage of a simpler problem $(\mathcal{Q})_{(b,\mu)}$ via a diffeomorphism Φ, which, in turn, is determined by the (commutative) fields g_i. The result in this section allows one to get a geometric insight into the results in [BC89], [BF93], [Fr98] as well, for, while the property of "being independent of x" is not chart-invariant, commutativity has an intrinsic meaning.

Second, in section 3 we present an example that reveals the crucial difference between the case with vanishing Lie brackets and the general case. Actually, in this example (the Lie brackets do not vanish and) two sequences $((b_n, b_n^2))_{n \in N}$ and

$((\tilde{b}_n, \tilde{b}_n^2))_{n\in N}$ converge to the *same* pair (b, μ), while the corresponding problems $(\mathcal{P})_{(\tilde{b}_n)}$ and $(\mathcal{P})_{(\tilde{b}_n)}$ converge to *different* limit problems. Hence, provided a limit problem exists (in some possibly extended sense), in order to characterize it one needs some "extra information" beside that contained in the assignment of the pair (b, μ).

The construction of an extended setting for problems with no commutativity assumptions is, in fact, the third aim of the paper. We pursue this objective in section 4 by redefining the minimum problems in the space of the graphs. Within this extended setting every minimum problem is identified by a triple of functions (B_0, B, γ) defined on $[0, 1]$, this triple replacing the role of the pair (b, μ). The map B_0, whose square root is the derivative of time t with respect to a *pseudotime parameter* s in the interval $[0, 1]$, assumes values greater than or equal to zero. A particular case is represented by the original problems $(\mathcal{P})_{(b_n)}$, which are identified with problems corresponding to triples of the form (B_{0_n}, B_n, B_n^2) with B_{0_n} *strictly greater than zero almost everywhere (a.e.) in* $[0, 1]$ and $B_n \doteq b_n B_{0_n}$. On the other hand, the extra information needed in order to single out the limit problem is provided by the restriction of γ to the subintervals of $[0, 1]$, where B_0 is equal to zero.

Last, in section 5 we prove some statements aiming to compose the (apparent) discrepancy between the case with vanishing Lie brackets—which is treated in section 2 in terms of the original time t—and the general case—which is addressed in section 4 in an extended framework. The key points consist in a projection of the set of triples (B_0, B, γ) onto the set of the pairs (b, μ) and in the consequent partition of the set of triples. Roughly speaking, when the commutativity hypothesis holds, all extended problems in a class of this partition correspond to a unique problem, namely, the one singled out by the (unique) projection (b, μ) of the triples in the class.

For the sake of self-consistency we conclude the paper with an appendix, where some basic facts from the general theory of Γ-convergence are briefly recalled.

Let us point out that a reader interested only in the case with vanishing Lie brackets may read just section 2. On the other hand, the construction of the extended setting for the general case, which is performed in section 4, is self-contained and independent of the antecedent material of the paper.

Notation. We will write $L^p(0, T; \mathbb{R}^M)$ to denote the space of p-integrable functions from $[0, T]$ into $\mathbb{R}^M$ endowed with the usual norm $\|\cdot\|_p$. Moreover, $\mathcal{M}([0, T]; \mathbb{R}^M)$ and $BV([0, T]; \mathbb{R}^M)$ will denote the space of $\mathbb{R}^M$-valued Borel measure on $[0, T]$ and the space of $\mathbb{R}^M$-valued functions with bounded variation on $[0, T]$, respectively. If $M = 1$, we write $L^p(0, T)$, $\mathcal{M}([0, T])$, $BV([0, T])$ instead of $L^p(0, T; \mathbb{R})$, $\mathcal{M}([0, T]; \mathbb{R})$, $BV([0, T]; \mathbb{R})$, respectively.

If $\mu \in \mathcal{M}([0, T])$, μ^a and μ^s stand for the absolutely continuous and the singular part of μ with respect to the Lebesgue measure dt, respectively. If μ_1 and μ_2 are a vector measure and a scalar measure on $[0, T]$, respectively, we write $\mu_1 << \mu_2$ to mean that μ_1 is absolutely continuous with respect to μ_2. Moreover, we denote the derivative of μ_1 with respect to μ_2 (in the sense of the Radon–Nikodym theorem) by $\frac{d\mu_1}{d\mu_2}$. Finally, by $\operatorname{supp}\mu$ we mean the support of the measure μ.

2. Null Lie brackets. We assume here the *commutativity condition* (HC) below, which, in particular, states that all Lie brackets $[g_i, g_j]$, $i, j = 1, \ldots, M$, are identically equal to zero. This hypothesis is crucial in order to prove a result of Γ-convergence (see the appendix for the definition of Γ-limit) analogous to the one proved in [Fr98], where the vectors multiplying the control were assumed x-independent. This fact allows one to get a geometric insight into the question, since the case with constant g_i

is nothing but a particular occurrence of the commutativity condition. We will see in the next sections that such a result does not hold when the commutativity assumption is not assumed.

Commutativity condition (HC). For every $i, j = 1, \dots, M$ the Lie bracket

$$[g_i, g_j](x) = Dg_j(x)g_i(x) - Dg_i(x)g_j(x)$$

(where $Dg(x)$ denotes the derivative of g at x) is identically equal to zero. Moreover, the maps h_i and k_i are constant. (See Remark 2.12 below for a comment on this latter condition.)

In order to define the Γ-limit, we introduce a suitable coordinate transformation which is induced by the fields $g_1, \dots, g_M$. This transformation is made possible by the crucial commutativity assumption (HC). Let us begin by adding the auxiliary equations $z_i(t) = \int_0^t b_{n_i}(s)u_i(s)ds$, $i = 1, \dots, M$. Then the state equation of $(\mathcal{P})_{(b_n)}$ reads as

$$\begin{pmatrix} \dot{z} \\ \dot{x} \end{pmatrix} = \tilde{g}_0(t, x) + \sum_{i=1}^{M} \tilde{g}_i(x) b_{n_i}(t) u_i(t),$$

where

$$\begin{array}{rccc} \tilde{g}_0 : & [0,T] \times \mathbb{R}^N & \to & \mathbb{R}^M \times \mathbb{R}^N, \\ & (t, x) & \mapsto & \begin{pmatrix} 0_M \\ g_0^1(t,x) \\ . \\ . \\ g_0^N(t,x) \end{pmatrix}, \end{array}$$

and, for every $i = 1, \dots, M$,

$$\begin{array}{rccc} \tilde{g}_i : & \mathbb{R}^N & \to & \mathbb{R}^M \times \mathbb{R}^N, \\ & x & \mapsto & \begin{pmatrix} e_i \\ g_i^1(x) \\ . \\ . \\ g_i^N(x) \end{pmatrix}, \end{array}$$

0_M and e_i being the zero vector and the ith (column) vector of the canonical basis in $\mathbb{R}^M$, respectively.

In the extended state space $\mathbb{R}^M \times \mathbb{R}^N$, problem $(\mathcal{P})_{(b_n)}$ is now formulated as

$$(\mathcal{P})_{(b_n)} \quad \begin{cases} \begin{pmatrix} \dot{z} \\ \dot{x} \end{pmatrix} = \tilde{g}_0(t, x) + \displaystyle\sum_{i=1}^{M} \tilde{g}_i(x) b_{n_i}(t) u_i(t), \ (z(0), x(0)) = (0, x_0), \\ \min\limits_{u} \left\{ J_n((z, x), u) = \displaystyle\int_0^T \left(l(t, x) + \sum_{i=1}^{M} k_i b_{n_i}(t) u_i(t) + \sum_{i=1}^{M} h_i^2 u_i^2(t) \right) dt \right\}. \end{cases}$$

(Notice that we use the same notation, namely, $(\mathcal{P})_{(b_n)}$, to mean both the problem in $\mathbb{R}^N$ and the corresponding one in $\mathbb{R}^M \times \mathbb{R}^N$.)

Let us set

$$\begin{aligned} &\Phi_1(z, x) = z, \\ &\Phi_2(z, x) = \exp(-z_M g_M) \circ \cdots \circ \exp(-z_1 g_1)x \end{aligned}$$

(where $\exp(sg)x$ stands for the value at time s of the solution of the Cauchy problem $\dot{y}(s) = g(y(s))$, $y(0) = x$), and let us consider the map Φ defined by

$$\begin{pmatrix} z \\ y \end{pmatrix} = \Phi(z,x) \doteq \begin{pmatrix} \Phi_1(z,x) \\ \Phi_2(z,x) \end{pmatrix}.$$

We shall also use the notations $(z, x(z,y))$ and $(z, y(z,x))$ instead of $\Phi^{-1}(z,y)$ and $\Phi(z,x)$, respectively. Notice that, since the maps g_i are of class C^2, Φ is a local diffeomorphism. Actually, Φ is a global diffeomorphism.

Let us define the vector fields $\check{g}_0 : (0,T) \times \mathbb{R}^M \times \mathbb{R}^N \to \mathbb{R}^M \times \mathbb{R}^N$ and $\check{g}_i : \mathbb{R}^M \times \mathbb{R}^N \to \mathbb{R}^M \times \mathbb{R}^N$, $i = 1, \dots, M$, by setting

$$\begin{aligned} \check{g}_0(t,z,y) &\doteq D\Phi(z,x)\, \tilde{g}_0(t,x), \\ \check{g}_i(z,y) &\doteq D\Phi(z,x)\, \tilde{g}_i(x), \end{aligned}$$

where $(z,x) = \Phi^{-1}(z,y)$. Notice that $\check{g}_0$ and $\check{g}_i$ are the expressions of $\tilde{g}_0$ and $\tilde{g}_i$, respectively, in the new coordinate (z,y).

PROPOSITION 2.1. *The first components of the vector field $\check{g}_0 : (0,T) \times \mathbb{R}^M \times \mathbb{R}^N \to \mathbb{R}^M \times \mathbb{R}^N$ are equal to zero, that is,*

$$\check{g}_0(t,z,y) = \begin{pmatrix} 0_M \\ g_0^\sharp(t,z,y) \end{pmatrix},$$

where the (column) vector field $g_0^\sharp(t,z,y)$ is given by $g_0^\sharp(t,z,y) = D_x\Phi_2(z,x)\tilde{g}_0(t,x)$ with $(z,x) = \Phi^{-1}(z,y)$. In particular, $\check{g}_0$ verifies (Hg0) *(with N replaced by $M+N$). Moreover, one has, for $i = 1, \dots, M$,*

$$\check{g}_i(z,y) = \begin{pmatrix} e_i \\ 0_N \end{pmatrix},$$

where 0_N stands for the (column) zero vector of $\mathbb{R}^N$.

A proof of this trivial proposition can be found in [BR91].

By means of this coordinate change, problem $(\mathcal{P})_{(b_n)}$ is transformed into the problem

$$(\mathcal{Q})_{(b_n)} \qquad \begin{cases} \begin{pmatrix} \dot{z} \\ \dot{y} \end{pmatrix} = \check{g}_0(t,z,y) + \displaystyle\sum_{i=1}^{M} \check{g}_i b_{n_i}(t) u_i(t), \qquad (z(0), y(0)) = (0, x_0), \\ \min\limits_u \left\{ \check{J}_n((z,y),u) \right\}, \end{cases}$$

$$\check{J}_n((z,y),u) = \int_0^T \left(l(t, x(z,y)) + \sum_{i=1}^{M} k_i b_{n_i}(t) u_i(t) + \sum_{i=1}^{M} h_i^2 u_i^2(t) \right) dt,$$

which, thanks to Proposition 2.1, displays the following, particularly simple, form:

$$(\mathcal{Q})_{(b_n)} \qquad \begin{cases} \dot{z}_1(t) = b_{n_1}(t) u_1(t), \\ \cdot \\ \cdot \\ \dot{z}_M(t) = b_{n_M}(t) u_M(t), \\ \\ \dot{y}(t) = g_0^\sharp(t,z,y), \\ \\ \min\limits_u \left\{ \check{J}_n((z,y),u) \right\}. \end{cases}$$

Remark 2.2. By saying that "$(\mathcal{P})_{(b_n)}$ is transformed into $(\mathcal{Q})_{(b_n)}$" we mean the following.

(i) A trajectory-control pair $((z,y),u)$ is admissible for the problem $(\mathcal{Q})_{(b_n)}$ if and only if the trajectory-control pair $((z,x),u) \doteq (\Phi^{-1}(z,y),u)$ is admissible for $(\mathcal{P})_{(b_n)}$.

(ii) For each trajectory-control pair $((z,y),u)$, if $((z,x),u) \doteq (\Phi^{-1}(z,y),u)$, then

$$J_n((z,x),u) = \check{J}_n(\Phi(z,x),u)$$

for every $u \in L^2(0,T;\mathbb{R}^M)$.

In particular, a trajectory-control pair $((z^\sharp,x^\sharp),u^\sharp)$ is optimal for $(\mathcal{P})_{(b_n)}$ if and only if $((z^\sharp,y^\sharp),u^\sharp)$ is optimal for $(\mathcal{Q})_{(b_n)}$, where $y^\sharp = \Phi_2(z^\sharp,x^\sharp)$.

In order to provide a representation of the Γ-limit of problems $(\mathcal{P})_{(b_n)}$ we shall be concerned with the set of *data pairs*

$$A = \left\{(b,\mu) \in L^2(0,T;\mathbb{R}^M) \times \mathcal{M}([0,T];\mathbb{R}^M) \ : \ \mu \geq b^2\right\},$$

where $\mu = (\mu_1,\dots,\mu_M)$, $b = (b_1,\dots,b_M)$, and the inequality has to be interpreted as $\mu_i \geq b_i^2$ for all $i = 1,\dots,M$ (in the measure-theoretical sense). In particular, we shall consider the subset $A_s \subset A$ defined by

$$A_s = \left\{(b,\mu) \in A \ : \ \mu = b^2\right\},$$

which we call the subset of *simple data pairs* of A. (We recall that b^2 denotes the vector $(b_1^2,\dots,b_M^2)$.)

DEFINITION 2.3. *Let $(b,\mu) \in A$, and let us set $\sigma = \sum_{i=1}^M \mu_i^s$. We consider the variational problem*

$$(\mathcal{Q})_{(b,\mu)} \qquad \min_{((z,y),u)} \left\{\check{J}((z,y),u) \ : \ \dot{z} << dt + \sigma, \quad \dot{y} = g_0^\sharp(t,z,y)\right\},$$

where the minimum is searched over the trajectory-control pairs $((z,y),u)$ in $BV([0,T];\mathbb{R}^M \times \mathbb{R}^N) \times L^2(0,T;\mathbb{R}^M)$ and the cost functional $\check{J}$ is defined by

$$\begin{aligned}
&\check{J}((z,y),u) \\
&\quad \doteq \int_0^T \left[l(t,w) + \sum_{i=1}^M \left(k_i \dot{z}_i^a(t) + h_i^2 u_i^2(t) + h_i^2 \frac{(b_i(t)u_i(t) - \dot{z}_i^a(t))^2}{(\mu_i^a(t) - b_i^2(t))} \right) \right] dt \\
&\quad + \int_{\Omega_s \setminus \{0,T\}} \sum_{i=1}^M \left(h_i^2 \left| \frac{d\dot{z}_i^s}{d\sigma} \right|^2 + k_i \left| \frac{d\dot{z}_i^s}{d\sigma} \right| \right) d\sigma \\
&\quad + \sum_{i=1}^M \left(h_i^2 \frac{|z_i(0^+) - z_i(0^-)|^2}{\sigma(\{0\})} + k_i \left| z_i(0^+) - z_i(0^-) \right| \right) \\
&\quad + \sum_{i=1}^M \left(h_i^2 \frac{|z_i(T) - z_i(T^-)|^2}{\sigma(\{T\})} + k_i \left| z_i(T^+) - z_i(T^-) \right| \right),
\end{aligned}$$

where we have set $w = x(z,y)$ and $\Omega_s = \operatorname{supp}\sigma$. (See section 1 for the notations in the above formula.)

Remark 2.4. We adopt here the convention (already used in [BC89], [BF93], [Fr98]) according to which the fractions appearing in the definition of $\check{J}$ are zero as soon as their denominators are zero.

Remark 2.5. If one has $b_i^2 = \mu_i$ for $i = 1, \dots, M$, then the limit problem $(\mathcal{Q})_{(b,\mu)}$ reduces to the standard form

$$\begin{cases} \begin{pmatrix} \dot z \\ \dot y \end{pmatrix} = \check g_0(t,z,y) + \sum_{i=1}^{M} \check g_i(z,y) b_i(t) u_i(t), \quad (z(0), y(0)) = (0, x_0), \\ \min\limits_u \big\{ \check J((z,y),u) \big\}, \end{cases}$$

$$\check J((z,y),u) = \int_0^T \left(l(t, x(z,y)) + \sum_{i=1}^{M} k_i b_i(t) u_i(t) + \sum_{i=1}^{M} h_i^2 u_i^2(t) \right) dt.$$

Definition 2.6. *Let us rewrite problem* $(\mathcal{Q})_{(b,\mu)}$ *in the form*

$$(\mathcal{Q})_{(b,\mu)} \qquad \min \left\{ \check F((z,y),u) : (z,y) \in BV([0,T]; \mathbb{R}^M \times \mathbb{R}^N),\ u \in L^2(0,T;\mathbb{R}^M) \right\},$$

where $\check F((z,y),u) \doteq \check J\,((z,y),u) + \chi_{\{\dot z << dt + \sigma\,,\ \dot y = g_0^\sharp(t,w)\}}$. *We define problem* $\Phi^{-1}((\mathcal{Q})_{(b,\mu)})$ *as follows:*

$$\Phi^{-1}((\mathcal{Q})_{(b,\mu)}) \qquad \min\{ \check F(\Phi((z,x)),u) : (z,x) \in BV([0,T]; \mathbb{R}^M \times \mathbb{R}^N), u \in L^2(0,T;\mathbb{R}^M) \}.$$

The next result states that problems $(\mathcal{P})_{(b_n)}$ converge to the variational problem $\Phi^{-1}((\mathcal{Q})_{(b,\mu)})$. For the basic facts concerning the Γ-convergence, see the appendix and the references therein.

Theorem 2.7. *If the* (b_n, b_n^2) *converge to* (b,μ) *as in* (1.1)*, then the problems* $(\mathcal{P})_{(b_n)}$ Γ*-converge to* $\Phi^{-1}((\mathcal{Q})_{(b,\mu)})$.

Proof. In view of Lemma 2.8 below we have to prove only that the $(\mathcal{Q})_{(b_n)}$ Γ-converge to $(\mathcal{Q})_{(b,\mu)}$. Now the optimal control problems $(\mathcal{Q})_{(b_n)}$ verify hypotheses (7.1)–(7.5) in [Fr98]. Moreover, assumption (1.1) here implies (7.17) and (7.18) therein. Hence, in view of the results in [Fr98], problems $((\mathcal{Q})_{(b_n)})_{n\in N}$ Γ-converge to the problem $(\mathcal{Q})_{(b,\mu)}$ introduced in Definition 2.3. □

Lemma 2.8. *If the sequence of problems* $((\mathcal{Q})_{(b_n)})_{n\in N}$ Γ*-converges to* $(\mathcal{Q})_{(b,\mu)}$*, then the sequence* $((\mathcal{P})_{(b_n)})_{n\in N}$ Γ*-converges to* $\Phi^{-1}((\mathcal{Q})_{(b,\mu)})$.

Proof. To begin with, for each $n \in N$ we set

$$\check F_n((z,y),u) \doteq \check J_n((z,y),u) + \chi_{\check C_n}((z,y),u),$$

where $\check C_n$ is the set of admissible trajectory-control pairs for $(\mathcal{Q})_{(b_n)}$ (see the appendix). By assumption we have (see Definition A.3 in the appendix)

$$\Gamma(N, U^-, Y^-) \lim_{n\to\infty} \check F_n((z,y),u) = \check F((z,y),u). \tag{2.1}$$

Now (see Remark 2.2)

$$\begin{aligned} F_n((z,x),u) &\doteq J_n((z,x),u) + \chi_{C_n}((z,x),u) \\ &= \check J_n(\Phi(z,x),u) + \chi_{\check C_n}(\Phi(z,x),u)) = \check F_n(\Phi(z,x),u), \end{aligned}$$

where C_n is the set of admissible trajectory-control pairs for $(\mathcal{P})_{(b_n)}$. Hence, by (2.1),

$$\begin{aligned} &\Gamma(N, U^-, Y^-) \lim_{n\to\infty} F_n((z,x),u) \\ &\quad = \Gamma(N, U^-, Y^-) \lim_{n\to\infty} \check F_n(\Phi(z,x),u) = \check F(\Phi(z,x),u), \end{aligned}$$

which proves the lemma. □

Theorem 2.7 says that the Γ-limit of a sequence of problems $(\mathcal{P})_{(b_n)}$ has the form $\Phi^{-1}((\mathcal{Q})_{(b,\mu)})$. Conversely, we have the following.

THEOREM 2.9. *For each problem* $\Phi^{-1}((\mathcal{Q})_{(b,\mu)})$ *with* $(b,\mu) \in A$, *there exists a sequence of problems* $((\mathcal{P})_{(b_n)})_{n\in N}$ *which* Γ*-converges to* $\Phi^{-1}((\mathcal{Q})_{(b,\mu)})$.

In order to prove this theorem, we need the following result.

LEMMA 2.10. *For each* $(b,\mu) \in A$ *(with* $M = 1$*) there exists a sequence* $(b_n)_{n\in N} \in L^2(0,T)$ *such that* $b_n \to b$ *weakly in* $L^2(0,T)$ *and* $b_n^2 \to \mu$ *weakly* in* $\mathcal{M}([0,T])$.

In the case where μ is an L^∞-function we can sharpen the above result as follows.

LEMMA 2.11. *If* $(b,\mu) \in A$ *(with* $M = 1$*) and* $\mu \in L^\infty(0,T)$, *then there exists a sequence* $(b_n)_{n\in N} \in L^2(0,T)$ *such that* $b_n \to b$ *weakly in* $L^2(0,T)$ *and* $b_n^2 \to \mu$ *weakly* in* $L^\infty(0,T)$.

We omit the proofs of both Lemmas 2.10 and 2.11, for they are mostly based on the same arguments as in the proof of Theorem 3.2 in [BR93].

Proof of Theorem 2.9. In view of Lemma 2.10, for each $(b,\mu) \in L^2(0,T;\mathbb{R}^M) \times \mathcal{M}([0,T];\mathbb{R}^M)$ such that $(b,\mu) \in A$ there exist sequences $(b_{n_i})_{n\in N}$ in $L^2(0,T)$ such that $b_{n_i} \to b_i$ weakly in $L^2(0,T)$ and $b_{n_i}^2 \to \mu_i$ weakly* in $\mathcal{M}([0,T])$ for $i = 1,\dots,M$. Hence, in view of Theorem 2.7, the sequence of problems $((\mathcal{P})_{(b_n)})_{n\in N}$ Γ-converges to $\Phi^{-1}((\mathcal{P})_{(b,\mu)})$. □

Remark 2.12. By the above arguments it is clear that we could replace hypothesis (HC) with the following more general assumption (GHC), which, on one hand, does not assume that the functions k_i and h_i are constant and, on the other hand, involves these functions in the zero-Lie bracket condition.

Generalized commutativity condition (GHC). For every $\alpha,\beta = 1,\dots,2M$

$$[\gamma_\alpha,\gamma_\beta] = 0,$$

where the vector fields γ_δ are defined on $\mathbb{R}^{N+2}$ by

$$\gamma_\delta = \begin{pmatrix} g_i^1 \\ . \\ . \\ g_i^N \\ k_\delta \\ 0 \end{pmatrix}$$

when $\delta = 1,\dots,M$, and

$$\gamma_\delta = \begin{pmatrix} 0 \\ . \\ . \\ 0 \\ 0 \\ h_\delta \end{pmatrix}$$

when $\delta = M+1,\dots,2M$.

3. Nonvanishing Lie brackets: An example. In the previous section it has been shown that whenever the vector fields commute the Γ-limit of problems $(\mathcal{P})_{(b_n)}$ for (b_n, b_n^2) converging to (b,μ) *does exist*. However, this is no longer true whenever some Lie bracket is not vanishing, as shown in the example below. In the next sections

we will provide a theoretical framework from which it will be clear that, in general, there exist infinitely many limit problems corresponding to the pair (b, μ).

In order to get rid of the suspicion that having a state's dimension larger than the control's dimension might matter with the convergence question, the state in this example is one-dimensional.

Let $N = 1$, $M = 2$, and consider the state equation

$$\begin{cases} \dot{x}(t) = b_{n_1}(t)u_1(t) + a(x(t))b_{n_2}(t)u_2(t), \\ x(0) = 0, \end{cases}$$

where $a(x)$ is a bounded C^2 function coinciding with the identity map in the interval $[-4, 4]$. Hence $g_1(x)$ coincides with the constant 1, and $g_2(x) = a(x)$. Let us assume (Hb), (Hu), $T = 1$, and let us consider the cost functional

$$J_{b_n}(x, u) = \int_0^1 \left(|u(t)|^2 + b_{n_1}(t)u_1(t) + a(x(t))b_{n_2}(t)u_2(t)\right) dt \left(= \int_0^1 |u(t)|^2 \, dt + x(1)\right).$$

If we set $h_1(x) = 1$, $h_2(x) = 1$, $k_1(t, x) = 1$, $k_2(t, x) = a(x)$, and $l(t, x) = 0$, the hypotheses in section 2 turns out to be satisfied.

Since $[g_1, g_2](0) = -1$, neither the commutativity condition (HC) nor its generalization (GHC) are fulfilled.

Let us consider the two sequences of coefficients

$$(b_{n_1}(t), b_{n_2}(t)) \doteq \left(\sqrt{2n}, 0\right) \mathbf{I}_{\left[1-\frac{1}{n}, 1-\frac{1}{2n}\right]}(t) + \left(0, \sqrt{2n}\right) \mathbf{I}_{\left[1-\frac{1}{2n}, 1\right]}(t),$$

$$\left(\tilde{b}_{n_1}(t), \tilde{b}_{n_2}(t)\right) \doteq \left(0, \sqrt{2n}\right) \mathbf{I}_{\left[1-\frac{1}{n}, 1-\frac{1}{2n}\right]}(t) + \left(\sqrt{2n}, 0\right) \mathbf{I}_{\left[1-\frac{1}{2n}, 1\right]}(t),$$

where $\mathbf{I}_{[a,b]} = 1$ if $t \in [a, b]$ and $\mathbf{I}_{[a,b]} = 0$ if $t \notin [a, b]$. Let us observe that

$$(b_{n_1}(t), b_{n_2}(t)) \to (0, 0) \qquad \text{weakly in } L^2(0, T),$$

$$\left(\tilde{b}_{n_1}(t), \tilde{b}_{n_2}(t)\right) \to (0, 0) \qquad \text{weakly in } L^2(0, T),$$

$$\left(b^2_{n_1}(t), b^2_{n_2}(t)\right) \to (\delta_1, \delta_1) \qquad \text{weakly}^* \text{ in } \mathcal{M}([0, T]),$$

$$\left(\tilde{b}^2_{n_1}(t), \tilde{b}^2_{n_2}(t)\right) \to (\delta_1, \delta_1) \qquad \text{weakly}^* \text{ in } \mathcal{M}([0, T]),$$

where δ_1 denotes the Dirac measure at $T = 1$. Hence the two sequences fulfill the convergence assumption (1.1) with the *same* limit $(b_1, b_2) = (0, 0)$ and $(\mu_1, \mu_2) = (\delta_1, \delta_1)$. Yet the corresponding sequences $((\mathcal{P})_{(b_n)})_{n\in N}$ and $((\mathcal{P})_{(\tilde{b}_n)})_{n\in N}$ *cannot* converge to the same Γ-limit. Indeed, if we implement the control

$$(u_1^n(t), u_2^n(t)) = \left(-\sqrt{\frac{n}{2}} \, \mathbf{I}_{\left[1-\frac{1}{n}, 1-\frac{1}{2n}\right]}(t), 0\right) + \left(0, \sqrt{n} \, \mathbf{I}_{\left[1-\frac{1}{2n}, 1\right]}(t)\right)$$

in the system driven by the (b_n), we obtain a trajectory x_n verifying

$$x_n(1) = -\frac{1}{2} \exp(2^{-1/2}).$$

Thus

$$J_{b_n}(x_n, u_n) = K \doteq -\frac{1}{2}\exp(2^{-1/2}) + \frac{3}{4}\left(< -\frac{1}{4}\right).$$

On the contrary, if we consider $(\mathcal{P})_{(\tilde{b}_n)}$, a simple application of the Pontryagin maximum principle shows that

$$(\hat{u}_1^n(t), \hat{u}_2^n(t)) = \left(-\frac{1}{2}\,\tilde{b}_{n_1}, 0\right)$$

is an optimal control. The corresponding optimal trajectory $\hat{x}_n$ solves

$$\dot{\hat{x}}_n(t) = -\frac{1}{2}\,\tilde{b}^2_{n_1}, \qquad \hat{x}_n(0) = 0.$$

Hence

$$-\frac{1}{4} = J_{\tilde{b}_n}(\hat{x}_n, \hat{u}_n) = \min_u\{J_{\tilde{b}_n}(x, u)\}.$$

In particular, one has

$$\liminf_{n\to\infty}\left(\inf_u\{J_{b_n}(x,u)\}\right) \le K < -\frac{1}{4} = \liminf_{n\to\infty}\left(\min_u\{J_{\tilde{b}_n}(x,u)\}\right).$$

Hence, although the (b_{n_1}, b_{n_2}) and $(\tilde{b}_{n_1}, \tilde{b}_{n_2})$ converge to the same (b, μ) in the sense of (1.1), in view of Theorem A.2 in the appendix the Γ-limit of the $(\mathcal{P})_{(b_n)}$ and $(\mathcal{P})_{(\tilde{b}_n)}$ are necessarily different.

4. Nonvanishing Lie brackets: An extended setting. In this section we still assume hypotheses (Hg0), (Hg1), (Hl), (Hk), and (Hh), but we *do not* assume the commutativity hypothesis (HC) made in section 2. The previous example shows that in order to determine the limit problem *it is not enough* to assume that $b_n \to b$ weakly in $L^2(0,T;\mathbb{R}^M)$ and $b_n^2 \to \mu$ weakly* in $\mathcal{M}([0,T];\mathbb{R}^M)$. In fact, due to the noncommutativity of the vector fields g_i $(i = 1, \dots, M)$, some extra information—related to the choice of the particular sequence (b_n, b_n^2) approximating (b, μ)—is needed. It turns out that this extra information can be represented neatly by first embedding the problem in the (t, x)-space and then reparameterizing time with a nondecreasing map whose derivative is zero for those values of t where μ is concentrated. In particular, this embedding allows one to keep track of the particular sequence (b_n, b_n^2) approximating (b, μ).

Let us begin with some definitions.

Definition 4.1. *The set of* data triples *is defined as*

$$\mathcal{A} \doteq \Bigg\{(B_0, B, \gamma) :\ B_0 : [0,1] \to \mathbb{R}^+ \cup \{0\}\ ,\ B : [0,1] \to \mathbb{R}^M,$$

$$\gamma : [0,1] \to (\mathbb{R}^+ \cup \{0\})^M \textit{ are Borel functions in } L^\infty(0,T;\mathbb{R}^M)\ :$$
$$\gamma_i \ge B_i^2 \ \textit{ for all } i = 1, \dots, M \textit{ and } \int_0^1 B_0^2(s)ds = T\Bigg\}.$$

The subset A_{NI} *of* nonimpulsive data triples *is defined as*

$$\mathcal{A}_{NI} \doteq \{(B_0, B, \gamma) \in \mathcal{A} \ :\ B_0 > 0 \ \textit{a.e. on } [0,1]\}.$$

The subset of simple data triples *is defined as*

$$\mathcal{A}_s \doteq \left\{(B_0, B, \gamma) \in \mathcal{A} \,:\, \gamma_i = B_i^2 \text{ a.e. on } [0,1]\,,\ i = 1, \dots, M\ \right\}.$$

We will denote the vector $(B_1^2, \dots, B_M^2)$ by B^2.

For each triple $(B_0, B, \gamma) \in \mathcal{A}$ let us consider the *space-time* optimal control problem

$$(\mathcal{P})_{(B_0,B,\gamma)} \qquad \begin{cases} y'(s) = g_0(s,y)B_0^2(s) + \displaystyle\sum_{i=1}^{M} g_i(y)V_i(s), \quad y(0) = y_0, \\ \min\limits_{U,V} \left\{\hat{J}(y,U,V)\right\}, \end{cases}$$

$$\begin{aligned} \hat{J}(y,U,V) \;=\; & \int_0^1 \left(l(s,y)B_0^2(s) + \sum_{i=1}^{M} k_i(s,y)V_i(s) + \sum_{i=1}^{M} h_i^2(y)U_i^2(s) \right. \\ & \left. +\ \sum_{i=1}^{M} h_i^2(y) \frac{(B_i(s)U_i(s) - V_i(s))^2}{(\gamma_i(s) - B_i^2(s))} \right) ds, \end{aligned}$$

where $U \in L^2(0,1;\mathbb{R}^M)$ and $V \in L^2(0,1;\mathbb{R}^M)$.

Remark 4.2. When $(B_0, B, \gamma) \in \mathcal{A}_s$, that is, $\gamma_i = B_i^2$, $i = 1, \dots, M$, the optimal control problem $(\mathcal{P})_{(B_0,B,\gamma)}$ reduces to the following standard form:

$$(\mathcal{P})_{(B_0,B,B^2)} \qquad \begin{cases} y'(s) = g_0(s,y)B_0^2(s) + \displaystyle\sum_{i=1}^{M} g_i(y)B_i(s)U_i(s), \quad y(0) = y_0, \\ \min\limits_{U} \left\{\hat{J}(y,U)\right\}, \end{cases}$$

$$\hat{J}(y,U) = \int_0^1 \left(l(s,y)B_0^2(s) + \sum_{i=1}^{M} k_i(s,y)B_i(s)U_i(s) + \sum_{i=1}^{M} h_i^2(y)U_i^2(s) \right) ds.$$

We shall show that the class of problems $(\mathcal{P})_{(b)}$—where $(\mathcal{P})_{(b)}$ stands for a problem like $(\mathcal{P})_{(b_n)}$ when b_n is replaced by b—can be put into one-to-one correspondence with the class of space-time problems $\{(\mathcal{P})_{(B_0,B,B^2)} \,:\, (B_0, B, B^2) \in \mathcal{A}_s \cap \mathcal{A}_{NI}\}$. Then we shall give sufficient conditions for the Γ-convergence of a sequence of problems $(\mathcal{P})_{(B_{0_n},B_n,B_n^2)}$ to a problem $(\mathcal{P})_{(B_0,B,\gamma)}$. Last, we shall see that *every* such problem is the Γ-limit of a suitable sequence of problems $(\mathcal{P})_{(B_{0_n},B_n,B_n^2)}$.

DEFINITION 4.3. *Given* $(B_0, B, \gamma) \in \mathcal{A}$, *let us define* $\alpha(B_0, B, \gamma) \doteq (b, \mu)$ *by setting the following.*

(i) $$t(s) \doteq \int_0^s B_0^2(u)du,$$

and, whenever there exists $\delta > 0$ *such that* $B_0 > 0$ *a.e. on* $[s-\delta, s+\delta] \cap [0,1]$,

$$b_i(t(s)) \doteq \frac{B_i(s)}{B_0(s)} \qquad (i = 1, \dots, M).$$

(ii) *For each Borel subset* $E \subseteq [0,T]$

$$\mu_i(E) \doteq \int_I \gamma_i(s)ds \qquad (i = 1, \dots, M)$$

when $E = t(I)$.

Remark 4.4. (a) The function $b(\cdot)$ is well defined. Indeed, the set of values of t such that t^{-1} is not a singleton is at most countable.

(b) For each $(B_0, B, \gamma) \in \mathcal{A}$ the pair $(b, \mu) = \alpha(B_0, B, \gamma)$ is in A; in particular, if $(B_0, B, B^2) \in \mathcal{A}_{NI} \cap \mathcal{A}_s$, then $\alpha(B_0, B, B^2) = (b, b^2) \in A_s$.

(c) The definition of μ is equivalent to

$$\int_{[0,T]} \phi(t)d\mu = \int_0^1 \langle \gamma_i(s), \phi(t(s))\rangle ds \quad \text{for all } \phi \in C([0,T]; \mathbb{R}^M), \tag{4.1}$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product in $\mathbb{R}^M$.

(d) The map α is not injective, unless it is restricted to $\mathcal{A}_{NI} \cap \mathcal{A}_s$. Let us show that it is surjective. Indeed, for every (b, μ) let us set

$$s(t) = \begin{cases} 0 & t = 0, \\ \dfrac{t + \int_{]0,t]} d\mu}{T + \int_{]0,T]} d\mu} & 0 < t < T, \\ 1 & t = T, \end{cases}$$

and let us define $t(s)$ as the unique nondecreasing continuous map such that $t \circ s(\tau) = \tau$ for all $\tau \in [0,T]$. Correspondingly, let us set $B_0^2(s) = t'(s)$, $B_i(s) = b_i(t(s))B_0(s)$, $s \in [0,1]$, $i = 1, \dots, M$. Finally, let us choose $\gamma_i(s)$ such that

$$\int_{s_1}^{s_2} \gamma_i(s)ds = \int_{t(s_1,s_2)} d\mu_i, \quad i = 1, \dots, M,$$

for each subinterval (s_1, s_2) of $[0,1]$. Then $\alpha(B_0, B, \gamma) = (b, \mu)$. We will call this data triple *the canonical preimage* of (b, μ). Let us notice that for every (b, μ), $\alpha^{-1}(b, \mu)$ turns out to be the class of data triples in $\mathcal{A}$ such that

$$(B_0, B, \gamma), (\tilde{B}_0, \tilde{B}, \tilde{\gamma}) \in \alpha^{-1}(b, \mu) \quad \Leftrightarrow \quad \begin{cases} B_0 = \tilde{B}_0 \text{ a.e. }, \\ B = \tilde{B} \text{ a.e. }, \\ \gamma_i(s) = \tilde{\gamma}_i(s) \text{ for a.e. } s \in [0,1] \setminus \cup I_j \\ \text{ and } \displaystyle\int_{I_j} \gamma_i(s)ds = \int_{I_j} \tilde{\gamma}_i(s)ds \quad \text{for all } j, \end{cases}$$

where $\{I_j\}$ is the (countable) family of (disjoint) subintervals of $[0,1]$ such that $B_0 = \tilde{B}_0 = 0$ on each I_j.

In the following two theorems we establish a one-to-one correspondence between the class of problems $(\mathcal{P})_{(b)}$, $b \in L^2(0,T;\mathbb{R}^M)$, and the class of problems $(\mathcal{P})_{(B_0,B,\gamma)}$, $(B_0, B, \gamma) \in \mathcal{A}_{NI} \cap \mathcal{A}_s$ (i.e., $B^2 = \gamma$ and $B_0 > 0$). Before stating these results let us notice that α is one-to-one from $\mathcal{A}_{NI} \cap \mathcal{A}_s$ onto A_s.

THEOREM 4.5. *Let* b *and* u *satisfy* (Hb) *and* (Hu), *and let* $x(\cdot)$ *be the corresponding solution of the state equation of* $(\mathcal{P})_{(b)}$

$$\begin{cases} \dot{x} = g_0(t,x) + \displaystyle\sum_{i=1}^M g_i(x)b_i(t)u_i(t), \\ x(0) = x_0. \end{cases} \tag{4.2}$$

Let $(B_0, B, B^2) = \alpha^{-1}(b, b^2)$*, and set* $U(s) \doteq [u \circ t(s)]B_0(s)$*. Let* y *be the solution of the state equation of* $(\mathcal{P})_{(B_0,B,B^2)}$

$$\begin{cases} y'(s) = g_0(s,y)B_0^2(s) + \sum_{i=1}^{M} g_i(y)B_i(s)U_i(s), \\ y(0) = y_0. \end{cases} \tag{4.3}$$

Then

$$y(s) = x(t(s)) \quad \text{for all } s \in [0,1].$$

Conversely, let $(B_0, B, B^2) \in \mathcal{A}_s \cap \mathcal{A}_{NI}$*,* $U \in L^2(0,1;\mathbb{R}^M)$*, and let* $y(\cdot)$ *be the corresponding solution of* (4.3)*. Setting* $(b, b^2) = \alpha(B_0, B, B^2)$*, let us define*

$$u_i(t) \doteq \frac{U_i(s(t))}{B_0(s(t))}, \quad i = 1, \dots, M.$$

If $x(\cdot)$ *is the solution of* (4.2) *corresponding to these* b_i *and* u_i*, then*

$$x(t) = y(s(t)) \quad \text{for all } t \in [0,T].$$

Proof. The proof of this theorem relies essentially on the uniqueness properties of (4.2) and (4.3). For this reason we omit it. □

An analogous result holds for the payoffs J and $\hat{J}$.

THEOREM 4.6. *Consider* b, u, x, B_0, B, U*, and* y *as in the first part of Theorem* 4.5*, and set*

$$J(x,u) = \int_0^T \left(l(t,x) + \sum_{i=1}^{M} k_i(t,x)b_i(t)u_i(t) + \sum_{i=1}^{M} h_i^2(x)u_i^2(t) \right) dt,$$

$$\hat{J}(y,U) = \int_0^1 \left(l(s,y)B_0(s) + \sum_{i=1}^{M} k_i(s,y)B_i(s)U_i(s) + \sum_{i=1}^{M} h_i^2(y)U_i^2(s) \right) ds.$$

Then $\hat{J}(y,U) = J(x,u)$*. Conversely, if* B_0, B, U, y *and* b, u, x *are as in the second part of Theorem* 4.5*, then* $J(x,u) = \hat{J}(y,U)$*.*

Proof. In view of Theorem 4.5 the proof of this theorem is straightforward. □

When the problems $(\mathcal{P})_{(b)}$ and $(\mathcal{P})_{(B_0,B,B^2)}$ are related as in the previous result, we say that they are *isomorphic*. In view of Theorems 4.5 and 4.6 the map $(\mathcal{P})_{(b)} \mapsto (\mathcal{P})_{(B_0,B,B^2)}$ with $\alpha(B_0, B, B^2) = (b, b^2)$ establishes a *one-to-one correspondence between the class of problems* $\{(\mathcal{P})_{(b)}\ ,\ b \in L^2(0,T;\mathbb{R}^M)\}$ *and the subset* $\{(\mathcal{P})_{(B_0,B,\gamma)}\ ,\ (B_0, B, \gamma) \in \mathcal{A}_s \cap \mathcal{A}_{NI}\} \subset \{(\mathcal{P})_{(B_0,B,\gamma)}\ ,\ (B_0, B, \gamma) \in \mathcal{A}\}$. This one-to-one correspondence can be regarded as an embedding of the original class of problems $\{(\mathcal{P})_{(b)}\ :\ b \in L^2(0,T;\mathbb{R}^M)\}$ in the larger class $\{(\mathcal{P})_{(B_0,B,\gamma)}\ ,\ (B_0, B, \gamma) \in \mathcal{A}\}$. In this extended setting we are now able to provide a convergence result, so giving an answer to the question raised with the example in section 3. In other words, we are going to replace the assumptions on the sequence $((\mathcal{P})_{(b_n)})_{n\in N}$ with hypotheses on the sequence of isomorphic problems $(\mathcal{P}_{(B_{0_n},B_n,B_n^2)})_{n\in N}$. And assigning the limit of the triples (B_{0_n}, B_n, B_n^2), we actually provide the extra information whose lack was revealed by the example of section 3.

Here is the main result.

THEOREM 4.7. *Let* $(B_{0_n}, B_n, B_n^2) \in \mathcal{A}_s \cap \mathcal{A}_{NI}$ *and* $(B_0, B, \gamma) \in \mathcal{A}$ *verify*

$$\lim_{n\to\infty} B_{0_n}(\cdot) = B_0(\cdot) \text{ a.e. on } [0,1], \tag{4.4}$$

$$\lim_{n\to\infty} B_n(\cdot) = B(\cdot) \text{ weakly in } L^1(0,1;\mathbb{R}^M), \tag{4.5}$$

$$\lim_{n\to\infty} B_n^2(\cdot) = \gamma(\cdot) \text{ weakly in } L^1(0,1;\mathbb{R}^M). \tag{4.6}$$

Then problems $\mathcal{P}_{(B_{0_n},B_n,B_n^2)}$ Γ*-converge to problem* $(\mathcal{P})_{(B_0,B,\gamma)}$.

Remark 4.8. In the previous statement one possibly has $\alpha(B_{0_n}, B_n, B_n^2) = (b_n, b_n^2)$ with $b_n \in L^2(0,T;\mathbb{R}^M)$. So, in particular, Theorem 4.7 can be regarded as a convergence result concerning the original problems $(\mathcal{P})_{(b_n)}$.

Proof of Theorem 4.7. Thanks to the performed rescaling of the problem, we can exploit the general results proved by Buttazzo and Cavazzuti in [BC89]. Actually, hypotheses (Hg0), (Hg1), (Hl), (Hk), (Hh), and (4.4)–(4.6) imply (3.6)–(3.10) and (3.12)–(3.15) in [BC89], respectively. Hence Propositions 3.2 and 3.3 of [BC89] state that $(\mathcal{P})_{(B_0,B,\gamma)}$ is the Γ-limit of the $\mathcal{P}_{(B_{0_n},B_n,B_n^2)}$. □

Similarly to what has been done in the case where the Lie brackets vanish, we now prove that *each* problem $(\mathcal{P})_{(B_0,B,\gamma)}$ with $(B_0, B, \gamma) \in \mathcal{A}$ is indeed the Γ-limit of a sequence of problems of the form $\mathcal{P}_{(B_{0_n},B_n,B_n^2)}$ with $B_{0_n} > 0$ (which, up to the introduced one-to-one correspondence, means that $(\mathcal{P})_{(B_0,B,\gamma)}$ is the Γ-limit of problems $(\mathcal{P})_{(b_n)}$ with $(b_n, b_n^2) = \alpha(B_{0_n}, B_n, B_n^2)$).

THEOREM 4.9. *For every* $(B_0, B, \gamma) \in \mathcal{A}$*, the problem* $(\mathcal{P})_{(B_0,B,\gamma)}$ *is the* Γ*-limit of a suitable sequence* $(\mathcal{P}_{(B_{0_n},B_n,B_n^2)})_{n\in N}$ *with* $(B_{0_n}, B_n, B_n^2) \in \mathcal{A}_s \cap \mathcal{A}_{NI}$.

Proof. By Lemma 2.11 for each $(B_0, B, \gamma) \in A$ there is a sequence $((B_0, B_n, B_n^2))_{n\in N}$ $(\in \mathcal{A}_s)$ such that $B_{n_i} \to B_i$ weakly in $L^2(0,1)$ and $B_{n_i}^2 \to \gamma_i$ weakly in $L^1(0,1)$ for $i = 1, \dots, M$.

Moreover, by setting

$$B_{0_n}(s) = \sqrt{\frac{T}{T + \frac{1}{n}} \left(B_0^2(s) + \frac{1}{n} \right)},$$

we find that the triples (B_{0_n}, B_n, B_n^2) (belong to $\mathcal{A}_s \cap \mathcal{A}_{NI}$ and) verify (4.4)–(4.6). Hence one concludes by Theorem 4.7. □

5. Revisiting sections 2 and 3 in the light of the extended setting. On one hand, sections 2 and 3 reveal a crucial discrepancy between the case when all the brackets $[g_i, g_j]$ vanish identically and the general case. On the other hand, in section 4 we have introduced an extended setting in order to state a convergence result in the general case. In this section we are going to revisit both the positive result of section 2 and the counterexample of section 3 in light of the theory developed in section 4. Let us recall that the map $\alpha : \mathcal{A} \to A$ induces a one-to-one correspondence between the subset $\mathcal{A}_{NI} \cap \mathcal{A}_s \subset \mathcal{A}$ and $A_s \subset A$.

Null Lie brackets. In Theorem 5.1 below we show—under the commutativity hypothesis (HC) in section 2—that when problems $(\mathcal{P})_{(B_{0_n},B_n,B_n^2)}$ Γ-converge to a problem $(\mathcal{P})_{(B_0,B,\gamma)}$, then the space-projected problems $(\mathcal{P})_{(b_n)}$ with $(b_n, b_n^2) = \alpha(B_{0_n}, B_n, B_n^2)$ Γ-converge to the projected limit $\Phi^{-1}((\mathcal{Q})_{(b,\mu)})$, where $(b, \mu) = \alpha(B_0, B, \gamma)$.

The most relevant point of this theorem consists in the fact that two sequences $(B_{0_n}, B_n, B_n^2), (\tilde{B}_{0_n}, \tilde{B}_n, \tilde{B}_n^2)$ converging to two *different* triples $(B_0, B, \gamma), (\tilde{B}_0, \tilde{B}, \tilde{\gamma})$ such that $(b, \mu) = \alpha(B_0, B, \gamma) = \alpha(\tilde{B}_0, \tilde{B}, \tilde{\gamma})$ give rise to problems $(\mathcal{P})_{(b_n)}$ and $(\mathcal{P})_{(\tilde{b}_n)}$ Γ-converging to the *same* limit problem $\Phi^{-1}((\mathcal{Q})_{(b,\mu)})$. In particular, this explains why as soon as all the brackets vanish there is in fact no need of the extended setting.

THEOREM 5.1. *Let us assume the hypotheses of section* 2 *(in particular, the commutativity hypothesis* (HC)*). Given* $(B_0, B, \gamma) \in \mathcal{A}$*, let us consider* any *sequence* $(B_{0_n}, B_n, B_n^2) \in \mathcal{A}_{NI} \cap \mathcal{A}_s$ *such that the* B_{0_n} *are equibounded and*

$$\lim_{n\to\infty} B_{0_n}(\cdot) = B_0(\cdot) \text{ a.e. on } [0,1], \tag{5.1}$$

$$\lim_{n\to\infty} B_n(\cdot) = B(\cdot) \text{ weakly in } L^1(0,1;\mathbb{R}^M), \tag{5.2}$$

$$\lim_{n\to\infty} B_n^2(\cdot) = \gamma(\cdot) \text{ weakly in } L^1(0,1;\mathbb{R}^M) \tag{5.3}$$

(so that, by Theorem 4.7, $(\mathcal{P})_{(B_{0_n}, B_n, B_n^2)}$ Γ*-converges to* $(\mathcal{P})_{(B_0,B,\gamma)}$*). Then, setting* $(b_n, b_n^2) = \alpha(B_{0_n}, B_n, B_n^2)$ *and* $(b, \mu) = \alpha(B_0, B, \gamma)$*, one has that the problems* $(\mathcal{P})_{(b_n)}$ Γ*-converge to* $\Phi^{-1}((\mathcal{Q})_{(b,\mu)})$.

Proof. In view of Theorem 2.7, we have to prove only that

$$b_n \to b \text{ weakly in } L^2(0,T;\mathbb{R}^M), \tag{5.4}$$

$$b_n^2 \to \mu \text{ weakly}^* \text{ in } \mathcal{M}([0,T];\mathbb{R}^M). \tag{5.5}$$

We begin by observing that hypotheses (5.1)–(5.3) imply

$$\lim_{n\to\infty} B_{0_n}^2(\cdot) = B_0^2(\cdot) \text{ a.e. on } (0,1),$$

$$\lim_{n\to\infty} B_n(\cdot)B_{0_n}(\cdot) = B(\cdot)B_0(\cdot) \text{ weakly in } L^1(0,1;\mathbb{R}^M),$$

and

$$\lim_{n\to\infty} B_n^2(\cdot) = \gamma(\cdot) \text{ weakly in } L^1(0,T;\mathbb{R}^M).$$

Set

$$t_n(s) = \int_0^s B_{0_n}^2(u)du \quad \text{and} \quad t(s) = \int_0^s B_0^2(u)du.$$

Since the t_n tend to t pointwise and each t_n is increasing, the t_n tend to t uniformly on [0,1]. Hence, for each $\varphi \in C([0,T];\mathbb{R}^M)$,

$$\lim_{n\to\infty} \left| \int_0^T \langle b_n^2(t), \varphi(t)\rangle dt - \int_{[0,T]} \varphi(t)d\mu \right|$$
$$= \lim_{n\to\infty} \left| \int_0^1 \langle B_n^2(s), \varphi(t_n(s))\rangle ds - \int_0^1 \langle \gamma(s), \varphi(t(s))\rangle ds \right| = 0,$$

which proves (5.5).

In order to prove (5.4), let us observe that for any function $\varphi \in C_c^\infty(0,T;\mathbb{R}^M)$

$$\lim_{n\to\infty}\left|\int_0^T \langle b_n(t),\varphi(t)\rangle dt - \int_0^T \langle b(t),\varphi(t)\rangle dt\right| = \lim_{n\to\infty}\left|\int_0^1 \langle B_n(s)B_{0_n}(s),\varphi(s)\rangle ds - \int_0^1 \langle B(s)B_0(s),\varphi(s)\rangle ds\right| = 0.$$

Since the B_n are uniformly bounded in $L^2(0,T;\mathbb{R}^M)$, (5.4) follows by the density of $C_c^\infty(0,T;\mathbb{R}^M)$ in $L^2(0,T;\mathbb{R}^M)$. □

A compactness result. We now examine the converse situation, where a sequence $(b_n)_{n\in N}$ is given such that $((b_n,b_n^2))_{n\in N}$ converges—with some regularity—to $(b,\mu)\in A$. We do not assume here that the Lie brackets vanish. It turns out that a subsequence of the corresponding triples (B_{0_n},B_n,B_n^2) converges to an element $(B_0,B,\gamma)\in\alpha^{-1}(b,\mu)$.

THEOREM 5.2. *Assume the hypotheses of section* 2, *with the exclusion of the commutativity hypothesis* (HC). *Let* $\mathcal{T}=\{t_i, i\in N\}$, *a (countable) subset of* $[0,T]$, *and let* $(b,\mu)\in A$ *and* $(b_n)_{n\in N}$ *be given such that*

(i) $\mu = b^2+\mu^\tau$ *with* μ^τ *a (positive) measure concentrated in* $\mathcal{T}$;

(ii) *for each* $n\in N$, $b_n\in C([0,T]\setminus\mathcal{T};\mathbb{R}^M)$, *and*

$$b_n(\cdot)\to b(\cdot) \text{ uniformly on the compact subsets of } [0,T]\setminus\mathcal{T}, \tag{5.6}$$

$$b_n \to b \text{ weakly in } L^2(0,T;\mathbb{R}^M),$$

$$b_n^2 \to \mu \text{ weakly}^* \text{ in } \mathcal{M}([0,T];\mathbb{R}^M). \tag{5.7}$$

(So, if the commutative hypothesis (HC) *is in force,* $(\mathcal{P})_{(b_n)}\Gamma$*-converges to* $(\mathcal{P})_{(b,\mu)}$.*) Then, setting* $(B_{0_n},B_n,B_n^2)\doteq\alpha^{-1}(b_n,b_n^2)$, *there exists a subsequence* $(\check{B}_{0_n},\check{B}_n,\check{B}_n^2)$ *and a data triple* $(B_0,B,\gamma)\in\alpha^{-1}(b,\mu)$, *such that the problems* $(\mathcal{P})_{(\check{B}_{0_n},\check{B}_n,\check{B}_n^2)}$ Γ*-converge to* $(\mathcal{P})_{(B_0,B,\gamma)}$.

The proof of this theorem relies essentially on the following lemma.

LEMMA 5.3. *Assume the hypotheses of Theorem* 5.2. *Let us set*

$$s(t)\doteq\begin{cases}0, & t=0,\\ \dfrac{t+\int_{]0,t]}d\mu}{T+\int_{]0,T]}d\mu}, & 0<t<T,\\ 1, & t=T,\end{cases}$$

$$s_n(t)\doteq\frac{\int_0^t(1+|b_n|^2(s))ds}{\int_0^T(1+|b_n|^2(s))ds},$$

and let us define $t_n(s)$ *and* $t(s)$ *as the inverse of* $s_n(t)$ *and the unique nondecreasing continuous map such that* $t\circ s(\tau)=id_{[0,T]}$, *respectively. Then* $t_n(\cdot)$ *and* $t(\cdot)$ *are equi-Lipschitz continuous and*

$$\lim_{n\to\infty} t_n(s)=t(s) \tag{5.8}$$

uniformly on $[0,1]$. *Moreover, setting* $B_0(s) \doteq \sqrt{t'(s)}$, *one has*

$$B_0(s) = 0 \quad \textit{for all } s \in \mathrm{int}(t^{-1}(\mathcal{T})) \tag{5.9}$$

and

$$B_{0_n}(s) \to B_0 \ \textit{a.e. on } [0,1]. \tag{5.10}$$

Proof. Since $t^{-1}(\mathcal{T}) = \cup_{t_i \in \mathcal{T}} t^{-1}(t_i)$, we immediately obtain (5.9). Let us observe that hypothesis (5.7) implies that

$$\lim_{n\to\infty} s_n(t) = s(t) \quad \text{a.e. on } [0,T]$$

(see, e.g., Proposition 7.19 in [Fo84]).

Actually, by the continuity of $s(t)$ on $[0,T] \setminus \mathcal{T}$, one has

$$\lim_{n\to\infty} s_n(t) = s(t) \quad \text{for all } t \in [0,T] \setminus \mathcal{T}.$$

Moreover, by the monotonicity of the $s_n(\cdot)$ and of $s(\cdot)$, it follows that

$$\lim_{\varepsilon\to 0} \lim_{n\to\infty} (s_n(t+\varepsilon) - s_n(t-\varepsilon)) = s(t^+) - s(t^-),$$

which yields

$$\lim_{n\to\infty} t_n(s) = t(s) \quad \text{uniformly on } [0,1]. \tag{5.11}$$

Since $B^2_{0_n} = t'_n \geq 0$ for every n, there exists a subsequence, still denoted with $(B_{0_n})_{n\in N}$, such that

$$B^2_{0_n} = t'_n \to 0 \quad \text{a.e. on } \mathrm{int}(t^{-1}(\mathcal{T})).$$

The convergence of B_{0_n} to B_0 on the set $\mathrm{int}(t^{-1}(\mathcal{T}))$ is proved. In order to conclude, let us prove this convergence for every $s \in [0,1] \setminus t^{-1}(\mathcal{T})$. Indeed, by (5.6) and (5.11) one has

$$\lim_{n\to\infty} B^2_{0_n}(s) = \lim_{n\to\infty} t'_n(s) = \lim_{n\to\infty} \frac{1}{\dot{s}(t_n(s))} = \lim_{n\to\infty} \frac{\int_0^T (1+b_n(u)^2)du}{1+b_n^2(t_n(s))}$$

$$= \frac{T + \int_{[0,T]} d\mu}{1+b^2(t(s))} = B_0^2(s).$$

The lemma is proved. □

Proof of Theorem 5.2. By Theorem 4.7 in section 4 we have to prove only that the there exists a subsequence $(\check{B}_{0_n}, \check{B}_n, \check{B}^2_n)$ of (B_{0_n}, B_n, B^2_n) and a data triple $(B_0, B, \gamma) \in \alpha^{-1}(b,\mu)$ such that

$$\lim_{n\to\infty} \check{B}_{0_n}(\cdot) = B_0(\cdot) \text{ a.e. on } [0,1], \tag{5.12}$$

$$\lim_{n\to\infty} \check{B}_n(\cdot) = B(\cdot) \text{ weakly in } L^1(0,1;\mathbb{R}^M), \tag{5.13}$$

$$\lim_{n\to\infty} \check{B}^2_n(\cdot) = \gamma(\cdot) \text{ weakly in } L^1(0,1;\mathbb{R}^M). \tag{5.14}$$

Let us define B_0 as in Lemma 5.3, which yields (5.12).

Moreover, since the B_n are equibounded, there is a subsequence $(\tilde{B}_n)_{n\in N}$ of $(B_n)_{n\in N}$ converging to a map B weakly in $L^1(0,1;\mathbb{R}^M)$.

By Ascoli–Arzela's theorem there exists a subsequence $(\phi_n)_{n\in N}$ of

$$\tilde{\phi}_n(s) \doteq \int_0^s \tilde{B}_n(\sigma)^2 d\sigma$$

converging to a Lipschitz continuous map ϕ. Then the subsequence $\check{B}^2_n(s) \doteq \frac{d\phi_n(s)}{ds}$ converges weakly* in $L^\infty(0,T;\mathbb{R}^M)$ to $\gamma(s) \doteq \frac{d\phi(s)}{ds}$, which implies (5.14).

We claim that $(B_0, B, \gamma) \in \alpha^{-1}(b,\mu)$. Indeed, $\check{B}_{0_n}(s)$ tends to $B_0(s)$ for every s such that $B_0 > 0$ a.e. on $[s-\delta, s+\delta]$ for a sufficiently small δ. Moreover, thanks to (5.11) and hypothesis (5.6), $b_n(t_n(s))$ converges to $b(t(s))$ for every such point s. Since $\check{B}_n(s)$ tends a.e. to $B(s)$, (i) in the definition of the mapping α (Definition 4.3) turns out to be satisfied. Finally, in view of (5.14) and (4.1), (ii) in the definition of α holds true as well. □

Revisiting the example of section 3. Let us conclude by framing the example of section 3 in the extended setting. This will clarify that the distinct limiting behavior of problems $(\mathcal{P})_{(b_n)}$ and problems $(\mathcal{P})_{(\tilde{b}_n)}$ arises from the fact that the corresponding triples $(B_{0_n}, B_n, B^2_n), (\tilde{B}_{0_n}, \tilde{B}_n, \tilde{B}^2_n)$ converge to different limits.

Let us recall that the state equation and the cost functional were given by

$$\begin{cases} \dot{x}(t) = b_{n_1}(t)u_1(t) + a(x(t))b_{n_2}(t)u_2(t), \\ x(0) = 0 \end{cases}$$

and

$$J_n(x,u) = \int_0^1 \left(|u(t)|^2 + b_{n_1}(t)u_1(t) + a(x(t))b_{n_2}(t)u_2(t)\right) dt \left(= \int_0^1 |u(t)|^2 dt + x(1)\right),$$

respectively. The problem of minimizing $J_n(x,u)$ over the controls $u \in L^2(0,T)$ was denoted by $(\mathcal{P})_{(b_n)}$ and $(\mathcal{P})_{(\tilde{b}_n)}$ when the parameters were identified with

$$(b_{n_1}(t), b_{n_2}(t)) = \left(\sqrt{2n}, 0\right)\mathbf{I}_{\left[1-\frac{1}{n}, 1-\frac{1}{2n}\right]}(t) + \left(0, \sqrt{2n}\right)\mathbf{I}_{\left[1-\frac{1}{2n}, 1\right]}(t)$$

and

$$\left(\tilde{b}_{n_1}(t), \tilde{b}_{n_2}(t)\right) = \left(0, \sqrt{2n}\right)\mathbf{I}_{\left[1-\frac{1}{n}, 1-\frac{1}{2n}\right]}(t) + \left(\sqrt{2n}, 0\right)\mathbf{I}_{\left[1-\frac{1}{2n}, 1\right]}(t),$$

respectively. Following the construction performed in section 4, let us compute the isomorphic problems $\mathcal{P}_{(B_0, B_n, B^2_n)}$ and $\mathcal{P}_{(\tilde{B}_0, \tilde{B}_n, \tilde{B}^2_n)}$ with $(B_0, B_n, B^2_n) = \alpha^{-1}(b_n, b^2_n)$ and $(\tilde{B}_0, \tilde{B}_n, \tilde{B}^2_n) = \alpha^{-1}(\tilde{b}_n, \tilde{b}^2_n)$. In both cases the optimal control problem turns out to have the form

$$\begin{cases} y'(s) = B_{n_1}(s)U_1(s) + a(y(s))B_{n_2}(s)U_2(s)\ ,\ y(0) = 0, \\ \min\limits_U \{\hat{J}_n(y,U)\}, \end{cases}$$

$$\hat{J}_n(y,U) = \int_0^1 \left(|U(s)|^2 + B_{n_1}(s)U_1(s) + a(y(s))B_{n_2}(s)U_2(s)\right) ds$$

with the parameters B identified with

$$\begin{aligned} B_{0_n}(s) &= \sqrt{3}\,\mathrm{I}_{\left[0,\frac{1}{3}\left(1-\frac{1}{n}\right)\right]} + \sqrt{\frac{3}{1+2n}}\,\mathrm{I}_{\left[\frac{1}{3}\left(1-\frac{1}{n}\right),1\right]}, \\ B_n(s) &= (B_{n_1}(s), B_{n_2}(s)) \\ &= \left(\sqrt{\frac{6n}{1+2n}}, 0\right)\,\mathrm{I}_{\left[\frac{1}{3}\left(1-\frac{1}{n}\right),\frac{1}{3}\left(2-\frac{1}{2n}\right)\right]} + \left(0, \sqrt{\frac{6n}{1+2n}}\right)\,\mathrm{I}_{\left[\frac{1}{3}\left(2-\frac{1}{2n}\right),1\right]}, \end{aligned}$$

and

$$\begin{aligned} \tilde{B}_{0_n}(s) &= \sqrt{3}\,\mathrm{I}_{\left[0,\frac{1}{3}\left(1-\frac{1}{n}\right)\right]} + \sqrt{\frac{3}{1+2n}}\,\mathrm{I}_{\left[\frac{1}{3}\left(1-\frac{1}{n}\right),1\right]}, \\ \tilde{B}_n(s) &= (\tilde{B}_{n_1}(s), \tilde{B}_{n_2}(s)) \\ &= \left(0, \sqrt{\frac{6n}{1+2n}}\right)\,\mathrm{I}_{\left[\frac{1}{3}\left(1-\frac{1}{n}\right),\frac{1}{3}\left(2-\frac{1}{2n}\right)\right]} + \left(\sqrt{\frac{6n}{1+2n}}, 0\right)\,\mathrm{I}_{\left[\frac{1}{3}\left(2-\frac{1}{2n}\right),1\right]}, \end{aligned}$$

respectively. In order to find the Γ-limit of problems $\mathcal{P}_{(B_0,B_n,B_n^2)}$ and $\mathcal{P}_{(\tilde{B}_0,\tilde{B}_n,\tilde{B}_n^2)}$, we need to compute the limits appearing in hypotheses (4.4), (4.5), and (4.6).

For the data triples (B_0, B_n, B_n^2), we obtain

$$\begin{aligned} \lim_{n\to\infty} B_{0_n}(s) = \quad & B_0(s) \doteq \sqrt{3}\,\mathrm{I}_{[0,1/3]} && \text{a.e. } [0,1], \\ \lim_{n\to\infty} B_{n_1}(s) = \quad & B_1(s) \doteq \sqrt{3}\,\mathrm{I}_{[1/3,2/3]} && \text{in } L^1(0,1), \\ \lim_{n\to\infty} B_{n_2}(s) = \quad & B_2(s) \doteq \sqrt{3}\,\mathrm{I}_{[2/3,1]} && \text{in } L^1(0,1), \\ \lim_{n\to\infty} B_{n_1}^2(s) = \quad & B_1^2(s) = 3\ \mathrm{I}_{[1/3,2/3]} && \text{in } L^1(0,1), \\ \lim_{n\to\infty} B_{n_2}^2(s) = \quad & B_2^2(s) = 3\,\mathrm{I}_{[2/3,1]} && \text{in } L^1(0,1), \end{aligned}$$

while, for the data triples $(\tilde{B}_0, \tilde{B}_n, \tilde{B}_n^2)$, we have

$$\begin{aligned} \lim_{n\to\infty} \tilde{B}_{0_n}(s) = \quad & \tilde{B}_0(s) \doteq \sqrt{3}\,\mathrm{I}_{[0,1/3]} && \text{a.e. } [0,1], \\ \lim_{n\to\infty} \tilde{B}_{n_1}(s) = \quad & \tilde{B}_1(s) \doteq \sqrt{3}\,\mathrm{I}_{[2/3,1]} && \text{in } L^1(0,1), \\ \lim_{n\to\infty} \tilde{B}_{n_2}(s) = \quad & \tilde{B}_2(s) \doteq \sqrt{3}\,\mathrm{I}_{[1/3,2/3]} && \text{in } L^1(0,1), \\ \lim_{n\to\infty} \tilde{B}_{n_1}^2(s) = \quad & \tilde{B}_1^2(s) = 3\,\mathrm{I}_{[2/3,1]} && \text{in } L^1(0,1), \\ \lim_{n\to\infty} \tilde{B}_{n_2}^2(s) = \quad & \tilde{B}_2^2(s) = 3\,\mathrm{I}_{[1/3,2/3]} && \text{in } L^1(0,1). \end{aligned}$$

Let us remark that the limits of (B_0, B_n, B_n^2) and $(\tilde{B}_0, \tilde{B}_n, \tilde{B}_n^2)$ are *different*. This explains why problems $(\mathcal{P})_{(b_n)}$ and $(\mathcal{P})_{(\tilde{b}_n)}$ in the example cannot converge to the same limit problem. More precisely, in view of Theorem 4.7, problems $\mathcal{P}_{(B_0,B_n,B_n^2)}$ and $\mathcal{P}_{(\tilde{B}_0,\tilde{B}_n,\tilde{B}_n^2)}$ Γ-converge to the optimal control problems

$$\begin{cases} y'(s) = B_1(s)U_1(s) + a(y(s))B_2(s)U_2(s), \quad y(0) = 0, \\ \min\limits_U \{\hat{J}(y,U)\}, \end{cases}$$

$$\hat{J}(y,U) = \int_0^1 \left(|U(s)|^2 + B_1(s)U_1(s) + a(y(s))B_2(s)U_2(s)\right) ds$$

and

$$\begin{cases} y'(s) = \tilde{B}_1(s)U_1(s) + a(y(s))\tilde{B}_2(s)U_2(s), \quad y(0) = 0, \\ \min\limits_{U}\{\hat{J}(y,U)\}, \end{cases}$$

$$\hat{J}(y,U) = \int_0^1 \big(|U(s)|^2 + \tilde{B}_1(s)U_1(s) + a(y(s))\tilde{B}_2(s)U_2(s)\big)ds,$$

respectively.

Appendix. Basic tools from Γ-convergence applied to control theory. Since the work of Wijsman [Wi64], [Wi66], many different concepts of convergence for sequences of functionals and operators have been appearing in the literature. These concepts were especially designed to approach the limit of sequences of variational problems. Each type of variational problem (minimization, maximization, min-max, etc.) has been associated to a particular concept of convergence.

In the case of the *minimization problem,* the first concept of convergence was the so-called *epiconvergence.* The epiconvergence of a sequence of functionals is equivalent to set-convergence of the corresponding epigraphs.

In turn, this concept was placed in the general framework of Γ-convergence theory by De Giorgi. The theory of Γ-convergence aims to deduce the asymptotic behavior of the solutions of a sequence of variational problems from the asymptotic behavior of the corresponding functionals. Typical examples of applications of Γ-convergence are the theories of homogenization, of singular perturbations, and of the limit behavior of elliptic problems with various obstacles. (We refer, e.g., to the books of Attouch [At84], Bensoussan, Lions, and Papanicolau [BLP78], Sanchez-Palencia [SP80], Dal Maso [DM93], and Buttazzo [Bu89].)

In this paper we have studied the Γ-convergence of sequences of optimal control problems. Let us sketch the general framework of this branch of the theory of Γ-convergence. For each $n \in N$ let $C_n \subseteq Y \times \mathcal{U}$ denote the set of *admissible trajectory-control pairs* defined by

$$C_n \doteq \{(y,u) \in Y \times \mathcal{U} \,:\, A_n(y) = B_n(u)\},$$

where A_n and B_n map Y and $\mathcal{U}$, respectively, in a third topological space V. Correspondingly, let us consider the optimal control problems

$$(\mathcal{P}_n) \qquad \min\{J_n(y,u) \,:\, (y,u) \in C_n\},$$

where J_n is a real operator defined on $Y \times \mathcal{U}$.

Setting $F_n(y,u) = J_n(y,u) + \chi_{C_n}(y,u)$ (where χ_E is 1 on E and $+\infty$ on $(Y \times \mathcal{U}) \setminus E$), let us rephrase problems $(\mathcal{P}_n)$ as follows:

$$(\mathcal{P}_n) \qquad \min\{F_n(y,u) \,:\, (y,u) \in Y \times \mathcal{U}\}.$$

We will say that (y_n, u_n) is an *optimal pair* for the problem $(\mathcal{P}_n)$ if

$$F_n(y_n,u_n) = \min_{Y\times\mathcal{U}} F_n.$$

Via Theorem A.2 below, the theory of Γ-convergence provides a notion of the limit problem guaranteeing the following property. *If (y_n, u_n) is an optimal pair of*

$(\mathcal{P}_n)$, or simply a minimizing sequence, and if (y_n, u_n) tends to (y, u) in $Y \times \mathcal{U}$, then (y, u) is an optimal pair for the limit problem $(\mathcal{P})$.

Theorem A.2 below provides a notion of Γ-*limit* problem $(\mathcal{P})$ such that this property holds. In order to state this theorem, we recall the definition of the multiple Γ-limit operator (see [BDM82]). We shall denote the "sup" and the "inf" operators by $Z(+)$ and $Z(-)$, respectively.

DEFINITION A.1. *Let X and W be two topological spaces, and let $F_n : X \times W \to \overline{\mathbb{R}}$ be a sequence of functions. For every $x \in X$, $w \in W$, and $\alpha, \beta, \gamma \in \{+, -\}$, let us define the* Γ-limit of the F_n *by setting*

$$\Gamma(N^\alpha, X^\beta, W^\gamma) \lim_{n\to\infty} F_n(w, x) = \mathop{\mathrm{Z}(\beta)}_{(x_n)\in S(x)} \mathop{\mathrm{Z}(\gamma)}_{(w_n)\in S(w)} \mathop{\mathrm{Z}(-\alpha)}_{k\in N} \mathop{\mathrm{Z}(\alpha)}_{n\geq k} F_n(w_n, x_n),$$

where $S(x)$ and $S(w)$ denote the sets of all sequences $x_n \to x$ in X and $w_n \to w$ in W, respectively. When the Γ*-limit does not depend on the sign $+$ or $-$, this sign is omitted. For example, if*

$$\Gamma(N^+, X^-, W^+) \lim_{n\to\infty} F_n(w, x) = \Gamma(N^+, X^+, W^+) \lim_{n\to\infty} F_n(w, x),$$

their common value will be indicated by $\Gamma(N^+, X, W^+) \lim_{n\to\infty} F_n(w, x)$.

In particular,

$$\Gamma(N, \mathcal{U}^-, Y^-) \lim_{n\to\infty} F_n(y, u) = \inf_{(u_n)\in S(u)} \inf_{(y_n)\in S(y)} \lim_{n\to\infty} F_n(y_n, u_n).$$

THEOREM A.2. *Let Y and $\mathcal{U}$ be two topological spaces, and let $F_n \colon Y \times \mathcal{U} \to \overline{\mathbb{R}}$ be a sequence of functions. For each $n \in N$, let (y_n, u_n) be a minimum point for F_n or simply a pair such that*

$$\lim_{n\to\infty} F_n(y_n, u_n) = \lim_{n\to\infty} [\inf_{Y\times\mathcal{U}} F_n].$$

Assume that the (y_n, u_n) converge to (y, u) in $Y \times \mathcal{U}$ and there exists

$$F(y, u) \doteq \Gamma(N, \mathcal{U}^-, Y^-) \lim_{n\to\infty} F_n(y_n, u_n). \tag{A.1}$$

Then

(i) *(y, u) is a minimum point for F on $Y \times \mathcal{U}$;*

(ii) $\lim_{n\to\infty} [\inf_{Y\times\mathcal{U}} F_n] = \min_{Y\times\mathcal{U}} F(y, u)$.

(For the proof see [BDM82, Proposition 2.1, p. 388].)

Note that if $F_n(y, u) \doteq F(u)$, then the Γ-limit $F(y, u)$ in (A.1) coincides with the so-called relaxed functional $\overline{F}$ (see, e.g., [Bu89]).

The above theorem motivates the following definition of the Γ-limit problem.

DEFINITION A.3. *When* (A.1) *is verified we say that the problem*

$$\min \{F(y, u) \,:\, (y, u) \in Y \times \mathcal{U}\} \tag{$\mathcal{P}$}$$

is the Γ-limit of problems $(\mathcal{P}_n)$.

See, e.g., [BDM82], [BC89], [BF93], [BF95], and [Fr98] for the explicit calculation of the Γ-limits in various interesting situations.

REFERENCES

[At84] H. Attouch, *Variational Convergence for Functions and Operators*, Pitman, London, 1984.

[BLP78] A. Bensoussan, J. L. Lions and G. Papanicolau, *Asymptotic Analysis for Periodic Structures*, North-Holland, Amsterdam, 1987.

[BR91] A. Bressan and F. Rampazzo, *Impulsive control systems with commutative vector fields*, J. Optim. Theory Appl., 71 (1991), pp. 67–83.

[BR93] A. Bressan and F. Rampazzo, *On differential systems with quadratic impulses and their applications to lagrangian mechanics*, SIAM J. Control Optim., 31 (1993), pp. 1205–1220.

[Bu89] G. Buttazzo, *Semicontinuity, Relaxation and Integral Representation in the Calculus of Variation*, Pitman Res. Notes Math. Ser. 207, Longman, Harlow, UK, 1989.

[BC89] G. Buttazzo and E. Cavazzuti, *Limit problems in optimal control theory*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 6 (1989), pp. 151–160.

[BDM82] G. Buttazzo and G. Dal Maso, *Γ-convergence and optimal control problems*, J. Optim. Theory Appl., 38 (1982), pp. 382–407.

[BF93] G. Buttazzo and L. Freddi, *Sequences of optimal control problems with measures as controls*, Adv. Math. Sci. Appl., 2 (1993), pp. 215–230.

[BF95] G. Buttazzo and L. Freddi, *Optimal control problems with weakly converging input operators*, Discrete Contin. Dynam. Systems, 1 (1995), pp. 401–420.

[DM93] G. Dal Maso, *An Introduction to Γ-Convergence*, Birkhäuser, Boston, 1993.

[Fo84] G. B. Folland, *Real Analysis Modern Techniques and Their Applications*, John Wiley and Sons, New York, 1984.

[Fr98] L. Freddi, *Optimal control problems with eakly converging input operators in a non reflexive framework*, Portugal. Math., 57 (2000), pp. 97–126.

[LPV85] P. L. Lions, G. Papanicolau, and S. R. S. Varadan, *Homogenization of Hamilton-Jacobi Equations*, Preprint, 1985.

[SP80] E. Sanchez-Palencia, *Nonhomogeneous Media and Vibration Theory*, Lecture Notes in Physics 127, Springer-Verlag, Berlin, New York, 1980.

[Wi64] R. A. Wijsman, *Convergence of sequences of convex sets, cones and functions* I, Bull. Amer. Math. Soc. (N.S.), 70 (1964), pp. 186–188.

[Wi66] R. A. Wijsman, *Convergence of sequences of convex sets, cones and functions* II, Trans. Amer. Math. Soc., 123 (1966), pp. 32–45.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 2, pp. 393–420

CONVERGENCE OF THE OPTIMAL FEEDBACK POLICIES IN A NUMERICAL METHOD FOR A CLASS OF DETERMINISTIC OPTIMAL CONTROL PROBLEMS*

PAUL DUPUIS† AND ADAM SZPIRO‡

Abstract. We consider a Markov chain based numerical approximation method for a class of deterministic nonlinear optimal control problems. It is known that methods of this type yield convergent approximations to the value function on the entire domain. These results do not easily extend to the optimal control, which need not be uniquely defined on the entire domain. There are, however, regions of strong regularity on which the optimal control is well defined and smooth. Typically, the union of these regions is open and dense in the domain. Using probabilistic methods, we prove that, on the regions of strong regularity, the Markov chain method yields a convergent sequence of approximations to the optimal feedback control. The result is illustrated with several examples.

Key words. optimal control, numerical approximation, rate of convergence, Markov chain approximation, feedback controls, finite difference approximation

AMS subject classifications. 35B37, 49L20, 60F05, 65N06, 65N12, 93E25

PII. S0363012998344968

1. Introduction.

In this paper, we prove that an efficient Markov chain based numerical approximation method for a general class of nonlinear optimal control problems yields feedback controls which converge (on most of the domain) to the optimal feedback control for the problem that is being approximated. We consider an infinite time horizon problem on a finite domain in $\mathbb{R}^n$ with deterministic dynamics which are affine in the control variable. The running cost $L(x,u)$ is quadratic in the control variable u and is fully nonlinear in the state variable x, and there is no exit cost. Any problem in this class can be reduced by a simple change of variables to one with dynamics of calculus of variations type, and we find it convenient in our analysis to consider that form.

In general, one cannot explicitly evaluate either the value function or the optimal control, so accurate numerical approximation methods are needed. The quantity of interest for applications is typically the optimal control, and considerations of robustness in implementation make it important to have the optimal control in feedback form. Furthermore, for many recent applications, including robust control [3] and problems in computer vision [11, 20], approximations to the optimal feedback control and to the closely related gradient of the value function are needed. Given the fact that the control need not be uniquely defined, however, almost all of the literature focuses on approximating the value function, which, under our assumptions, is well defined and Lipschitz on the entire domain.

*Received by the editors September 23, 1998; accepted for publication (in revised form) January 19, 2001; published electronically July 19, 2001. This research was supported in part by the National Science Foundation (NSF-DMS-9704426), the Army Research Office (DAAH04-96-1-0075), and the Office of Naval Research (ONR-N000014-96-1-0276).

http://www.siam.org/journals/sicon/40-2/34496.html

†Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, RI 02912 (dupuis@cfm.brown.edu).

‡Lincoln Laboratory, Massachusetts Institute of Technology, 244 Wood Street, Lexington, MA 02173 (adam@ll.mit.edu).

A natural class of numerical methods, first described by Kushner [18], involves replacing the limit control problem with an approximating problem whose state variable takes values on a finite grid. The deterministic dynamics are replaced by a Markov chain so that movement in an arbitrary direction can be approximated by appropriate probabilities of jumping to neighboring gridpoints. As the underlying grid is refined, the value function for the Markov chain control problem becomes an increasingly good approximation to the value function for the limit problem.

As we noted earlier, the optimal feedback control for the limit problem is not uniquely defined at all points, and this makes it difficult to construct an approximate optimal control on the entire domain. However, there are large subsets of the domain, called regions of strong regularity, on which it is uniquely defined, smooth, and in feedback form. Our main theorem states that the numerical method described in [5] yields a convergent sequence of approximations to the limit optimal feedback control in the regions of strong regularity.

We remark that our proof is applicable to a larger class of control problems than the one considered here. The quadratic structure of the running cost is not essential and can be replaced by suitable smoothness and convexity conditions. We restrict our attention to the quadratic case in order to streamline the presentation. Furthermore, the class of problems that we consider is important in that the infima in the discrete dynamic programming equation (DPE) can be evaluated analytically, eliminating the need for computationally intensive numerical minimizations.

To our knowledge, there are no other general results of this type. Almost all of the literature, both probabilistic [5, 18, 19] and analytic using viscosity solution methods [1, 8], is dedicated to proving convergence of the value functions on the entire domain, and convergence of the controls does not follow naturally from those proofs. That is not surprising, since our proof strongly exploits smoothness properties which hold only in the regions of strong regularity. Some results regarding the convergence of controls in the Markov chain approximation method have been obtained in [12], but the situation there is quite specialized and is restricted to one dimension. In general, one dimensional problems are qualitatively easier to deal with because the control can point only in two directions, while the number of possible directions for $n \geq 2$ is uncountable. In fact, for the present problem, the following startling result can be shown for the case where $n = 1$. For any point x at which the limit problem is regular and for a sufficiently refined grid, the optimal feedback control for the approximating problem is exactly equal to the limit value! This follows from the fact that identical one dimensional quadratic equations for the gradients of the limit and prelimit value functions can be obtained from the DPEs (2.3) and (3.5) by solving for the optimal feedback policies in terms of the respective gradients. Although this observation has limited practical value, it does serve as a powerful reminder of the unique nature of one dimensional problems.

In our development, we draw liberally on ideas presented by Fleming in [14]. There, a similar problem is considered with a small variance Brownian motion perturbation of the deterministic dynamics taking the place of the Markov chain approximations in our problem. In a future paper, we will apply the present result to obtain a full asymptotic expansion of the limit value function in the regions of strong regularity, analogous to the expansion obtained in [14]. Using this expansion, we will present a new numerical method which, under some additional assumptions, will be proved to yield approximations which are second order accurate in the regions of strong regularity.

The outline of this paper is as follows. In section 2, we state our assumptions, introduce the limit optimal control problem, and define the regions of strong regularity on which our results will hold. Section 3 is dedicated to defining the approximating optimal control problems and their associated Markov chain dynamics, while in section 4 we establish some preliminary convergence results. The main theorem is stated and proved in section 5, and we conclude in section 6 with computational examples.

We end this section with some notation. Let $\mathbb{R}^n$ be an n-dimensional Euclidian space, and let $\mathbb{Z}^n$ be the subset of $\mathbb{R}^n$ consisting of n-tuples of integers. For vectors $x, y \in \mathbb{R}^n$, $\langle x, y\rangle$ is the scalar product, $\|x\| = \sqrt{\langle x, x\rangle}$ is the Euclidean norm, $\|x\|_1 = \Sigma_{i=0}^n |x_i|$ is the l^1-vector norm, and $|x| = (|x_1|, \ldots, |x_n|)$ is the componentwise absolute value. For a process $X(\cdot)$ taking values in $\mathbb{R}^n$ and for $S < +\infty$, $\|X(\cdot)\|_S = \int_0^S \|X(t)\| dt$ is the integrated l^2-norm, and $\|\!| X(\cdot) |\!\|_S = \sup_{0\le t\le S} \|X(t)\|$ is the uniform l^2-norm. For any two subsets A and A' of $\mathbb{R}^n$, $d(A, A')$ denotes minimum Euclidean distance between $\bar{A}$ and $\bar{A}'$, while $B_\varepsilon(A)$ is the open ball of radius ε around A.

For a smooth function f mapping $\mathbb{R}^n$ to $\mathbb{R}$, $D_i f(x) = \frac{\partial}{\partial x_i} f(x)$, and the gradient is $Df(x) = (D_1 f(x), \ldots, D_n f(x))$. For $h > 0$, the operators $D^{h,\pm}$ are finite difference approximations to the gradient operator. So the ith component of $D^{h,+} f(x)$ is

$$D_i^{h,+} f(x) = \frac{f(x + he_i) - f(x)}{h},$$

while the ith component of $D^{h,-} f(x)$ is

$$D_i^{h,-} f(x) = \frac{f(x) - f(x - he_i)}{h}.$$

The positive part of a scalar is $a^+ = \max(a, 0)$, and its negative part is $a^- = -\min(a, 0)$. For a vector, the positive and negative parts are taken componentwise, so that $x^\pm = (x_1^\pm, \ldots, x_n^\pm)$.

2. Deterministic control problem. In this section we describe a deterministic optimal control problem on a bounded domain with zero exit cost. Since our goal is to obtain the solution to this problem as the limit of numerical approximations, we will refer to it as the limit problem. Let $G \subset \mathbb{R}^n$ be open with compact closure, and assume that G satisfies uniform interior and exterior cone conditions (see [5] for definitions). Let b and c be C^∞ functions from $\mathbb{R}^n$ to $\mathbb{R}$, and let a be a C^∞ function from $\mathbb{R}^n$ to the space of symmetric positive definite $n \times n$ matrices. Notice that a is uniformly positive definite on G. Assume that $c(x) \ge c_0 > 0$ on G. For a control $\underline{u}^0(t)$ which is in $L^2([0, S]; \mathbb{R}^n)$ for all $S < +\infty$ and for an initial condition $x \in G$, we define $\underline{X}^0(t)$ by the dynamics

$$\underline{X}^0(t) = x + \int_0^t \underline{u}^0(s) ds, \tag{2.1}$$

up to the time when it exits from the domain G. We define the exit time $\underline{\tau}^0 = \inf[t : \underline{X}^0(t) \notin G]$. For the running cost

$$L(x, u) = \frac{1}{2}\Big\langle (u - b(x)), a^{-1}(x)(u - b(x))\Big\rangle + c(x),$$

we define the payoff functional

$$J^0(x, \underline{u}^0) = \int_0^{\underline{\tau}^0} L(\underline{X}^0(t), \underline{u}^0(t)) dt.$$

The problem is to minimize the payoff by choosing a suitable control. Define the value function

$$V^0(x) = \inf_{\underline{u}^0} J^0(x, \underline{u}^0),$$

where the infimum is over controls $\underline{u}^0$ which are in $L^2([0,S];\mathbb{R}^n)$ for all $S < +\infty$. We employ the underscore notation here to indicate trajectories which are obtained from an arbitrary control. The same notations, without the underscores, will be used later to refer to trajectories which are obtained through the application of an optimal control.

We note that our analysis subsumes a much larger class of deterministic control problems. Namely, any problem with smooth dynamics which depend affinely on the control variable u and with a cost structure of the type described above can be made to fit within our framework by a simple change of variables.

The dynamics in (2.1) involve an open loop control $\underline{u}^0(t)$, which is defined for all $t > 0$. It is generally desirable, from the point of view of robustness and for convenience of implementation, to consider controls which can be represented in the feedback form

$$\underline{X}^0(t) = x + \int_0^t \underline{u}^0(\underline{X}^0(s))ds. \tag{2.2}$$

A key feature of the regions of strong regularity is that the optimal open loop controls for all initial conditions in a region of strong regularity correspond to a unique smooth feedback function $u^0(x)$. That is the quantity that we wish to approximate.

The following lemma allows us to regard the limit control problem as one with a finite time horizon and a compact control space, when it is convenient to do so. Thus it follows from [2, Theorem 6.1] that V^0 is the unique nonnegative viscosity solution on G to the DPE

$$\inf_u \left[\langle u, DV^0(x)\rangle + L(x,u)\right] = 0 \tag{2.3}$$

with the continuous boundary condition $V^0(x) = 0$ on ∂G. See [1] and [15] for a thorough account of the relationship between viscosity solutions of Hamilton–Jacobi PDEs and the value functions for various types of optimal control problems.

LEMMA 2.1. *$V^0(x)$ is bounded and uniformly Lipschitz on G, and there exists $T < +\infty$ such that every optimal trajectory exits from G by time $T-1$. Furthermore, there exists $U^0 < +\infty$ such that the norm of every optimal open loop control is bounded by U^0 for each $0 \le t \le T-1$.*

Proof. We begin by observing that the fact that $a(x)$ is uniformly positive definite implies that it is possible to move with unit velocity in any direction with bounded running cost. It follows immediately from this observation that the value function $V^0(x)$ is bounded uniformly by a finite multiple of $\sup_{x\in G} d(x, \partial G)$. Furthermore, the principle of optimality thus implies for all $x, y \in G$ the relation $V^0(x) \le V^0(y) + C\|y - x\|$ for some fixed $C < +\infty$, and this gives a uniform Lipschitz property.

To obtain the bound on the optimal controls, it suffices to find $U^0 < +\infty$ such that any control $\underline{u}^0(t)$ which has norm exceeding U^0 on some measurable set $A \subset [0, +\infty)$ can be replaced by one with a smaller maximum norm, resulting in a lower cost. Since the running cost $L(x,u)$ is uniformly convex in the control variable u, we can

accomplish this for U^0 sufficiently large by constructing a modified control $\tilde{u}^0(t)$ as follows. Let

$$s(t) = \int_0^t \left[I_{A^c}(r) + \frac{1}{2} I_A(r) \right] dr,$$

where $I_{A'}(\cdot)$ is the indicator function for a set A'. Now define

$$\tilde{u}^0(t) = u^0(s(t)) I_{A^c}(s(t)) + \frac{1}{2} u^0(s(t)) I_A(s(t)).$$

This results in following the same trajectory at a slower speed, and a straightforward calculation indicates that it yields a lower cost. □

It turns out that V^0 is smooth on most of the domain G. Let Q be a relatively open subset of $\bar{G}$. We call Q a region of strong regularity if the following hold.

1. For each initial condition $x \in Q$, there is a unique optimal open loop control, and the corresponding trajectory is contained in Q up to its exit time. The optimal trajectory meets ∂G nontangentially.
2. $V^0 \in C^\infty(Q)$.
3. There is a unique $u^0 \in C^\infty(Q)$ such that the optimal control can be represented in feedback form and is given by $u^0(x)$ for each $x \in Q$.

For a discussion of the classical method of characteristics and its application to proving the existence of regions of strong regularity for the present problem, see the appendices in [14] and [16]. At least in the case where ∂G is of class C^∞, the union of the regions of strong regularity is open and dense in the domain [16]. Detailed information on the structure of the singularity sets for closely related problems can be found in [13], [7], and [6]. Since V^0 is a classical solution to the DPE (2.3) on the regions of strong regularity, the optimal feedback control can be explicitly evaluated there:

$$u^0(x) = -a(x) DV^0(x) + b(x). \tag{2.4}$$

Let B_0 be a subset of $\bar{G}$ such that $\bar{B}_0 \subset Q$, and consider a nested sequence of three regions of strong regularity B, N, and Q such that

$$\bar{B}_0 \subset B \subset \bar{B} \subset N \subset \bar{N} \subset Q.$$

The main convergence results will be stated in terms of uniform limits on the set B_0. We assume the following.

ASSUMPTION 2.2. *The boundary section $Q \cap \partial G$ is parallel to one of the coordinate hyperplanes. Furthermore, the minimum distance in the outward normal direction from $Q \cap \partial G$ to $\partial G / Q$ is equal to $\tilde{\delta} > 0$.*

The foregoing assumption is a significant restriction in that it limits our results to regions Q such that $Q \cap \partial G$ is flat and such that ∂G does not make an acute angle at any of the extremal points of $Q \cap \partial G$. We suspect that the assumption about $Q \cap \partial G$ being flat can be relaxed. With regard to acute angles at the corners, we note that Assumption 2.2 does not preclude ∂G itself from making an acute angle, and it does not even preclude showing that the optimal controls converge for points arbitrarily close to such a corner. It does, however, restrict our uniform convergence estimates to regions which are bounded away from such corners.

It is convenient to have $u^0(x)$ defined and Lipschitz on all of $\mathbb{R}^n$, so we abuse notation by extending $u^0(x)$ to $\mathbb{R}^n$ and changing its values on the complement of $\bar{N}$. Let $\delta > 0$ be such that $\delta < d(N, \partial Q \cap G)$ and such that $\delta \leq \tilde{\delta}$, where $\tilde{\delta}$ is as in

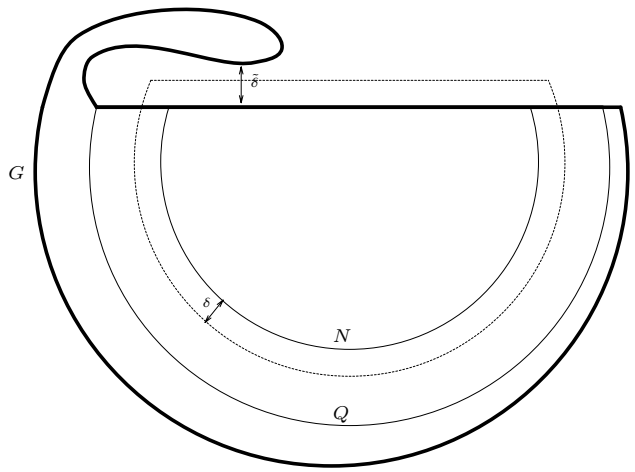


FIG. 1. *Region for smooth extension of* u^0.

Assumption 2.2; see Figure 1. We can define a Lipschitz function $\tilde{u}^0(x)$ on $B_\delta(\bar{N})$ by setting $\tilde{u}^0(x) = u^0(x)$ on $B_\delta(\bar{N}) \cap G$ and by extending it to be constant across the boundary section $Q \cap \partial G$. Now let $\phi(x)$ be a C^∞ function on $\mathbb{R}^n$ taking values in $[0,1]$ such that $\phi(x) = 1$ on $B_{\delta/2}(\bar{N})$ and $\phi(x) = 0$ outside of $B_\delta(\bar{N})$. Such a function can be constructed by standard methods using a smooth convolution kernel [17, Theorem 0.17]. We can now redefine $u^0(x)$ to be equal to $\phi(x)\tilde{u}^0(x)$ on $B_\delta(\bar{N})$ and zero everywhere else. This new $u^0(x)$ is Lipschitz on $\mathbb{R}^n$ and satisfies (2.4) on the region N. Furthermore, $\|u^0(x)\| \le U^0$ for each $x \in \mathbb{R}^n$, where $U^0 < +\infty$ is the bound from Lemma 2.1.

For any $x \in N$, let $X_x^0(t)$ be the trajectory obtained by applying the optimal feedback control u^0 with initial condition x. Since we use the extended version of u^0, we can define $X_x^0(t)$ by (2.2) for all $t \ge 0$. Let τ_x^0 be the first exit time of $X_x^0(t)$ from G, and let z_x^0 be its exit location. Notice that the definition of regions of strong regularity implies that $z_x^0 \in N$ for each $x \in N$ and that τ_x^0 is also the first exit time from the interior of N. We will often suppress the initial conditions in the subscripts of these notations.

LEMMA 2.3. *For each sufficiently small* $\varepsilon > 0$, *there exists* $\eta > 0$ *such that the following holds. Let* X *be a trajectory with initial condition in* N, *and let* τ_N *and* z_N *be its exit time and location from the interior of* N. *If* $\| X - X_x^0 \|_T \le \eta$ *holds for some* $x \in N$, *then* $\tau_N \le \tau_x^0 + \varepsilon$. *If, in addition,* $x \in B$, *then it also follows that* $|\tau_N - \tau_x^0| \le \varepsilon$ *and* $\|z_N - z_x^0\| \le \varepsilon$.

Proof. We first consider the case $x \in N$. Recall from Lemma 2.1 the bounds T and U^0 on the exit times and on the optimal controls, respectively. Given the way we extended u^0 beyond N and given the nontangential exit property for the regions of strong regularity, we have that

$$X_x^0(t) \in B_{\delta/2}(N)/G$$

for all $\tau_x^0 \le t \le \tau_x^0 + \Delta$, where $\Delta = \min(1, \delta/2U^0)$. Furthermore, there is $0 < \gamma \le \delta/2$ such that the component of $\dot{X}_x^0(t)$ in the outward normal direction away from the boundary segment $Q \cap \partial G$ is at least equal to γ for $\tau_x^0 \le t < \tau_x^0 + \Delta$. On account of

the second part of Assumption 2.2, it follows that

$$d(X_x^0(t), G) \geq \gamma(t - \tau_x^0)$$

for $\tau_x^0 \leq t < \tau_x^0 + \Delta$; see Figure 1. Thus, if $\| X - X_x^0 \|_T < \gamma\varepsilon$, then $\tau_N \leq \tau_x^0 + \varepsilon$. For the remainder of this proof, we consider only those η such that $\eta \leq \gamma\varepsilon$, so we may assume that $\tau_N \leq \tau_x^0 + \varepsilon$.

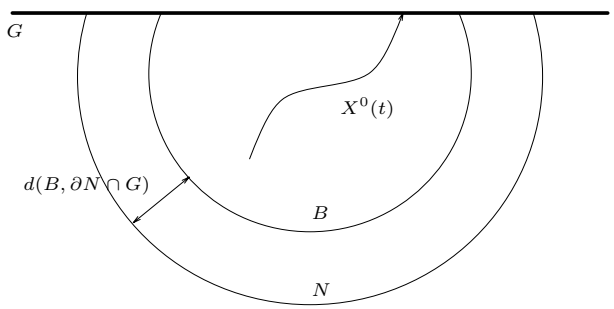


FIG. 2. *Regions of strong regularity.*

Suppose now that $x \in B$. To establish the lower bound for τ_N, we begin by observing that the nontangential exit property for regions of strong regularity implies

$$d(X_x^0(t), \partial N \cap G) \geq d(B, \partial N \cap G) - \varepsilon U^0$$

for all $0 \leq t \leq \tau_x^0 + \varepsilon$; see Figure 2. Thus, if $\varepsilon > 0$ is sufficiently small and if

$$\| X - X_x^0 \|_T < d(B, \partial N \cap G) - \varepsilon U^0,$$

then it follows that $X(\tau_N) \in \partial G$. We observe that $\tau_y^0 \leq C d(y, \partial G)$ holds for any $y \in G$, where $C = \sup_{x \in G} V^0(x)/c_0$ and $c_0 > 0$ is the lower bound on the running cost. Thus, if $\| X - X_x^0 \|_T < \varepsilon/C$, then the previous display implies

$$\tau_x^0 \leq \tau_N + C d(X_x^0(\tau_N), \partial G) \leq \tau_N + \varepsilon.$$

We have shown that $|\tau_N - \tau_x^0| \leq \varepsilon$ is satisfied for sufficiently small $\eta > 0$. Now we observe that

$$\begin{aligned} \|z_N - z_x^0\| &\leq \|X(\tau_N) - X_x^0(\tau_N)\| + \|X_x^0(\tau_N) - X_x^0(\tau_x^0)\| \\ &\leq \eta + U^0 |\tau_N - \tau_x^0|. \end{aligned}$$

Thus we can use the above argument to select a possibly smaller $\eta > 0$ such that $|\tau_N - \tau_x^0| < (\varepsilon - \eta)/U^0$ and so establish the bound $\|z_N - z_x^0\| \leq \varepsilon$. □

The following lemma deals with the continuity of the trajectories with respect to the initial condition. It will be useful in establishing uniformity in the pathwise convergence results of section 4.

LEMMA 2.4. *Let $x_k \in B$ be such that $x_k \to x \in B$, and let T be as in Lemma 2.1. Then*

$$\| X_{x_k}^0 - X_x^0 \|_T, \quad \|u^0(X_{x_k}^0) - u^0(X_x^0)\|_T, \quad |\tau_{x_k}^0 - \tau_x^0|, \quad \textit{and} \quad \|z_{x_k}^0 - z_x^0\|$$

all converge to zero as $k \to \infty$.

Proof. Since the vector field u^0 is globally Lipschitz, the convergence of $\| X^0_{x_k} - X^0_x \|_T$ to zero can be established by a routine application of Gronwall's inequality to the dynamics in (2.2). Given that, the remaining parts of the lemma follow from the uniform continuity of the feedback control u^0 and from Lemma 2.3. □

3. Markov chain approximations. We employ the method of approximating Markov chains developed by Kushner [18] to compute approximate solutions to the deterministic optimal control problem described above. For an up-to-date treatment of this subject, see the book of Kushner and Dupuis [19]. Our approximation is essentially the one used by Boué and Dupuis in [5]. In order to numerically approximate the value function V^0 and the optimal control u^0, we need to define a process which takes values on a finite lattice and which approximates the continuous dynamics. We circumvent the problem of only being able to move in the lattice directions by introducing jump probabilities which give rise to arbitrary mean velocities. The value function corresponding to this process, with the same cost structure as above, satisfies on the lattice a DPE analogous to (2.3). Thus it is possible to numerically compute the value function and the optimal feedback control for the approximating problem. We will show that, at least in the compact set B_0, these are good approximations to V^0 and u^0.

Let $h > 0$ be a discretization parameter, and define the discrete domain $G^h = h\mathbb{Z}^n \cap G$. For any $A \subset \mathbb{R}^n$, we define $A^h = h\mathbb{Z}^n \cap A^o$, where A^o is the interior of A. We consider limits as $h \to 0$ with the h chosen such that the hyperplane in which the boundary section $Q \cap \partial G$ lies lines up with the lattice $\mathbb{Z}^h$ (see Assumption 2.2). We will construct a continuous time controlled jump Markov process on G^h which approximates the deterministic dynamics in (2.2). This process will give rise to the same DPE obtained in [5] by using a discrete time Markov chain. For our purposes, however, it is more convenient to work with a continuous time jump Markov process.

Let $\underline{u}^h$ be any feedback control on G^h, and extend $\underline{u}^h$ to be equal to u^0 on $\mathbb{Z}^h/G^h$. Let $\underline{X}^h$ be the Markov process with controlled generator given by

$$\mathcal{L}^h_u f = \langle u^+, D^{h,+} f \rangle - \langle u^-, D^{h,-} f \rangle \tag{3.1}$$

for any smooth function f mapping $\mathbb{R}^n$ to $\mathbb{R}$. See section 1 for the notation in this definition. The stochastic dynamics corresponding to this generator will be called the h-dynamics. As in the description of the limit problem, we employ the underscore notation to indicate objects which are obtained from the application of an arbitrary possibly suboptimal feedback control.

Since we consider only feedback controls, it is straightforward to construct $\underline{X}^h$, as in section 4.3 of [19] and in [9]. We define a sequence of independent and identically distributed exponential random fields parameterized by u with mean values specified as follows:

$$\overline{\Delta t}^h(u) = \begin{cases} \frac{h}{\|u\|_1}, & u \neq 0, \\ h, & u = 0. \end{cases}$$

Suppose that after $m-1$ jumps, $\underline{X}^h(s)$ is defined for $0 \le s \le t$ and that $\underline{X}^h(t) = x$. Then we take $\underline{X}^h(s) = x$ for all $t \le s < t + \eta$, where the waiting time η is the exponential random variable obtained by evaluating the mth random field with the parameter value $u = \underline{u}^h(\underline{X}^h(s))$. If $u = 0$, then $\underline{X}^h(t + \eta) = x$, but otherwise it is

conditionally distributed according to the jump probabilities

$$p^h(x,y|u) = \begin{cases} \frac{u_i^{\pm}}{\|u\|_1} & \text{if } y = x \pm he_i, \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to verify that the mean velocity of $\underline{X}^h$ at time t conditioned on $\underline{X}^h(t) = x$ is equal to $\underline{u}^h(x)$, so this is a consistent approximation to the limit dynamics in (2.2).

We consider the semimartingale decomposition of $\underline{X}^h$. For a given feedback control $\underline{u}^h$ and fixed initial condition $x \in G^h$, we write

$$\underline{X}^h(t) = \underline{Y}^h(t) + \underline{m}^h(t), \tag{3.2}$$

where the stochastic process $\underline{Y}^h$ is defined with probability one (w.p.1) by

$$\underline{Y}^h(t) = x + \int_0^t \underline{u}^h(\underline{X}^h(s))ds.$$

The consistency of the jump dynamics guarantees that $\underline{m}^h(t)$ is a martingale with mean zero. Furthermore, the variance of $\underline{m}^h(t)$ is controlled by the parameter h. That is the content of the following lemma.

LEMMA 3.1. *Fix $h > 0$, and let $\underline{u}^h$ be any feedback control which satisfies $\|\underline{u}^h(x)\| \le K < +\infty$ for all $x \in h\mathbb{Z}^n$. Then the bound*

$$E_x\|\underline{m}^h(\sigma)\|^2 \le hKE_x\sigma$$

holds for any bounded stopping time σ.

Proof. The triple $(\underline{m}^h, \underline{X}^h, \underline{Y}^h)$ is Markov and measurable with respect to the σ-algebra generated by $\underline{X}^h$. We consider its generator $\tilde{\mathcal{L}}^h$. Since $\underline{m}^h = \underline{X}^h - \underline{Y}^h$, we have

$$\tilde{\mathcal{L}}^h f = \langle \underline{u}^{h,+}(\underline{X}^h), D^{h,+}f\rangle - \langle \underline{u}^{h,-}(\underline{X}^h), D^{h,-}f\rangle - \langle \underline{u}^h(\underline{X}^h), Df\rangle$$

for any smooth function of the form $f(m,x,y) = f(m)$. Given the fact that $\underline{m}^h(t)$ takes values in a bounded set for bounded values of $t \ge 0$, the general theory of piecewise deterministic processes [9, Theorem 5.5] implies that for any smooth function f on $\mathbb{R}^n$, for any initial condition $x \in G$, and for any bounded stopping time σ,

$$E_x\left[f(\underline{m}^h(\sigma)) - f(0) - \int_0^\sigma \tilde{\mathcal{L}}^h f(\underline{m}^h(t))dt\right] = 0. \tag{3.3}$$

Taking $f(m) = \|m\|^2$, we use (3.3) and the fact that for this choice of f,

$$|\tilde{\mathcal{L}}^h f(\underline{m}^h(s))| \le 2h\|\underline{u}^h(\underline{X}^h(s))\|_1$$

to obtain

$$E_x\|\underline{m}^h(\sigma)\|^2 \le 2hE_x\int_0^\sigma \|\underline{u}^h(\underline{X}^h(s))\|_1 ds.$$

Given the bound on $\|\underline{u}^h\|$, this completes the proof. □

We now formulate the discrete approximation to the optimal control problem discussed in section 2. Define the value function

$$V^h(x) = \inf_{\underline{u}^h} E_x\int_0^{\underline{\tau}^h} L(\underline{X}^h(t), \underline{u}^h(\underline{X}^h(t)))dt, \tag{3.4}$$

where the exit time is $\underline{\tau}^h = \inf[t : \underline{X}^h(t) \notin G^h]$, and the infimum is over feedback controls $\underline{u}^h$. Using standard methods [19, section 4.3] it can be shown that V^h is the unique solution on G^h to the DPE

$$\inf_u \left[\left\langle u^+, D^{h,+}V^h(x) \right\rangle - \left\langle u^-, D^{h,-}V^h(x) \right\rangle + L(x,u) \right] = 0 \tag{3.5}$$

with zero boundary condition on $h\mathbb{Z}^n/G^h$. It is straightforward to verify that (3.5) is equivalent to

$$V^h(x) = \inf_u \left[\sum_{y \in \mathbb{R}^n} p^h(x,y|u)V^h(y) + \overline{\Delta t}^h(u)L(x,u) \right] \tag{3.6}$$

for $x \in G^h$, and that the minimizing values of u are the same for these two equations. As suggested by the form of (3.6), the fixed point and an optimal feedback control can be found numerically using either Jacobi or Gauss–Seidel iteration schemes. We note that (3.6) is the DPE for a different approximating control problem, where a discrete time Markov chain is used to approximate the deterministic dynamics. That is the approach taken in [5], where the time step $\overline{\Delta t}^h(u)$ is used to interpolate the Markov chain into continuous time. As discussed in [5], the choice of one-sided transition probabilities and of a control dependent time step facilitates rapid convergence of the iterative schemes used to solve (3.6), and the required infima at each step can be evaluated analytically.

The DPE (3.5) gives rise to an optimal feedback control u^h on G^h. It is convenient at this point to abuse notation and to redefine u^h to be equal to u^0 on $h\mathbb{Z}^n/N^h$. Then, for each initial condition $x \in N^h$, there is a unique process X_x^h defined for all $t \geq 0$ which is optimally controlled by u^h until it exits from N^h. We define the exit time $\tau_{x,N}^h = \inf[t : X_x^h(t) \notin N^h]$ and the exit location $z_{x,N}^h = X_x^h(\tau_{x,N}^h)$. Recall that N^h is defined to be $h\mathbb{Z}^n \cap N^o$, where N^o is the interior of N. Let m_x^h be the martingale part of the decomposition for X_x^h given by (3.2). As in the limit problem, we will often suppress the initial conditions in the subscripts of all of these notations.

The following remark and lemma simplify some of the analysis by allowing us to consider a compact domain and a compact control space.

REMARK 3.2. *We extended $u^0(x)$ to all of $\mathbb{R}^n$ in such a way that it is equal to zero off of the neighborhood $B_\delta(N)$. Thus the same is now true of $u^h(x)$. Consequently, for all initial conditions $x \in N$, the trajectories X_x^0 and X_x^h never leave the closed neighborhood $\overline{B_{\delta+h}(N)}$.*

LEMMA 3.3. *There exists a compact set $U \subset \mathbb{R}^n$ such that the extended optimal feedback controls $u^0(x)$ and $u^h(x)$ take values in U for all $h > 0$ and for all $x \in \mathbb{R}^n$ on which they are defined. Furthermore, the value functions $V^0(x)$ and $V^h(x)$ are bounded, uniformly in h and x.*

Proof. Recall from Lemma 2.1 that we obtained $U^0 < +\infty$ such that $\|u^0(x)\| \leq U^0$ for all $x \in \mathbb{R}^n$. From the DPE (3.5), it follows that for $x \in G^h$, each component of $u^h(x)$ either is equal to zero or is given by a bounded linear functional of $D^{h,\pm}V^h(x)$. Thus, in order to find the set U, it suffices to establish a bound on $D^{h,\pm}V^h(x)$ which is uniform for all $h > 0$ and $x \in G^h$.

Without loss of generality, we consider the case of bounding $D_i^{h,+}V^h(x)$. The principle of optimality implies that the minimal cost starting at $x \in G^h$ can be no larger than the minimal cost starting at $x + he_i$ plus the expected cost of getting from x to $x + he_i$ under any suboptimal control. The fact that a is uniformly positive

definite while b and c are bounded implies that there exists a constant $K < \infty$ such that $L(x,u) \le K$ whenever $\|u\| = 1$. Taking $u = he_i$, we obtain $V^h(x) \le V^h(x+he_i) + KE_x\eta$, where the waiting time η is exponential with mean h. Thus we have shown $V^h(x) \le V^h(x+he_i)+hK$. The reverse inequality is established by using $x+he_i$ as the initial condition, and it follows that $|D_i^{h,+}V^h(x)| < K$. That concludes the proof for the optimal controls. A uniform bound on the value functions $V^h(x)$ follows from the above argument and from the boundedness of the domain G. Along with the bound on $V^0(x)$ from Lemma 2.1, this finishes the proof. □

4. Preliminary convergence results. The main results of this section are contained in Lemma 4.1. It states that in the region of strong regularity B, the optimal trajectories, open loop controls, exit times, and exit locations converge in probability, uniformly with respect to initial conditions. Since the limit objects are deterministic, we are able to use convergence in distribution arguments to establish the desired convergence in probability. Uniformity with respect to initial conditions is a consequence of the continuity properties in Lemma 2.4.

LEMMA 4.1. *Let T be the bound on the exit times from Lemma* 2.1. *For every $\varepsilon > 0$, there exists $h_0 > 0$ such that*

$$\text{(i)} \quad P_x\left[\|\!|\!|\, X_x^h - X_x^0 \,|\!|\!|_T > \varepsilon\right] < \varepsilon,$$

$$\text{(ii)} \quad P_x\left[\|u^h(X_x^h) - u^0(X_x^0)\|_T > \varepsilon\right] < \varepsilon,$$

$$\text{(iii)} \quad P_x\left[\tau_{x,N}^h > \tau_x + \varepsilon\right] < \varepsilon$$

holds for all $0 < h \le h_0$ and for all initial conditions $x \in N^h$, and such that

$$\text{(iv)} \quad P_x\left[|\tau_{x,N}^h - \tau_x^0| > \varepsilon\right] < \varepsilon,$$

$$\text{(v)} \quad P_x\left[\|z_{x,N}^h - z_x^0\| > \varepsilon\right] < \varepsilon$$

holds for all $0 < h \le h_0$ and for all initial conditions $x \in B^h$.

We will use the following convergence result [5]. In fact, we will repeat part of the argument to prove this theorem in our proof of Lemma 4.1, but the exposition is made more transparent by assuming convergence of the value functions.

THEOREM 4.2. (i) *Let $x^h \in G^h$ for $h > 0$ be such that $x^h \to x \in \bar{G}$ as $h \to 0$. Then $V^h(x^h) \to V^0(x)$ as $h \to 0$.* (ii) *For any $\varepsilon > 0$, there exists $h_0 > 0$ such that $|V^h(x) - V^0(x)| < \varepsilon$ for all $0 < h \le h_0$ and all $x \in G^h$.*

Proof. Part (i) is proved, in a somewhat more general setting, as Theorem 5.4 in [5]. If part (ii) is false, then there are $\varepsilon > 0$ and a sequence $x^h \in G^h$ with $h \to 0$ such that $|V^h(x^h) - V^0(x^h)| > \varepsilon$ for each h. Since $\bar{G}$ is compact and V^0 is uniformly continuous on $\bar{G}$, we can extract a subsequence such that $x^h \to x \in \bar{G}$ and $|V^h(x^h) - V^0(x)| > \varepsilon/2$ for each h, which contradicts part (i) of the theorem. □

To facilitate treating the optimal trajectories and controls in the framework of convergence in distribution, we adopt some standard definitions. We treat the processes X^h as random variables taking values in $\mathcal{D}([0,\infty);\mathbb{R}^n)$, the space of $\mathbb{R}^n$-valued functions that are continuous from the right and have limits on the left. With the Skorokhod metric, $\mathcal{D}([0,\infty);\mathbb{R}^n)$ is a complete separable metric space [4], and convergence of a sequence in $\mathcal{D}([0,\infty);\mathbb{R}^n)$ is equivalent to convergence of that sequence

in $\mathcal{D}([0,S];\mathbb{R}^n)$ for each $S<+\infty$. If a sequence in $\mathcal{D}([0,\infty);\mathbb{R}^n)$ converges to a continuous function under the Skorokhod metric, then it also converges in the uniform norm $|||\cdot|||_S$ for each $S<+\infty$.

We also consider the space of relaxed controls. A relaxed control is an element of $\mathcal{R}(U\times[0,\infty))$, the space of all Borel measures ν on $U\times[0,\infty)$ such that $\nu(\mathbb{R}^n\times[0,S])=S$ for each $S\le+\infty$, where $U\subset\mathbb{R}^n$ is the compact control set from Lemma 3.3. This space can be metrized as a complete separable metric space with a metric such that $\nu_k\to\nu$ if and only if the restriction of ν_k to $U\times[0,S]$ converges weakly to the restriction of ν to $U\times[0,S]$ for all $S\le+\infty$ [19, section 9.5]. The second marginal of any measure $\nu\in\mathcal{R}(U\times[0,\infty))$ is a Lebesgue measure, so the decomposition $\nu(du\times dt)=\nu_t(du)dt$ holds, where ν_t is a probability measure for each $t\ge0$. If ν is a random variable, then this decomposition can be done so that it holds almost surely and so that for all $t\ge0$, ν_t is a random variable. We note that the following version of Fatou's lemma holds [4, Theorems 5.1 and 5.3]. If $\nu_k\to\nu$, then

$$\liminf_{k\to\infty}\int_{U\times[0,S]}f\,d\nu_k\ge\int_{U\times[0,S]}f\,d\nu \tag{4.1}$$

for any $S<+\infty$ and for any continuous nonnegative function f on the space $U\times[0,\infty)$.

For $h>0$ and for each $t\ge0$, let $\nu_t^h=\delta_{u^h(X^h(t))}$, where δ_u is the probability measure on U that places unit mass at the point u, and u^h is the optimal feedback control for the prelimit problem with parameter h. The corresponding optimal relaxed control is the measure valued random variable given by $\nu^h(A\times A')=\int_{A'}\nu_t^h(A)dt$ for Borel sets $A\subset U$ and $A'\subset[0,\infty)$. In terms of the optimal relaxed control measures, the inequality

$$V^h(x)\ge E_x\int_{U\times[0,\tau_N^h]}L(X^h(t),u)\nu^h(du\times dt) \tag{4.2}$$

holds for each $x\in N^h$. Equality may not hold in (4.2) because it is possible to have $X^h(\tau_N^h)\notin\partial G$; for equality to hold, we would need to add $E_xV^h(X^h(\tau_N^h))\ge0$ to the right-hand side of expression (4.2). For initial conditions in the region of strong regularity N, we can similarly define ν^0 to be the measure in $\mathcal{R}(U\times[0,\infty))$ with first marginals $\nu_t^0=\delta_{u^0(X^0(t))}$. Since ν^0 is not a random variable, an analogue to (4.2) holds for V^0 without the expectation:

$$V^0(x)=\int_{U\times[0,\tau^0]}L(X^0(t),u)\nu^0(du\times dt). \tag{4.3}$$

Notice that the inequality in (4.2) is replaced by equality in (4.3) because τ^0 is the exit time of X^0 from G. The proof of the following lemma is nearly identical to the proof of Lemma 5.3 in [5]. The only necessary modification is to use the martingale estimate from Lemma 3.1 in place of an analogous estimate obtained by applying a standard conditioning argument to the discrete time processes in [5].

LEMMA 4.3. *For $h>0$, let $x^h\in N^h$ be such that $x^h\to x\in\bar N$ as $h\to0$. Then, using these initial conditions, the random variables (X^h,ν^h) are tight. Furthermore, for any subsequence along which the limit*

$$(X^h,\nu^h)\to(X,\nu)$$

holds in the sense of distributions,

$$X(\cdot) = x + \int_0^{\cdot} \int_U u\, \nu_s(du)ds \tag{4.4}$$

*is valid w.p.*1.

Given the tightness from Lemma 4.3 and the convergence of the value functions from Theorem 4.2, we can use the uniqueness of optimal trajectories in regions of strong regularity for the limit problem to prove that the optimal trajectories and controls converge in distribution to the appropriate limit quantities. That is the conclusion of the next lemma.

LEMMA 4.4. *Let $x^h \in N^h$ be such that $x^h \to x \in \bar{N}$ as $h \to 0$. Then, using these initial conditions, the limit $(X^h, \nu^h) \to (X^0, \nu^0)$ holds in the sense of distributions as $h \to 0$.*

Proof. We consider the τ_N^h as random variables taking values in the compactified space $[0, \infty]$. Then Lemma 4.3 implies that the random variables (X^h, ν^h, τ_N^h) are tight. Thus, given the continuity of the process in expression (4.4), for any subsequence there is a further subsequence along which the weak convergence $(X^h, \nu^h, \tau_N^h) \to (X, \nu, \tilde{\tau})$ holds for some limit random variable taking values in

$$\mathcal{C}([0,\infty); \mathbb{R}^n) \times \mathcal{R}(U \times [0,\infty)) \times [0,\infty].$$

We will show that for any such limit, (X, ν) is w.p.1 equal to (X^0, ν^0).

By the Skorokhod representation theorem [10], we can consider a probability space on which the convergence is w.p.1. Since the limit trajectory X is continuous, the convergence $X^h \to X$ is uniform on compact intervals. Thus it is easy to verify that w.p.1 $\tilde{\tau} \geq \tau_N$, where τ_N is the first exit time of $X(t)$ from the interior of N. We obtain the following series of inequalities, each line of which is explained after the display:

$$\begin{aligned}
V^0(x) &= \lim_{h\to 0} V^h(x^h) \\
&\geq \lim_{h\to 0} E_{x^h} \int_{U\times[0,\tau_N^h)} L(X^h(t), u)\nu^h(du \times dt) \\
&\geq E_x \int_{U\times[0,\tilde{\tau})} L(X(t), u)\nu(du \times dt) \\
&\geq E_x \int_{U\times[0,\tau_N)} L(X(t), u)\nu(du \times dt) \\
&\geq E_x \int_0^{\tau_N} L(X(t), \dot{X}(t))dt \\
&\geq V^0(x).
\end{aligned}$$

The first line is due to part (i) of Theorem 4.2; the second line comes from the representation in (4.2); the third line is obtained by applying (4.1) along with the standard version of Fatou's lemma; the fourth line uses the fact that w.p.1 $\tilde{\tau} \geq \tau_N$; the fifth line follows from Jensen's inequality and the relation (4.4); and the final line is a consequence of the definition of $V^0(x)$.

Evidently, all of the inequalities in the previous display are in fact equalities. Thus, given the uniqueness of optimal trajectories in the regions of strong regularity, the last line implies that w.p.1, $X(t) = X^0(t)$ for $0 \leq t \leq \tau_N = \tau^0$. Recall that for

equality to occur in Jensen's inequality with a strictly convex function, the probability measure must be a point mass. Thus equality in the fifth line implies that w.p.1 $\nu_t = \nu_t^0$ for almost every (a.e.) $0 \le t \le \tau^0$. It remains to show that w.p.1, $X(t) = X^0(t)$ and $\nu_t = \nu_t^0$ for a.e. $\tau^0 \le t \le T$.

Since $|||\ X^h - X^0\ |||_{\tau^0}$ converges to zero w.p.1 it follows that $\|X^h(\tau^0) - X^0(\tau^0)\|$ converges to zero w.p.1 and hence in probability. Thus we can use the optimality of X^h, along with a uniform Lipschitz type bound on the V^h (see Lemma 3.3) and the lower bound on the running cost, to conclude that $\tau_N^h - \tau^0$ is small with arbitrarily high probability. Since the optimal controls are bounded, it follows that with high probability $\|X^h(t) - X^0(t)\|$ is arbitrarily small up to time $\tau^0 \vee \tau_N^h$. Furthermore, since $u^h = u^0$ outside of N, Gronwall's inequality implies that if $\|X^h(\tau^0 \vee \tau_N^h) - X^0(\tau^0 \vee \tau_N^h)\|$ is small, if $X^h(t)$ stays uniformly close to its mean after the stopping time $\tau^0 \vee \tau_N^h$, and if X^h does not return to N after it exits, then $\|X^h(t) - X^0(t)\|$ is small for all $\tau^0 \vee \tau_N^h \le t \le T$. This event occurs with arbitrarily high probability, so we conclude that $X(t) = X^0(t)$ w.p.1 for all $\tau^0 \le t \le T$. The verification of the needed fact that with high probability X^h does not return to N after it exits uses the observation that the extended optimal controls u^0 and u^h point away from the region N near the boundary section $N \cap \partial G$ and that the h-dynamics are one sided, so that whenever X^h exits from N at a point in $N \cap \partial G$ (which happens with high probability), it reaches the region where $u^h = 0$ before returning to N; see Figure 1. Finally, $\nu_t = \nu_t^0$ for a.e. $\tau^0 \le t \le T$ follows from the above argument since $u^h = u^0$ outside of N and u^0 is uniformly Lipschitz on $\mathbb{R}^n$. □

Proof of Lemma 4.1. Suppose that part (i) of the lemma is false. Then there exists $\varepsilon > 0$ along with a sequence $x^h \in B^h$ with $h \to 0$ such that

$$P\left[|||\ X_{x^h}^h - X_{x^h}^0\ |||_T > \varepsilon\right] \ge \varepsilon$$

for each h. Using the continuity of X^0 as a function of its initial condition from Lemma 2.4, we can extract a subsequence such that $x^h \to x \in \bar{B}$ and

$$P\left[|||\ X_{x^h}^h - X_x^0\ |||_T > \varepsilon/2\right] \ge \varepsilon$$

for each h. This is a contradiction, since the convergence in distribution of $X_{x^h}^h$ to the deterministic limit X_x^0 in Lemma 4.4 implies that $|||\ X_{x^h}^h - X_x^0\ |||_T \to 0$ in probability. Parts (iii)–(v) follow from part (i) and from Lemma 2.3.

The proof of part (ii) is slightly more subtle because we need to parlay the convergence of relaxed control measures from Lemma 4.4 into a statement about the convergence in $L^2([0,T];\mathbb{R}^n)$ of the controls $u^h(X^h(t))$. Consider a sequence of initial conditions $x^h \in B^h$ such that $x^h \to x \in \bar{B}$ as $h \to 0$. Using Lemma 4.4 and the Skorokhod representation theorem, we consider a probability space on which $\nu^h \to \nu^0$ w.p.1. Since $\nu^0(du \times dt) = \delta_{u^0(X^0(t))}(du)dt$ and $\nu^h(du \times dt) = \delta_{u^h(X^h(t))}(du)dt$, the w.p.1 convergence $\nu^h \to \nu^0$ implies

$$\begin{aligned}\int_0^T \|u^h(X^h(t)) - u^0(X^0(t))\|^2 dt &= \int_{U\times[0,T]} \|u - u^0(X^0(t))\|^2 \nu^h(du \times dt) \\ &\longrightarrow \int_{U\times[0,T]} \|u - u^0(X^0(t))\|^2 \nu^0(du \times dt) \\ &= \int_0^T \|u^0(X^0(t)) - u^0(X^0(t))\|^2 dt \\ &= 0,\end{aligned}$$

where the limit in the second line holds as $h \to 0$ w.p.1. Thus, switching from the Skorokhod space back to the original random variables, we can conclude that $\|u^h(X^h) - u^0(X^0)\|_T$ converges to zero in probability for any sequence of initial conditions $x^h \in B^h$ such that $x^h \to x \in \bar{B}$ as $h \to 0$. Now, as in the proof of part (i), this implies the convergence asserted by the lemma. □

It is useful to identify the suboptimal processes obtained by applying the limit optimal feedback control u^0 in the h-dynamics. For an initial condition in N^h, let $X^{h,0}$ be the process obtained by taking $\underline{u}^h = u^0$ in section 3, and let $m^{h,0}$ and $Y^{h,0}$ be the corresponding martingale and bounded variation parts indicated by the decomposition (3.2). Finally, define the exit time $\tau_N^{h,0} = \inf[t : X^{h,0}(t) \notin N^h]$ and the exit location $z_N^{h,0} = X^{h,0}(\tau_N^{h,0})$.

LEMMA 4.5. *Let T be the bound on the exit times from Lemma* 2.1. *For every $\varepsilon > 0$, there exists $h_0 > 0$ such that*

$$\text{(i)} \qquad P_x\left[\|\!| X_x^{h,0} - X_x^0 |\!\|_T > \varepsilon\right] < \varepsilon,$$

$$\text{(ii)} \qquad P_x\left[\tau_{x,N}^{h,0} > \tau_x + \varepsilon\right] < \varepsilon$$

holds for all $0 < h \le h_0$ and for all initial conditions $x \in N^h$, and such that

$$\text{(iii)} \qquad P_x\left[|\tau_{x,N}^{h,0} - \tau_x^0| > \varepsilon\right] < \varepsilon,$$

$$\text{(iv)} \qquad P_x\left[\|z_{x,N}^{h,0} - z_x^0\| > \varepsilon\right] < \varepsilon$$

holds for all $0 < h \le h_0$ and for all initial conditions $x \in B^h$.

Proof. For an initial condition $x \in B^h$, let $Z^h(t) = X^{h,0}(t) - X^0(t)$. Then by (3.2) we have

$$Z^h(t) = \int_0^t \left[u^0(X^{h,0}(s)) - u^0(X^0(s))\right] ds + m^{h,0}(t)$$

holding w.p.1 for any $t < +\infty$. Thus, if K is the uniform Lipschitz constant for u^0, then for any $0 \le \sigma < +\infty$

$$\|Z^h(t)\| \le \int_0^t K\|Z^h(s)\| ds + \|\!| m^{h,0}(t) |\!\|_\sigma$$

holds w.p.1 for each $0 \le t \le \sigma$. We can apply a version of Gronwall's inequality [10, Theorem A.6.4] to get the w.p.1 bound

$$\| X^{h,0}(t) - X^0(t) \|_\sigma \le \| m^{h,0}(t) \|_\sigma e^{K\sigma}. \tag{4.5}$$

Now, letting $\sigma = T$ in (4.5) and applying Lemma 3.1 with a standard submartingale inequality, we obtain part (i) of the lemma. Parts (ii)–(iv) follow directly from part (i) and from Lemma 2.3. □

5. Convergence of the feedback controls. The main results of this paper are Theorem 5.5 and Corollary 5.6. They state that in the set B_0, the optimal feedback controls $u^h(x)$ for the approximating control problems converge uniformly to $u^0(x)$, the optimal feedback control for the limit problem. Once we establish the analogous convergence of the approximate gradients $D^{h,\pm}V^h(x)$ to $DV^0(x)$, we will be able to use the uniqueness of the optimal control $u^0(x)$ to prove Theorem 5.5. Thus most of this section is devoted to establishing the following lemma.

LEMMA 5.1. *Let $x^h \in B_0^h$ be such that $x^h \to x \in \bar{B}_0$ as $h \to 0$. Then, as $h \to 0$,*

$$D^{h,\pm}V^h(x^h) \to DV^0(x).$$

There are two main steps in the proof of Lemma 5.1. First, we obtain the convergence of $D^{h,\pm}V^h(x)$ to $DV^0(x)$ in a neighborhood of $B \cap \partial G$. Then we use representations of $V^0(x)$ and $V^h(x)$ in terms of integrals along optimal trajectories to obtain the convergence of the $D^{h,\pm}V^h(x)$ to $DV^0(x)$ on the interior of the smaller region B_0. Our arguments are very similar in spirit to those used in the proof of [14, Lemma 5.5]. It is useful in what follows to define a compact notation for the running cost under a feedback control $u(x)$ by

$$L_u(x) = L(x, u(x)). \tag{5.1}$$

The following lemma establishes a geometric bound on the difference between V^h and V^0 on the set B.

LEMMA 5.2. *For any $m > 0$, there exists $h_0 > 0$ such that*

$$(1-m)V^0(x) \;\le\; V^h(x) \;\le\; (1+m)V^0(x)$$

holds for all $0 < h \le h_0$ and for all $x \in B^h$.

Proof. We prove this lemma in two steps, first considering the upper bound on $V^h(x)$ and then the lower bound.

Upper bound. Let $\mu > m$, and put $W^0 = (1+\mu)V^0$. It follows from the DPE (2.3) that

$$\langle u^{0,+}, D^{h,+}W^0 \rangle - \langle u^{0,-}, D^{h,-}W^0 \rangle + \tilde{L}_{u^0} = 0 \tag{5.2}$$

holds on N, where the modified cost $\tilde{L}_{u^0}(x)$ is defined on N by

$$\tilde{L}_{u^0} = (1+\mu)L_{u^0} + \langle u^0, DW^0 \rangle - \langle u^{0,+}, D^{h,+}W^0 \rangle + \langle u^{0,-}, D^{h,-}W^0 \rangle.$$

Note that since $L_{u^0}(x) \ge c_0 > 0$ and $W^0(x)$ is smooth, $\tilde{L}_{u^0}(x) \ge L_{u^0}(x)$ for $h > 0$ sufficiently small and for all $x \in N$. Since the generator in (5.2) corresponds to applying the feedback control u^0 in the h-dynamics, we can use a standard verification argument to establish for all $x \in B^h$ the representation

$$W^0(x) = E_x \left[\int_0^{\tau_N^{h,0}} \tilde{L}_{u^0}(X^{h,0}) dt + W^0(z_N^{h,0}) \right]. \tag{5.3}$$

We use part (ii) of Lemma 4.5 to obtain the uniform integrability of $\tau_N^{h,0}$ needed for the right-hand side of (5.3) to be finite.

For $x \in B^h$, we define

$$V^{h,0}(x) = E_x \left[\int_0^{\tau_N^{h,0}} L_{u^0}(X^{h,0}) dt + V^h(z_N^{h,0}) \right]. \tag{5.4}$$

Since the feedback control u^0 is suboptimal in the control problem with the h-dynamics, it follows from the strong Markov property that $V^h(x) \leq V^{h,0}(x)$ for all $x \in B^h$. Thus it suffices to establish the bound $V^{h,0}(x) \leq (1+m)V^0(x)$.

Let K be the bound on $V^h(x)$ from Lemma 3.3. Then the following series of inequalities holds for all sufficiently small $h > 0$ and for all $x \in B^h$:

$$\begin{aligned} V^{h,0}(x) &\leq W^0(x) + E_x \left[V^h(z_N^{h,0}) - W^0(z_N^{h,0}) \right] \\ &\leq (1+\mu)V^0(x) + E_x V^h(z_N^{h,0}) \\ &\leq (1+\mu)V^0(x) + K P_x \left[z_N^{h,0} \in G \right]. \end{aligned} \tag{5.5}$$

The first line uses the representations (5.3) and (5.4), along with the fact that $\tilde{L}_{u^0}(x) \geq L_{u^0}(x)$ for all $x \in N$; the second line uses the definition of $W^0(x)$ and the nonnegativity of $W^0(x)$ for all $x \in G$; and the third line uses the fact that $V^h(x) = 0$ for all $x \in \partial G$.

We now turn our attention to bounding the final term in the last display. Let $\varepsilon > 0$ be equal to $d(B, \partial N \cap G)$, so that $z_N^{h,0} \in G$ implies that $z_N^{h,0}$ is at least a distance of ε away from $z^0 \in B$; see Figure 3 before Lemma 5.3. Choose $0 < \delta < \varepsilon/2$ such that once an optimal trajectory for the limit problem with initial condition $x \in B$ gets to within δ of the boundary ∂G, it can travel no further than distance $\varepsilon/2$ before exiting. The existence of such a $\delta > 0$ is guaranteed by the nontangential exit property for the regions of strong regularity. Finally, let $0 < \eta < \delta < \varepsilon/2$ be chosen so that the conclusions of Lemma 2.3 hold. We obtain the following series of inequalities holding for all $x \in B^h$, each line of which is explained after the display:

$$\begin{aligned} P_x \left[z_N^{h,0} \in G \right] &\leq P_x \left[\| z_N^{h,0} - z^0 \| > \varepsilon \right] \\ &\leq P_x \left[\| \! | X^{h,0} - X^0 | \! \|_{\tau_N^{h,0} \wedge T} \geq \eta \right] \\ &\leq P_x \left[\| \! | m^{h,0}(t) | \! \|_{\tau_N^{h,0} \wedge T} \geq \eta e^{-K'T} \right] \\ &\leq (\eta e^{-K'T})^{-2} E_x \left[\| m^{h,0}(\tau_N^{h,0} \wedge T) \|^2 \right] \\ &\leq hCE_x \tau_N^{h,0} \\ &\leq hc_0^{-1} C V^{h,0}(x). \end{aligned} \tag{5.6}$$

The first line is a consequence of the fact that $z_N^{h,0} \in G$ can occur only if $\| z_N^{h,0} - z^0 \| > \varepsilon$; the second line follows from the choice of δ and η; the third line follows from (4.5)

with K' equal to the Lipschitz constant for $u^0(x)$; the fourth line is obtained by a standard submartingale inequality; the fifth line follows from Lemma 3.1 with C a composite finite constant; and the last line is a consequence of the definition of $V^{h,0}$ and of the lower bound c_0 on the running cost. We can combine the last lines of (5.5) and (5.6) to obtain the bound

$$(1 - hc_0^{-1}CK)V^{h,0}(x) \leq (1+\mu)V^0(x)$$

for sufficiently small $h > 0$ and for all $x \in B^h$. Since $V^h(x) \leq V^{h,0}(x)$, we complete the proof of the upper bound by taking $h > 0$ sufficiently small so that $(1+\mu)/(1 - hc_0^{-1}CK) \leq 1 + m$.

Lower bound. Let $\mu < m$, and this time put $W^0 = (1-\mu)V^0$. It follows from the DPE (2.3) that the relation

$$\langle u^h, DV^0\rangle + L_{u^h} - \phi^h = 0$$

holds on N^h for some nonnegative function ϕ^h. This, in turn, implies that

$$\langle u^{h,+}, D^{h,+}W^0\rangle - \langle u^{h,-}, D^{h,-}W^0\rangle + \tilde{L}^h_{u^h} = 0, \tag{5.7}$$

where $\tilde{L}^h_{u^h}(x)$ is defined on N^h by

$$\begin{aligned}\tilde{L}^h_{u^h} &= (1-\mu)L_{u^h} - (1-\mu)\phi^h \\ &\quad + \langle u^h, DW^0\rangle - \langle u^{h,+}, D^{h,+}W^0\rangle + \langle u^{h,-}, D^{h,-}W^0\rangle.\end{aligned}$$

Since $L_{u^h}(x) \geq c_0 > 0$ and $W^0(x)$ is smooth, the nonnegativity of $\phi^h(x)$ implies that $L_{u^h}(x) \geq \tilde{L}^h_{u^h}(x)$ for $h > 0$ sufficiently small and for all $x \in N^h$. The generator in (5.7) corresponds to applying the feedback control u^h in the h-dynamics, so we can use a standard verification argument to establish for all $x \in B^h$ the representation

$$W^0(x) = E_x\left[\int_0^{\tau_N^h} \tilde{L}^h_{u^h}(X^h)dt + W^0(z_N^h)\right]. \tag{5.8}$$

We use part (iii) of Lemma 4.1 to obtain the uniform integrability of τ_N^h needed for the right-hand side to be finite. The strong Markov property implies the representation

$$V^h(x) = E_x\left[\int_0^{\tau_N^h} L_{u^h}(X^h)dt + V^h(z_N^h)\right]. \tag{5.9}$$

Thus the following series of inequalities holds for all sufficiently small $h > 0$ and for all $x \in B^h$:

$$\begin{aligned} V^h(x) &\geq W^0(x) + E_x\left[V^h(z_N^h) - W^0(z_N^h)\right] \\ &\geq (1-\mu)V^0(x) - \sup_{y\in G^h}\left|V^h(y) - V^0(y)\right| P_x\left[z_N^h \notin \partial G\right] \\ &= (1-\mu)V^0(x) - o_h(1)P_x\left[z_N^h \notin \partial G\right]. \end{aligned} \tag{5.10}$$

The first line follows from the representations (5.8) and (5.9), along with the fact that $L_{u^h}(x) \geq \tilde{L}^h_{u^h}(x)$ for all $x \in N^h$; the second line uses the definition of $W^0(x)$,

the nonnegativity of $V^0(x)$ for all $x \in G$, and the fact that $V^h(x) = V^0(x) = 0$ for all $x \in \partial G$; finally, the third line uses part (ii) of Theorem 4.2, and the $o_h(1)$ term converges to zero as $h \to 0$, uniformly for all $x \in B^h$.

Let $\varepsilon > 0$ be equal to $d(B, \partial N \cap G)/2$. Notice that any trajectory with an initial condition $x \in B$ must travel a distance of at least 2ε if it is to exit N at a point which is not in ∂G; see Figure 3. Let $K \geq \sup_{u \in U} \|u\|$ be such that $\varepsilon/K \leq T$. Then, given the semimartingale decomposition in (3.2), $z_N^h \in \partial G$ will follow if $\tau_N^h < \varepsilon/K \leq T$ and if $\| \, m^h \, \|_{\tau_N^h \wedge T} < \varepsilon$. By Chebyshev's inequality and by the lower bound c_0 on the running cost, for each $x \in B^h$ we have

$$\begin{aligned} P_x \left[\tau_N^h \geq \varepsilon/K \right] &\leq (\varepsilon/K)^{-1} E_x \tau_N^h \\ &\leq (\varepsilon c_0/K)^{-1} V^h(x). \end{aligned}$$

As in (5.6), we can use a standard submartingale inequality and Lemma 3.1 to verify that

$$P_x \left[\| \, m^h \, \|_{\tau_N^h \wedge T} \geq \varepsilon \right] \leq hCV^h(x)$$

for all $x \in B^h$, where C is a finite constant. Thus, for a composite constant C', we conclude that

$$P_x \left[z_N^h \notin \partial G \right] \leq C' V^h(x)$$

for all $x \in B^h$. Combining this last bound with (5.10), we obtain

$$(1 + o_h(1)) V^h(x) \geq (1 - \mu) V^0(x)$$

for all $h > 0$ sufficiently small and for all $x \in B^h$. By taking $h > 0$ sufficiently small so that $(1 - \mu)/(1 + o_h(1)) \geq 1 - m$ on B^h, we complete the proof of the lower bound. □

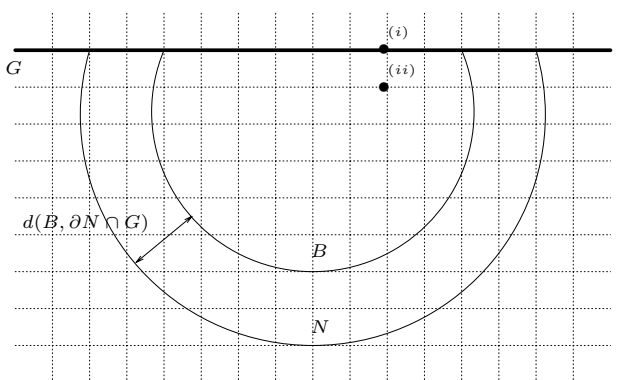


FIG. 3. *Boundary points.*

We are now able to prove that the approximations $D^{h,\pm} V^h(x)$ converge uniformly to the gradient $DV^0(x)$ at appropriate points x near the boundary of B. The two cases in the following lemma are illustrated in Figure 3.

LEMMA 5.3. *For $\varepsilon > 0$ there exists $h_0 > 0$ such that for all $0 < h \le h_0$ and for each $i = 1, \dots, n$,*

$$|D_i^{h,+}V^h(x) - D_iV^0(x)| \le \varepsilon \quad (resp.,\ |D_i^{h,-}V^h(x) - D_iV^0(x)| \le \varepsilon)$$

for each $x \in \mathbb{R}^n$ such that either (i) *$x \in \partial G$ and $x + he_i \in B^h$ (resp., $x - he_i \in B^h$), or* (ii) *$x \in B^h$ and $x + he_i \in \partial G$ (resp., $x - he_i \in \partial G$).*

Proof. For simplicity, we treat only case (i) with $x \in \partial G$ and $x + he_i \in B^h$. Given the smoothness of D_iV^0, the other cases follow easily from the same argument. Fix $\varepsilon > 0$, and let K be the uniform Lipschitz constant for $V^0(x)$. Then by Lemma 5.2, there exists $h_0 > 0$ such that

$$\left| \frac{V^h(x + he_i) - V^0(x + he_i)}{V^0(x + he_i)} \right| \le \frac{\varepsilon}{2} K^{-1} \tag{5.11}$$

for all $0 < h \le h_0$ and for all $x \in \mathbb{R}^n$ satisfying condition (i). For such x and h, put $y = x + he_i$. Then, using (5.11) along with the fact that $V^0(x)$ and $V^h(x)$ satisfy zero boundary conditions on ∂G, we obtain

$$\begin{aligned} |D_i^{h,+}V^h(x) - D_i^{h,+}V^0(x)| &= \left| \frac{V^h(y) - V^0(y)}{h} \right| \\ &= \left| \frac{V^0(y)}{h} \right| \left| \frac{V^h(y) - V^0(y)}{V^0(y)} \right| \\ &\le K \left| \frac{V^h(y) - V^0(y)}{V^0(y)} \right| \\ &\le \varepsilon/2. \end{aligned} \tag{5.12}$$

Now let $h_0 > 0$ be sufficiently small so that

$$|D_i^{h,+}V^0(x) - D_iV^0(x)| \le \varepsilon/2$$

for all $0 < h \le h_0$ and $x \in \bar{B}$. Then the result follows from (5.12). □

The first step in extending the result of Lemma 5.3 to the interior of B_0 is to establish a representation for $DV^0(x)$ in terms of an integral of the gradient in x of the running cost $L(x, u)$ along the optimal trajectories. The proof we give for this representation in the next lemma is fairly simple because it involves only deterministic trajectories. An analogous argument, involving stochastic trajectories, will be used to establish the convergence of $D^{h,\pm}V^h(x)$ to $DV^0(x)$ in the proof of Lemma 5.1. We also note that the representation in Lemma 5.4 can be obtained by the classical method of characteristics [13]. Recall the notation $L_u(x) = L(x, u(x))$ for a feedback control $u(x)$.

LEMMA 5.4. *The representation*

$$DV^0(x) = \int_0^{\tau_x^0} DL_{u^0}(X_x^0)dt + DV^0(z_x^0) \tag{5.13}$$

holds for all initial conditions $x \in \bar{B}_0$.

Proof. We fix $x \in \bar{B}_0$ and establish the representation separately for each component $D_iV^0(X)$ of the gradient $DV^0(x)$. Without loss of generality, assume that

$x + he_i \in N$ for all sufficiently small $h > 0$. Since $D_i^{h,+}V^0(x) \to D_iV^0(x)$ as $h \to 0$, we can establish the representation by proving separately the upper bound

$$\limsup_{h\to 0} D_i^{h,+}V^0(x) \le \int_0^{\tau_x^0} D_iL_{u^0}(X_x^0)dt + D_iV^0(z_x^0)$$

and the lower bound

$$\liminf_{h\to 0} D_i^{h,+}V^0(x) \ge \int_0^{\tau_x^0} D_iL_{u^0}(X_x^0)dt + D_iV^0(z_x^0).$$

Upper bound. Let $\hat{\tau}_h$ be the minimum of τ_x^0 and the first exit time from N of the shifted trajectory $X_x^0(t) + he_i$, and let $\hat{z}_h = X_x^0(\hat{\tau}_h)$. Using the fact that $X_x^0(t) + he_i$ is a suboptimal trajectory for the initial condition $x + he_i$, we obtain the relations

$$\begin{aligned} D_i^{h,+}V^0(x) &\le \frac{1}{h}\left[\int_0^{\hat{\tau}_h} L(X_x^0 + he_i, u^0(X_x^0))dt + V^0(\hat{z}_h + he_i)\right. \\ &\quad \left. - \int_0^{\hat{\tau}_h} L(X_x^0, u^0(X_x^0))dt - V^0(\hat{z}_h)\right] \\ &= \int_0^{\hat{\tau}_h} D_i^{h,+}L_{u^0}(X_x^0)dt + D_i^{h,+}V^0(\hat{z}_h). \end{aligned}$$

Lemma 2.3 implies that $|\hat{\tau}_h - \tau_x^0|$ and $\|\hat{z}_h - z_x^0\|$ both converge to zero as $h \to 0$, so we can apply the Lebesgue dominated convergence theorem to obtain the upper bound.

Lower bound. This time, let $\hat{\tau}_h$ be the minimum of $\tau_{x+he_i}^0$ and the first exit time from N of the shifted trajectory $X_{x+he_i}^0(t) - he_i$, and let $\hat{z}_h = X_{x+he_i}^0(\hat{\tau}_h)$. Using the fact that $X_{x+he_i}^0(t) - he_i$ is a suboptimal trajectory for the initial condition x, we obtain the relations

$$\begin{aligned} D_i^{h,+}V^0(x) &\ge \frac{1}{h}\left[\int_0^{\hat{\tau}_h} L(X_{x+he_i}^0, u^0(X_{x+he_i}^0))dt + V^0(\hat{z}_h)\right. \\ &\quad \left. - \int_0^{\hat{\tau}_h} L(X_{x+he_i}^0 - he_i, u^0(X_{x+he_i}^0))dt - V^0(\hat{z}_h - he_i)\right] \\ &= \int_0^{\hat{\tau}_h} D_i^{h,-}L_{u^0}(X_{x+he_i}^0)dt + D_i^{h,-}V^0(\hat{z}_h). \end{aligned}$$

By Lemma 2.4, $|||\, X_{x+he_i}^0 - X_x^0 \,|||_T$ converges to zero as $h \to 0$. Thus Lemma 2.3 implies that $|\hat{\tau}_h - \tau_x^0|$ and $\|\hat{z}_h - z_x^0\|$ both converge to zero as $h \to 0$, and we can apply the Lebesgue dominated convergence theorem to obtain the lower bound. □

We can now use the representation for $DV^0(x)$ obtained in Lemma 5.4 to prove Lemma 5.1. In fact, the proofs are essentially analogous. The primary difference is that the trajectories which arise in the proof of Lemma 5.1 are stochastic, so the analysis of the limits as $h \to 0$ is more involved.

Proof of Lemma 5.1. We give a detailed argument only for the convergence $D_i^{h,+}V^h(x^h) \to D_iV^0(x)$. Fix $x \in \bar{B}_0$, and let $x^h \in B_0^h$ be such that $x^h \to x$ as

$h \to 0$. As usual, we prove separately the upper bound

$$\limsup_{h\to 0} D_i^{h,+} V^h(x^h) \le D_i V^0(x)$$

and the lower bound

$$\liminf_{h\to 0} D_u^{h,+} V^h(x^h) \ge D_i V^0(x).$$

Upper bound. Let $\hat{\tau}_h$ be the minimum of $\tau^h_{x^h,N}$ and the first exit time from the interior of N of the shifted trajectory $X^h_{x^h}(t) + he_i$, and let $\hat{z}_h = X^h_{x^h}(\hat{\tau}_h)$. Using the fact that the trajectory $X^h_{x^h}(t) + he_i$ results when the suboptimal feedback control $\tilde{u}^h(\cdot) = u^h(\cdot - he_i)$ is applied in the h-dynamics with initial condition $x^h + e_i$, we obtain the relations

$$\begin{aligned} D_i^{h,+} V^h(x^h) &\le \frac{1}{h} E_x^h \left[\int_0^{\tilde{\tau}_h} L_{\tilde{u}^h}(X^h_{x^h} + he_i)dt + V^h(\tilde{z}_h + he_i) \right. \\ &\qquad \left. - \int_0^{\tilde{\tau}_h} L_{u^h}(X^h_{x^h})dt - V^h(\tilde{z}_h) \right] \\ &= E_{x^h} \left[\int_0^{\hat{\tau}_h} D_i^{h,+} L_{u^h}(X^h_{x^h})dt + D_i^{h,+} V^h(\hat{z}_h) \right]. \end{aligned}$$

In light of the representation (5.13), this implies that we can establish the upper bound by showing that

$$E_{x^h} \left| \int_0^{\hat{\tau}_h} D_i^{h,+} L(X^h_{x^h}, u^h(X^h_{x^h}))dt - \int_0^{\tau^0_x} D_i L(X^0_x, u^0(X^0_x))dt \right| \tag{5.14}$$

and

$$E_{x^h} \left| D_i^{h,+} V^h(\hat{z}_h) - D_i V^0(z^0_{x,N}) \right| \tag{5.15}$$

both converge to zero as $h \to 0$. Recall from Remark 3.2 and Lemma 3.3 that the pair $(X^h, u^h(X^h))$ takes values in a compact set for all initial conditions and for all $h \ge 0$. Thus the smoothness of L implies that we can use the triangle inequality to bound the quantity in (5.14) by a constant times

$$E_{x^h} \left[\|X^h_{x^h} - X^0_x\|_T + \|u^h(X^h_{x^h}) - u^0(X^0_x)\|_T + |\hat{\tau}_h - \tau^0_x| + hT \right]. \tag{5.16}$$

By combining parts (i) and (ii) of Lemma 4.1 with Lemma 2.4, we can establish that the first two terms in (5.16) converge to zero as $h \to 0$. Uniform integrability of the $\hat{\tau}_h$ follows from the strong Markov property and from the fact that Lemmas 2.3 and 2.4 together imply that $\hat{\tau}_h \le T$ with positive probability, uniformly in h. Thus, since $\hat{\tau}_h \to \tau^0_x$ in probability, the third term in (5.16) converges to zero as $h \to 0$, and this implies that the expression in (5.14) converges to zero as $h \to 0$.

Applying the triangle inequality to (5.15), we find that it is bounded by

$$E_{x^h} \left[|D_i^{h,+} V^h(\hat{z}_h) - D_i V^0(\hat{z}_h)| + |D_i V^0(\hat{z}_h) - D_i V^0(z^0_x)| \right]. \tag{5.17}$$

Combining part (i) of Lemma 4.1 with Lemmas 2.3 and 2.4, we conclude that $\hat{z_h}$ converges to z_x^0 in probability. The continuity result in Lemma 2.4 and the fact that B is a region of strong regularity imply that the set $\{z_y^0 : y \in \bar{B}_0\}$ is compactly contained in B, so the probability of $\hat{z}^h$ satisfying the conditions of Lemma 5.3 increases to one as $h \to 0$. Thus we can use Lemma 5.3 and the smoothness of $D_i V^0$ to conclude that each of the terms in (5.17) converges to zero as $h \to 0$. That, in turn, implies that (5.14) converges to zero as $h \to 0$ and completes the proof of the upper bound.

Lower bound. This time, we let $\hat{\tau}_h$ be the minimum of $\tau^h_{x+he_i}$ and the first exit time from N of the shifted trajectory $X^h_{x+he_i}(t) - he_i$, and we let $\hat{z}_h = X^h_{x+he_i}(\hat{\tau}_h)$. As in the proof of the upper bound, we obtain

$$D_i^{h,+}V^h(x^h) \quad \geq \quad E_{x^h}\left[\int_0^{\hat{\tau}_h} D_i^{h,-}L_{u^h}(X^h_{x^h+he_i})dt + D_i^{h,-}V^h(\hat{z}_h)\right].$$

See also the analogous relation which appears in the proof of the lower bound for Lemma 5.4. Just as in the proof of the upper bound, we show that the right-hand side of the above expression converges to $D_i V^0(x)$ as $h \to 0$. Notice that we need $x^h + he_1 \in B^h$ in order to apply Lemma 4.1. Since $x^h \in B_0$, this condition is satisfied for all sufficiently small $h > 0$. □

THEOREM 5.5. *Let $x^h \in B_0^h$ be such that $x^h \to x \in \bar{B}_0$ as $h \to 0$. Then, as $h \to 0$,*

$$u^h(x^h) \to u^0(x).$$

Proof. For each $u \in U$, we define

$$F^0(u) = \langle u, DV^0(x)\rangle + L(x,u)$$

and

$$F^h(u) = \langle u^+, D^{h,+}V^h(x^h)\rangle - \langle u^-, D^{h,-}V^h(x^h)\rangle + L(x^h,u).$$

Recall from Lemma 3.3 that the optimal controls u^0 and u^h take values in the compact set U. Thus there exists $\tilde{u} \in U$ such that in a subsequence, $u^h(x^h) \to \tilde{u}$ as $h \to 0$. It suffices to show that $\tilde{u} = u^0(x)$. Since x is in a region of strong regularity, it follows from the DPE (2.3) that the unique minimizer of $F^0(u)$ is given by $u^0(x)$. Similarly, the DPE (3.5) implies that $u^h(x^h)$ is a minimizer of $F^h(u)$. Also, notice that $F^0(u)$ is a continuous function of u and that Lemma 5.1 implies that $F^h(u)$ converges to $F^0(u)$ for each $u \in U$. Thus we obtain the following relations:

$$F^0(u^0(x)) \;=\; \lim_{h\to 0} F^h(u^0(x)) \;\geq\; \lim_{h\to 0} F^h(u^h(x^h)) \;=\; F^0(\tilde{u}).$$

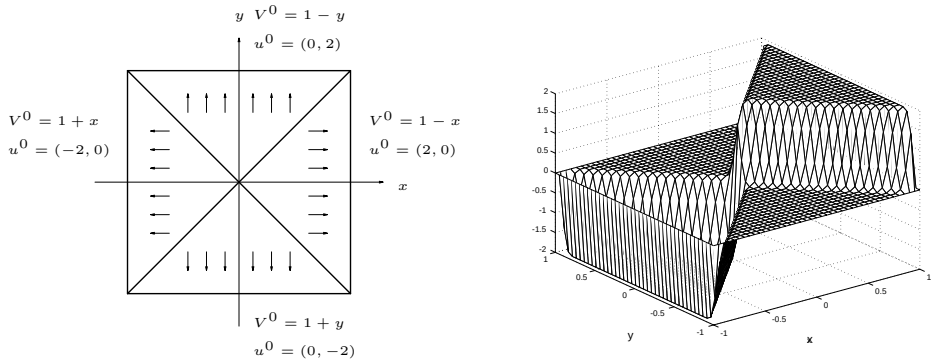
Since $u^0(x)$ is the unique minimizer of F^0, it follows that $\tilde{u} = u^0(x)$. That completes the proof of the theorem. □

COROLLARY 5.6. *For any $\varepsilon > 0$, there exist $h_0 > 0$ such that*

$$\|u^h(x) - u^0(x)\| < \varepsilon$$

for all $0 < h \leq h_0$ and for all $x \in B_0^h$.

Proof. Since $u^0(x)$ is uniformly continuous on $\bar{B}_0$, the result follows from Theorem 5.5 and a standard argument by contradiction. □



(a) (b)

FIG. 4. (a) *Exact solution;* (b) *approximate control.*

6. Computational examples. In this section, we present examples of approximations obtained by numerically solving the DPE (3.5) for the Markov chain optimal control problem. The fixed point of that equation is taken as the approximation to the value function $V^0(x)$, and the infimizing feedback control is taken as an approximate feedback control for the limit problem. The solution of (3.5) is obtained by using either Jacobi or Gauss–Seidel iteration in the equivalent equation (3.6), and we note that the infima at each step can be evaluated analytically [5]. When the Gauss–Seidel method is used, we observe in our examples that the number of iterations required to find the fixed point is essentially independent of the parameter h.

Example 1. Our first example is a minimum escape time problem on the unit square $G = [-1, 1] \times [-1, 1]$ in $\mathbb{R}^2$ with running cost

$$L(x, u) = \frac{1}{4}\|u\|^2 + 1.$$

The domain can be decomposed into four regions of strong regularity on which the value function is smooth and on which there is a smooth optimal feedback control. This decomposition and the optimal values are indicated in Figure 4(a). Arrows indicate the direction of the optimal velocity field. Figure 4(b) displays the values of the first component of the feedback control $u^h(x)$ for $h = 0.05$. We see that the discontinuities in the optimal control are resolved very sharply by our approximation scheme. In Table 1, we indicate errors in the approximations to the controls. The L^1 errors reported there are for the entire domain, while the L^∞ errors are for points inside the regions of strong regularity and a distance of at least 0.1 from the discontinuities. We also indicate the number of iterations required for the Gauss–Seidel method to achieve a residual of less than 10^{-8}, our standard tolerance. We see that the errors in the optimal controls become machine zero on the regions of strong regularity, which is in part a consequence of the fact that the value function is linear in those regions. Additionally, the optimal controls appear to converge in L^1 on the entire domain. Without detailed assumptions about the structure of the regions of strong regularity, our results do not necessarily predict that type of convergence. However, since we know for the present problem that the complement of the regions of strong regularity has Lebesgue measure zero, convergence in L^1 on the entire domain is, in fact, expected. Finally, similar error values are indicated in the second part of Table 1 for

TABLE 1
Escape time problem errors.

	$n = 2$			$n = 3$		
h	Iter	L^1	L^∞ RSR	Iter	L^1	L^∞ RSR
0.2	8	1.74 e − 00	5.91 e − 01	9	4.36 e − 00	5.91 e − 01
0.1	8	1.01 e − 00	8.77 e − 02	17	2.77 e − 00	8.77 e − 02
0.05	8	5.43 e − 01	1.92 e − 03	17	1.55 e − 00	1.92 e − 03
0.025	8	2.80 e − 01	2.31 e − 13	17	8.22 e − 01	3.94 e − 13

the escape time problem on the unit cube in $\mathbb{R}^3$.

Example 2. Our next example involves a value function which is obtained by perturbing the value function for the escape time problem in $\mathbb{R}^2$. We take care to modify the value function and the cost structure in such a way that we obtain a new problem with smooth data and with a solution that can be evaluated analytically. To that end, we introduce the C^∞ double bump function defined by

$$\chi(\xi) = \begin{cases} e^{-\lambda((\xi-m)^2-\sigma^2)^{-2}}, & \xi \in [m-\sigma, m+\sigma], \\ e^{-\lambda((-\xi-m)^2-\sigma^2)^{-2}}, & \xi \in [-m-\sigma, -m+\sigma], \\ 0 & \text{otherwise}, \end{cases}$$

where we use the parameter values $m = 0.7$, $\sigma = 0.5$, and $\lambda = 0.07$. Now we define a mollifier by

$$\Phi(x, y) = \chi(x + y)\chi(x - y)$$

for all (x, y) in the unit square, and then we define the value function $V^0(x, y)$ by multiplying the value function for the escape time problem by $1 + \Phi(x, y)$. The resulting function has the same regions of strong regularity as indicated in Figure 4(a), and it maintains the linear structure in a neighborhood of the discontinuities. In a similar spirit, we define

$$a(x, y) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + 3\sin(2\pi x)^2 \begin{bmatrix} 2 & 5 \\ 5 & 18 \end{bmatrix} \Phi(x, y)$$

and

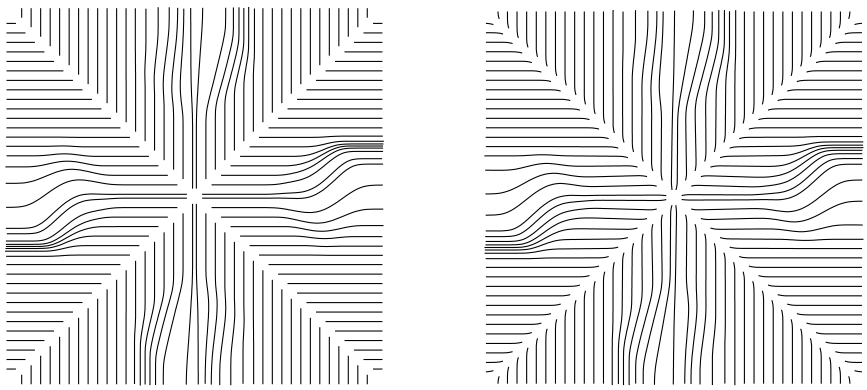
$$b(x, y) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} + 5 \begin{bmatrix} x \\ y\sin((x^2 + y^2)^{1/2} - 1/2) \end{bmatrix} \Phi(x, y),$$

so that $a(x, y)$ is the identity and $b(x, y)$ is the zero vector in a neighborhood of the discontinuities. Now we define $c(x, y)$ on the regions of strong regularity by

$$c(x, y) = (1/2)\Big\langle DV^0(x, y), a(x, y)DV^0(x, y)\Big\rangle - \Big\langle b(x, y), DV^0(x, y)\Big\rangle.$$

Our use of a mollifier in defining all of the relevant functions ensures that the cost function $c(x, y)$ extends smoothly to $c(x, y) = 1$ at the discontinuities, and it turns out that $V^0(x)$ solves the limit control problem for the indicated cost structure.

The optimal trajectories for this problem are indicated in Figure 5(a), while trajectories computed using the approximate optimal controls with $h = 0.025$ are shown in Figure 5(b). Clearly, the controls computed with our algorithm yield an excellent approximation to the optimal trajectories. In Table 2, we exhibit error values for the



(a) (b)

FIG. 5. (a) *Characteristics;* (b) *approximate characteristics.*

TABLE 2
Perturbed escape time problem errors.

h	Iter	L^1	L^∞ RSR
0.1	10	5.83 e − 01	2.66 e − 01
0.05	12	3.24 e − 01	1.59 e − 01
0.025	12	1.72 e − 01	8.94 e − 02
0.0125	12	8.93 e − 02	4.84 e − 02
0.00625	12	4.55 e − 02	2.54 e − 02

approximations to the optimal control with the L^1 errors being on the entire domain and the L^∞ errors being for points a distance of at least 0.1 from the discontinuities. Evidently, both measures of the error are approximately proportional to h, and it is also worth noting that the number of iterations required for the Gauss–Seidel procedure to converge to the fixed point is essentially constant. In Figure 6, we display an approximation to the first component of the optimal control with $h = 0.05$ and the errors in the approximation to the control for $h = 0.05$. The discontinuities are resolved very sharply, and we can see that the errors are uniformly small within the regions of strong regularity

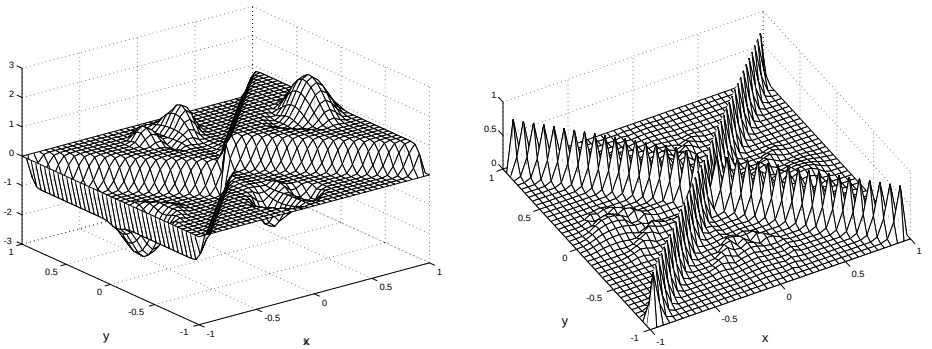
Example 3. Our final example is an application to the problem of finding geodesics on a surface, suggested in [20]. Given a surface $z(x, y)$ on the unit square in $\mathbb{R}^2$, the problem is to find the shortest path along the surface from a given point to the boundary. It is shown in [20, section 16.5] that the solution to this problem can be obtained from our optimal control problem with running cost specified by

$$a(x,y) = \begin{bmatrix} 1+z_y^2 & z_x z_y \\ z_x z_y & 1+z_x^2 \end{bmatrix}, \qquad b(x,y) = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

and

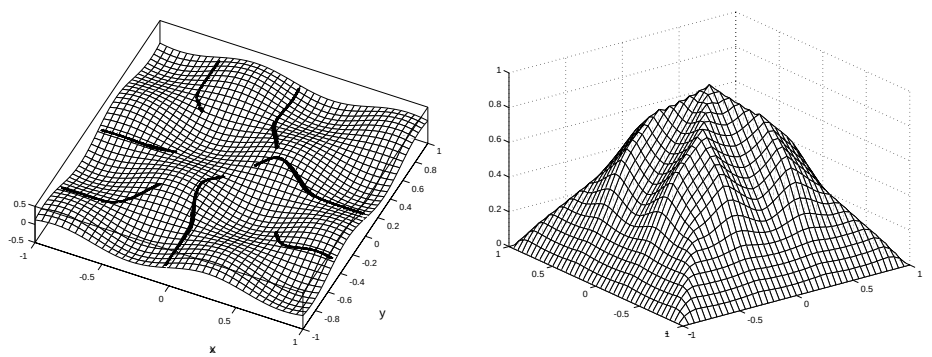
$$c(x,y) = (1/2)(1+z_x^2+z_y^2).$$

Geodesics are obtained by following the optimal trajectories from points on the interior of the unit square to the boundary. In Figure 7(a), we show several approximate



(a) (b)

FIG. 6. (a) *Approximate control;* (b) *control error.*



(a) (b)

FIG. 7. (a) *Geodesics;* (b) *value function.*

geodesics computed using our algorithm for the sinusoidal surface

$$z(x,y) = (1/4)\sin\left(\frac{7\pi}{4}x\right)\sin\left(\frac{7\pi}{4}y\right).$$

The approximate controls are computed on a grid with spacing $h = 0.05$, and the trajectories are integrated by a simple Euler method with linear interpolation. In Figure 7(b), we show the value function computed with $h = 0.05$ for the corresponding control problem, illustrating the fairly complex structure of the regions of strong regularity. Since we do not know the exact solution for this problem, it is not possible for us to present a numerical measure of accuracy for the approximate geodesics in Figure 7(a). However, Theorem 5.5 guarantees that for initial conditions in a region of strong regularity, the approximations will converge to the correct geodesics as $h \to 0$. Being that a more refined grid does not result in discernible changes to the paths indicated in Figure 7(a), we conclude that these are, in fact, good approximations to the exact geodesic curves.

REFERENCES

[1] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser Boston, Boston, 1997.

[2] M. Bardi and P. Soravia, *Hamilton-Jacobi equations with singular boundary conditions on a free boundary and applications to differential games*, Trans. Amer. Math. Soc., 325 (1991), pp. 205–229.

[3] A. Bensoussan and H. Nagai, *Min-max characterization of a small noise limit on risk-sensitive control*, SIAM J. Control Optim., 35 (1997), pp. 1093–1115.

[4] P. Billingsley, *Convergence of Probability Measures*, John Wiley and Sons, New York, 1968.

[5] M. Boué and P. Dupuis, *Markov chain approximations for deterministic control problems with affine dynamics and quadratic cost in the control*, SIAM J. Numer. Anal., 36 (1999), pp. 667–695.

[6] P. Cannarsa, A. Mennucci, and C. Sinestrari, *Regularity results for solutions of a class of Hamilton-Jacobi equations*, Arch. Ration. Mech. Anal., 140 (1997), pp. 197–223.

[7] P. Cannarsa and C. Sinestrari, *Convexity properties of the minimum time function*, Calc. Var. Partial Differential Equations, 3 (1995), pp. 273–298.

[8] M. G. Crandall and P. L. Lions, *Two approximations of solutions of Hamilton-Jacobi equations*, Math. Comp., 43 (1984), pp. 1–19.

[9] M. H. A. Davis, *Piecewise deterministic Markov processes: A general class of non-diffusion stochastic models*, J. Roy. Statist. Soc. Ser. B, 46 (1984), pp. 353–388.

[10] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*, John Wiley and Sons, New York, 1997.

[11] P. Dupuis and J. Oliensis, *An optimal control formulation and related numerical methods for a problem in shape reconstruction*, Ann. Appl. Probab., 4 (1994), pp. 287–346.

[12] B. Fitzpatrick and W. H. Fleming, *Numerical methods for an optimal investment-consumption model*, Math. Oper. Res., 16 (1991), pp. 823–841.

[13] W. H. Fleming, *The Cauchy problem for a nonlinear first order partial differential equation*, J. Differential Equations, 5 (1967), pp. 515–530.

[14] W. H. Fleming, *Stochastic control for small noise intensities*, SIAM J. Control, 9 (1971), pp. 473–517.

[15] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[16] W. H. Fleming and P. E. Souganidis, *Asymptotic series and the method of vanishing viscosity*, Indiana Univ. Math. J., 35 (1986), pp. 425–447.

[17] G. Folland, *Introduction to Partial Differential Equations*, Princeton University Press, Princeton, 1976.

[18] H. J. Kushner, *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*, Academic Press, New York, 1977.

[19] H. J. Kushner and P. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time*, Springer-Verlag, New York, 1992.

[20] J. A. Sethian, *Level Set Methods: Evolving Interfaces in Geometry, Fluid Mechanics, Computer Vision, and Materials Science*, Cambridge University Press, Cambridge, UK, 1996.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 2, pp. 421–449

ON A BOUNDARY CONTROL APPROACH TO DOMAIN EMBEDDING METHODS*

L. BADEA† AND P. DARIPA‡

Abstract. In this paper, we propose a domain embedding method associated with an optimal boundary control problem with boundary observations to solve elliptic problems. We prove that the optimal boundary control problem has a unique solution if the controls are taken in a finite dimensional subspace of the space of the boundary conditions on the auxiliary domain.

Using a controllability theorem due to J. L. Lions, we prove that the solutions of Dirichlet (or Neumann) problems can be approximated within any prescribed error, however small, by solutions of Dirichlet (or Neumann) problems in the auxiliary domain taking an appropriate subspace for such an optimal control problem. We also prove that the results obtained for the interior problems hold for the exterior problems. Some numerical examples are given for both the interior and the exterior Dirichlet problems.

Key words. domain embedding methods, optimal control

AMS subject classifications. 93B05, 93B07, 93B40, 65N30, 65P05, 65R20

PII. S0363012999357380

1. Introduction.

The embedding or fictitious domain methods, which were developed especially in the seventies (see [6], [2], [34], [35], [28], or [14]), have been a very active area of research in recent years because of their appeal and potential for applications in solving problems in complicated domains very efficiently. In these methods, complicated domains ω, where solutions of problems may be sought, are embedded into larger domains Ω with simple enough boundaries so that solutions in these embedded domains can be constructed more efficiently. The use of these embedding methods is now commonplace for solving complicated problems arising in science and engineering. To this end, it is worth mentioning the domain embedding methods for Stokes equations (Borgers [5]), for fluid dynamics and electromagnetics (Dinh et al. [12]), and for the transonic flow calculation (Young et al. [36]).

In [3], an embedding method is associated with a distributed optimal control problem. There the problem is solved in an auxiliary domain Ω using a finite element method on a fairly structured mesh which allows the use of fast solvers. The auxiliary domain Ω contains the domain ω, and the solution in Ω is found as a solution of a distributed optimal control problem such that it satisfies the prescribed boundary conditions of the problem in the domain ω. The same idea is also used in [10], where a least squares method is used. In [13], an embedding method is proposed in which a combination of Fourier approximations and boundary integral equations is used. Essentially, there a Fourier approximation for a particular solution of the inhomogeneous equation in Ω is found, and then the solution in ω for the homogeneous equation is sought using the boundary integral methods.

*Received by the editors June 14, 1999; accepted for publication (in revised form) January 24, 2001; published electronically July 19, 2001.
http://www.siam.org/journals/sicon/40-2/35738.html

†Institute of Mathematics, Romanian Academy of Sciences, P.O. Box 1-764, RO-70700 Bucharest, Romania (lbadea@imar.ro).

‡Department of Mathematics, Texas A&M University, College Station, TX 77843 (prabir.daripa@math.tamu.edu). The research of this author was supported by the Texas Advanced Research Program, grant TARP-97010366-030.

In recent years, progress in this field has been substantial, especially in the use of the Lagrange multiplier techniques. In this connection, the works of Girault, Glowinski, Hesla, Joseph, Kuznetsov, Lopez, Pan, and Périaux (see [15], [16], [17], [18], and [19]) should be cited.

There are many problems for which an exact solution on some particular domains may be known or computed numerically within a good approximation very efficiently. In these cases, an embedding domain method associated with a boundary optimal control problem allows one to find solutions of the problems very efficiently in complicated domains. Specifically, the particular solution of the inhomogeneous equation can be used to reduce the problem to solving a homogeneous equation in ω subject to appropriate conditions on the boundary of the domain ω. This solution in the complicated domain ω can be obtained via an optimal boundary control problem where one finds the solution of the same homogeneous problem in the auxiliary domain Ω that would satisfy appropriate boundary conditions on the domain ω. We mention that the boundary control approach already has been used by Mäkinen, Neittaanmäki, and Tiba for optimal shape design and two-phase Stefan-type problems (see [29], [32]). Moreover, recently there has been enormous progress in shape optimization using the fictitious domain approaches. We can cite here, for instance, the works of Daňková, Haslinger, Klarbring, Makinen, Neittaanmäki, and Tiba (see [9], [22], [23], and [33]) among many others.

In section 2, an optimal boundary control problem involving an elliptic equation is formulated. In this formulation, the solution on the auxiliary domain Ω is sought such that it satisfies the boundary conditions on the domain ω. In general, such an optimal control problem leads to an ill posed problem, and, consequently, it may not have a solution.

Using a controllability theorem of J. L. Lions, it is proved here that the solutions of the problems in ω can be approximated within any specified error, however small, by the solutions of the problems in Ω for appropriate values of the boundary conditions. In section 3, it is shown that the optimal control problem has a unique solution in a finite dimensional space. Consequently, considering a family of finite dimensional subspaces with their union dense in the whole space of controls, we can approximate the solution of the problem in ω with the solutions of the problems in Ω using finite dimensional optimal boundary control problems. Since the values of the solutions in Ω are approximately calculated on the boundary of the domain ω, we study the optimal control problem with boundary observations in a finite dimensional subspace in section 4. In section 5, we extend the results obtained for the interior problems to the exterior problems. In section 6, we give some numerical examples for both bounded and unbounded domains. The numerical results are presented to show the validity and high accuracy of the method. Finally, in section 7 we provide some concluding remarks. There is still a large room for further improvement and numerical tests. In future works, we will apply this method in conjunction with fast algorithms (see [4], [7], [8]) to solve other elliptic problems in complicated domains.

2. Controllability. Let ω, $\Omega \in \mathcal{N}^{(1),1}$ (i.e., the maps defining the boundaries of the domains and their derivatives are Lipschitz continuous) be two bounded domains in $\mathbf{R}^N$ such that $\bar{\omega} \subset \Omega$. Their boundaries are denoted by γ and Γ, respectively.

In this paper, we use domain embedding and the optimal boundary control approach to solve the elliptic equation

$$Ay = f \quad \text{in } \omega, \tag{2.1}$$

subject to either Dirichlet boundary conditions

$$y = g_\gamma \quad \text{on } \gamma \tag{2.2}$$

or Neumann boundary conditions

$$\frac{\partial y}{\partial n_A(\omega)} = h_\gamma \quad \text{on } \gamma, \tag{2.3}$$

where $\frac{\partial}{\partial n_A(\omega)}$ is the outward conormal derivative associated with A.

We assume that the operator A is of the form

$$A = -\sum_{i,j=1}^{N} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial}{\partial x_j}\right) a_0$$

with $a_{ij} \in C^{(1),1}(\bar{\Omega})$, $a_0 \in C^{(0),1}(\bar{\Omega})$, $a_0 \geq 0$ in Ω, and there exists a constant $c > 0$ such that $\sum_{i,j=1}^{N} a_{ij}\xi_i\xi_j \geq c(\xi_1^2 + \cdots + \xi_N^2)$ in Ω for any $(\xi_1, \ldots, \xi_N) \in \mathbf{R}^N$. Also, we assume that $f \in L^2(\Omega)$, $g_\gamma \in L^2(\gamma)$, and $h_\gamma \in H^{-1}(\gamma)$.

For later use, we define the following. A function $y \in H^{1/2}(\omega)$ is called a solution of the Dirichlet problem (2.1)–(2.2) if it satisfies (2.1) in the sense of distributions and the boundary conditions (2.2) in the sense of traces in $L^2(\gamma)$. A function $y \in H^{1/2}(\omega)$ is called a solution of the Neumann problem (2.1), (2.3) if it satisfies (2.1) in the sense of distributions and the boundary conditions (2.3) in the sense of traces in $H^{-1}(\gamma)$ (see [27, Chap. 2, section 7]).

The Dirichlet problem (2.1)–(2.2) has a unique solution which depends continuously on the data

$$|y|_{H^{1/2}(\omega)} \leq C\{|f|_{L^2(\omega)} + |g_\gamma|_{L^2(\gamma)}\}. \tag{2.4}$$

If there exists a constant $c_0 > 0$ such that $a_0 \geq c_0$ in ω, then the Neumann problem (2.1), (2.3) has a unique solution which depends continuously on the data

$$|y|_{H^{1/2}(\omega)} \leq C\{|f|_{L^2(\omega)} + |h_\gamma|_{H^{-1}(\gamma)}\}. \tag{2.5}$$

If $a_0 = 0$ in ω, then the Neumann problem (2.1), (2.3) has a solution if

$$\int_\omega f + \int_\gamma h_\gamma = 0. \tag{2.6}$$

In this case, the problem has a unique solution in $H^{1/2}(\omega)/\mathbf{R}$ and

$$\inf_{r \in \mathbf{R}} |y + r|_{H^{1/2}(\omega)} \leq C\{|f|_{L^2(\omega)} + |h_\gamma|_{H^{-1}(\gamma)}\}. \tag{2.7}$$

We also remark that the solution of problem (2.1)–(2.2) can be viewed (see [27, Chap. 2, section 6]) as the solution of the problem

$$\begin{aligned} &y \in H^{1/2}(\omega) : \int_\omega yA^*\psi = \int_\omega f\psi - \int_\gamma g_\gamma \frac{\partial \psi}{\partial n_{A^*}(\omega)} \\ &\text{for any } \psi \in H^2(\omega),\ \psi = 0 \text{ on } \gamma, \end{aligned} \tag{2.8}$$

and that a solution of problem (2.1), (2.3) is also solution of the problem

$$\begin{aligned} &y \in H^{1/2}(\omega) : \int_\omega yA^*\psi = \int_\omega f\psi + \int_\gamma h_\gamma\psi \\ &\text{for any } \psi \in H^2(\omega),\ \frac{\partial \psi}{\partial n_{A^*}(\omega)} = 0 \text{ on } \gamma, \end{aligned} \tag{2.9}$$

where A^* is the adjoint operator of A given by

$$A^* = -\sum_{i,j=1}^{N} \frac{\partial}{\partial x_i}\left(a_{ji}\frac{\partial}{\partial x_j}\right) + a_0.$$

Evidently, the above results also hold for problems in the domain Ω.

We consider in the following only the cases in which the above problems have unique solutions, i.e., the Dirichlet problems, and we assume in the case of the Neumann problems that there exists a constant $c_0 > 0$ such that $a_0 \geq c_0$ in Ω.

Below we use the notations and the notions of optimal control from Lions [26]. First, we study the controllability of the solutions of the above two problems (defined by (2.1)–(2.3)) in ω with the solutions of a Dirichlet problem in Ω. Let

$$\mathcal{U} = L^2(\Gamma) \tag{2.10}$$

be the space of controls. The state of the system for a control $v \in L^2(\Gamma)$ is given by the solution $y(v) \in H^{1/2}(\Omega)$ of the following Dirichlet problem:

$$\begin{array}{lll} Ay(v) = f & \text{in} & \Omega, \\ y(v) = v & \text{on} & \Gamma. \end{array} \tag{2.11}$$

In the case of the Dirichlet problem (2.1)–(2.2), the space of observations is taken to be

$$\mathcal{H} = L^2(\gamma), \tag{2.12}$$

and the cost function is given by

$$J(v) = \frac{1}{2}|y(v) - g_\gamma|^2_{L^2(\gamma)}, \tag{2.13}$$

where $v \in L^2(\Gamma)$ and $y(v)$ is the solution of problem (2.11). For the Neumann problem given by (2.1) and (2.3), the space of observations is taken to be

$$\mathcal{H} = H^{-1}(\gamma), \tag{2.14}$$

and the cost function is given by

$$J(v) = \frac{1}{2}\left|\frac{\partial y(v)}{\partial n_A(\omega)} - h_\gamma\right|^2_{H^{-1}(\gamma)}. \tag{2.15}$$

Remark 2.1. Since $y(v) \in H^{1/2}(\Omega)$ and $Ay(u) = f \in L^2(\Omega)$, we have $y(v) \in H^2(D)$ for any domain D which satisfies $\bar{\omega} \subset D \subset \bar{D} \subset \Omega$ (see [30, Chap. 4, section 1.2, Theorem 1.3], for instance). Therefore, $y(v) \in H^{3/2}(\gamma)$ with the same values on both the sides of γ. Also, $\frac{\partial y(v)}{\partial n_A(\omega)} \in H^{1/2}(\gamma)$, $\frac{\partial y(v)}{\partial n_A(\Omega-\bar{\omega})} \in H^{1/2}(\gamma)$, and $\frac{\partial y(v)}{\partial n_A(\omega)} + \frac{\partial y(v)}{\partial n_A(\Omega-\bar{\omega})} = 0$. Consequently, the above two cost functions make sense.

PROPOSITION 2.1. *A control $u \in L^2(\Gamma)$ satisfies $J(u) = 0$, where the control function is given by* (2.13)*, if and only if the solution of* (2.11) *for $v = u$, $y(u) \in H^{1/2}(\Omega)$ satisfies*

$$\begin{array}{ll} Ay(u) = f & \text{in } \Omega - \bar{\omega}, \\ y(u) = y & \text{on } \gamma, \\ \frac{\partial y(u)}{\partial n_A(\Omega-\bar{\omega})} + \frac{\partial y}{\partial n_A(\omega)} = 0 & \text{on } \gamma, \end{array} \tag{2.16}$$

and

$$y(u) = y \quad \text{in } \omega, \tag{2.17}$$

where y is the solution of the Dirichlet problem defined by (2.1) *and* (2.2) *in the domain ω. The same result holds if the control function is given by* (2.15) *and y is the solution of the Neumann problem* (2.1) *and* (2.3).

Proof. Let $y(u) \in H^{1/2}(\Omega)$ be the solution of problem (2.11) corresponding to an $u \in L^2(\Gamma)$ such that $J(u) = 0$ with the control function given by (2.13). Consequently, $y(u)$ verifies (2.1) in the sense of distributions and the boundary condition (2.2) in the sense of traces. It gives $y(u) = y$ in ω. Since $y(u)$ satisfies (2.11) in $\Omega - \bar{\omega}$ in the sense of distributions, then, evidently, $y(u)$ is a solution of the equation in (2.16). From (2.17) and Remark 2.1, we obtain that $y(u)$ also satisfies the two boundary conditions of (2.16). The reverse implication is evident.

The same arguments also hold for the Neumann problem defined by (2.1) and (2.3) and the control function given by (2.15). □

Since (2.16) is not a properly posed problem, it follows from the above proposition that the optimal control might not exist. However, J. L. Lions proves in [26, Chap. 2, section 5.3, Theorem 5.1] a controllability theorem which can be directly applied to problem (2.11). We mention this theorem below.

LIONS'S CONTROLLABILITY THEOREM. *The set $\{\frac{\partial z_0(v)}{\partial n_A(\Omega-\bar{\omega})} \in H^{-1}(\gamma) : v \in L^2(\Gamma)\}$ is dense in $H^{-1}(\gamma)$, where $z_0(v) \in H^{1/2}(\Omega - \bar{\omega})$ is the solution of the problem*

$$\begin{aligned} &Az_0(v) = 0 && \text{in } \Omega - \bar{\omega}, \\ &z_0(v) = v && \text{on } \Gamma, \\ &z_0(v) = 0 && \text{on } \gamma. \end{aligned}$$

Now, we can easily prove the following lemma.

LEMMA 2.2. *For any $g \in L^2(\gamma)$, the set $\{\frac{\partial z(v)}{\partial n_A(\Omega-\bar{\omega})} \in H^{-1}(\gamma) : v \in L^2(\Gamma)\}$ is dense in $H^{-1}(\gamma)$, where $z(v) \in H^{1/2}(\Omega - \bar{\omega})$ is the solution of the problem*

$$\begin{aligned} &Az(v) = f && \text{in } \Omega - \bar{\omega}, \\ &z(v) = v && \text{on } \Gamma, \\ &z(v) = g && \text{on } \gamma. \end{aligned} \tag{2.18}$$

Proof. Let $z \in H^{1/2}(\Omega - \bar{\omega})$ be the solution of the problem

$$\begin{aligned} &Az = f && \text{in } \Omega - \bar{\omega}, \\ &z = 0 && \text{on } \Gamma, \\ &z = g && \text{on } \gamma. \end{aligned}$$

Using $z_0(v) = z(v) - z$ in the Lions controllability theorem, we get that the set $\{\frac{\partial(z(v)-z)}{\partial n_A(\Omega-\bar{\omega})} \in H^{-1}(\gamma) : v \in L^2(\Gamma)\}$ is dense in $H^{-1}(\gamma)$. Hence the lemma follows. □

The following theorem proves controllability of the solutions of problems in ω by the solutions of Dirichlet problems in Ω. In the proof of this theorem below, we use the spaces Ξ^s introduced in Lions and Magenes [27, Chap. 2, section 6.3]. For the sake of completeness, we give definitions of these spaces Ξ^s.

Let $\rho(x)$ be a function in $\mathcal{D}(\bar{\Omega})$ which is positive in Ω and vanishes on Γ. We also assume that for any $x_0 \in \Gamma$, the limit

$$\lim_{x \to x_0 \in \Gamma} \frac{\rho(x)}{d(x, \Gamma)}$$

exists and is positive, where $d(x,\Gamma)$ is the distance from $x \in \Omega$ to the boundary Γ. Then, for $s = 0, 1, 2, \ldots$, the space Ξ^s is defined by

$$\Xi^s(\Omega) = \{u : \rho^{|\alpha|} D^\alpha u \in L^2(\Omega), |\alpha| \le s\}.$$

With the norm

$$||u||_{\Xi^s(\Omega)} = \sum_{|\alpha| \le s} ||\rho^{|\alpha|} D^\alpha u||_{L^2(\Omega)},$$

the space $\Xi^s(\Omega)$ is a Hilbert space, and

$$\Xi^0(\Omega) = L^2(\Omega), \quad H^s(\Omega) \subset \Xi^s(\Omega) \subset L^2(\Omega), \quad \text{and} \quad \mathcal{D}(\Omega) \text{ is dense in } \Xi^s(\Omega).$$

Now, for a positive noninteger real $s = k+\theta$ with k the integer part of s and $0 < \theta < 1$, the space Ξ^s is, as in the case of the spaces H^s, the intermediate space

$$\Xi^s(\Omega) = [\Xi^{k+1}(\Omega), \Xi^k(\Omega)]_{1-\theta}.$$

Finally, for negative real values $-s$, $s > 0$, the space $\Xi^{-s}(\Omega)$ is the dual space of $\Xi^s(\Omega)$, $(\Xi^s(\Omega))'$.

THEOREM 2.3. *The set $\{y(v)_{|\omega} : v \in L^2(\Gamma)\}$ is dense, using the norm of $H^{1/2}(\omega)$, in $\{y \in H^{1/2}(\omega) : Ay = f \text{ in } \omega\}$, where $y(v) \in H^{1/2}(\Omega)$ is the solution of the Dirichlet problem* (2.11) *for a given $v \in L^2(\Gamma)$.*

Proof. Let us consider $y \in H^{1/2}(\omega)$ such that $Ay = f$ in ω, and a real number $\varepsilon > 0$. We denote the traces of y on γ by $y = g \in L^2(\gamma)$ and $\frac{\partial y}{\partial n_A(\omega)} = h \in H^{-1}(\gamma)$. From the previous lemma, it follows that there exists $v_\varepsilon \in L^2(\Gamma)$ such that the solution $z(v_\varepsilon) \in H^{1/2}(\Omega - \bar{\omega})$ of problem (2.18) satisfies

$$\left| \frac{\partial z(v_\varepsilon)}{\partial n_A(\Omega - \bar{\omega})} + h \right|_{H^{-1}(\gamma)} < \varepsilon.$$

Let $y(v_\varepsilon)$ be the solution of the Dirichlet problem (2.11) corresponding to v_ε, and let us define

$$y_\varepsilon = \begin{cases} y & \text{on } \omega, \\ z(v_\varepsilon) & \text{on } \Omega - \bar{\omega}. \end{cases}$$

Then $(y(v_\varepsilon) - y_\varepsilon) \in H^{1/2}(\Omega)$ and satisfies in the sense of distributions the equation

$$A(y(v_\varepsilon) - y_\varepsilon) = \tfrac{\partial z(v_\varepsilon)}{\partial n_A(\Omega - \bar{\omega})} + h \quad \text{in } \Omega$$

and the boundary conditions

$$y(v_\varepsilon) - y_\varepsilon = 0 \quad \text{on } \Gamma.$$

Consider, as in Remark 2.1, a fixed domain D such that $\bar{\omega} \subset D \subset \bar{D} \subset \Omega$. Then, for any $\psi \in \mathcal{D}(\Omega)$, we have $\int_\Omega A(y(v_\varepsilon) - y_\varepsilon)\psi = \int_\gamma (\frac{\partial z(v_\varepsilon)}{\partial n_A(\Omega - \bar{\omega})} + h)\psi \le |\frac{\partial z(v_\varepsilon)}{\partial n_A(\Omega - \bar{\omega})} + h|_{H^{-1}(\gamma)} |\psi|_{H^1(\gamma)} \le C(D)|\psi|_{H^{3/2}(D)}\varepsilon \le C(D)|\psi|_{\Xi^{3/2}(\Omega)}\varepsilon$, where $C(D)$ depends only on the domain D. Therefore,

$$|A(y(v_\varepsilon) - y_\varepsilon)|_{\Xi^{-3/2}(\Omega)} \le C(D)\varepsilon.$$

Taking into account the continuity of the solution on the data (see Lions and Magenes [27, Chap. 2, section 7.3, Theorem 7.4]), we get

$$|y(v_\varepsilon) - y_\varepsilon|_{H^{1/2}(\Omega)} \leq C(D)\varepsilon. \qquad \square$$

Below, the controllability of the solutions of the Dirichlet and the Neumann problems (given by (2.1), (2.2), and (2.1), (2.3), respectively) in ω by Neumann problems in Ω is discussed.

Now as a set of controls we can take the space

$$\mathcal{U} = H^{-1}(\Gamma), \tag{2.19}$$

and for a $v \in H^{-1}(\Gamma)$, the state of the system is the solution $y(v) \in H^{1/2}(\Omega)$ of the problem

$$\begin{aligned} Ay(v) &= f \quad \text{in} \quad \Omega, \\ \tfrac{\partial y(v)}{\partial n_A(\Omega)} &= v \quad \text{on} \quad \Gamma. \end{aligned} \tag{2.20}$$

We remark that

$$\begin{aligned} &i : \{y(v) \in H^{1/2}(\Omega) : v \in L^2(\Gamma),\ y(v) \text{ solution of problem (2.11)}\} \to \\ &\quad \{y(w) \in H^{1/2}(\Omega) : w \in H^{-1}(\Gamma),\ y(w) \text{ solution of problem (2.20)}\}, \\ &i(y(v)) = y(w) \Leftrightarrow y(v) = y(w) \text{ in } \Omega \end{aligned} \tag{2.21}$$

establish a bijective correspondence. Consequently, Proposition 2.1 also holds if the space of controls there is changed to $H^{-1}(\Gamma)$ and the states $y(v)$ of the system are solutions of problem (2.20). Theorem 2.3 in this case becomes the following theorem.

THEOREM 2.4. *The set* $\{y(v)_{|\omega} : v \in H^{-1}(\Gamma)\}$ *is dense, using the norm of* $H^{1/2}(\omega)$, *in* $\{y \in H^{1/2}(\omega) : Ay = f \text{ in } \omega\}$, *where* $y(v) \in H^{1/2}(\Omega)$ *is a solution of the Neumann problem* (2.20) *for a given* $v \in H^{-1}(\Gamma)$.

3. Controllability with finite dimensional spaces. Let $\{U_\lambda\}_\lambda$ be a family of finite dimensional subspaces of the space $L^2(\Gamma)$ such that, given (2.10) as a space of controls with the Dirichlet problems, we have

$$\bigcup_\lambda U_\lambda \text{ is dense in } \mathcal{U} = L^2(\Gamma). \tag{3.1}$$

For a $v \in L^2(\Gamma)$ we consider the solution $y'(v) \in H^{1/2}(\Omega)$ of the problem

$$\begin{aligned} Ay'(v) &= 0 \quad \text{in } \Omega, \\ y'(v) &= v \quad \text{on } \Gamma. \end{aligned} \tag{3.2}$$

We fix a U_λ. The cost functions J defined by (2.13) and (2.15) are differentiable and convex. Consequently, an optimal control

$$u_\lambda \in U_\lambda : J(u_\lambda) = \inf_{v \in U_\lambda} J(v) \tag{3.3}$$

exists if and only if it is a solution of the equation

$$u_\lambda \in U_\lambda : (y(u_\lambda), y'(v))_{L^2(\gamma)} = (g_\gamma, y'(v))_{L^2(\gamma)} \text{ for any } v \in U_\lambda, \tag{3.4}$$

when the control function is (2.13), and

$$u_\lambda \in U_\lambda : \left(\frac{\partial y(u_\lambda)}{\partial n_A(\omega)}, \frac{\partial y'(v)}{\partial n_A(\omega)}\right)_{H^{-1}(\gamma)} = \left(h_\gamma, \frac{\partial y'(v)}{\partial n_A(\omega)}\right)_{H^{-1}(\gamma)} \text{ for any } v \in U_\lambda, \tag{3.5}$$

when the control function is (2.15). Above, $y(u_\lambda)$ is the solution of problem (2.11) corresponding to u_λ, and $y'(v)$ is the solution of problem (3.2) corresponding to v. If $y_f \in H^2(\Omega)$ is the solution of the problem

$$\begin{array}{ll} Ay_f = f & \text{in } \Omega, \\ y_f = 0 & \text{on } \Gamma, \end{array} \tag{3.6}$$

then, for a $v \in L^2(\Gamma)$, we have

$$y(v) = y'(v) + y_f, \tag{3.7}$$

where $y(v)$ and $y'(v)$ are the solutions of problems (2.11) and (3.2), respectively. Therefore, we can rewrite problems (3.4) and (3.5) as

$$u_\lambda \in U_\lambda : (y'(u_\lambda), y'(v))_{L^2(\gamma)} = (g_\gamma - y_f, y'(v))_{L^2(\gamma)} \tag{3.8}$$

for any $v \in U_\lambda$, and

$$u_\lambda \in U_\lambda : \left(\frac{\partial y'(u_\lambda)}{\partial n_A(\omega)}, \frac{\partial y'(v)}{\partial n_A(\omega)}\right)_{H^{-1}(\gamma)} = \left(h_\gamma - \frac{\partial y_f}{\partial n_A(\omega)}, \frac{\partial y'(v)}{\partial n_A(\omega)}\right)_{H^{-1}(\gamma)} \tag{3.9}$$

for any $v \in U_\lambda$, respectively. Next, we prove the following lemma.

LEMMA 3.1. *For a fixed λ, let $\varphi_1, \ldots, \varphi_{n_\lambda}$, $n_\lambda \in \mathbf{N}$, be a basis of U_λ, and let $y'(\varphi_i)$ be the solution of problem* (3.2) *for $v = \varphi_i$, $i = 1, \ldots, n_\lambda$. Then $\{y'(\varphi_1)_{|\gamma}, \ldots, y'(\varphi_{n_\lambda})_{|\gamma}\}$ and $\{\frac{\partial y'(\varphi_1)}{\partial n_A(\omega)}|\gamma, \ldots, \frac{\partial y'(\varphi_{n_\lambda})}{\partial n_A(\omega)}|\gamma\}$ are linearly independent sets.*

Proof. From Remark 2.1, we have $y'(v) \in H^2(D)$ for any domain D which satisfies $\bar{\omega} \subset D \subset \bar{D} \subset \Omega$, and, consequently, $y'(v) \in H^{3/2}(\gamma)$ for any $v \in L^2(\Gamma)$. Assume that for $\xi_1, \ldots, \xi_{n_\lambda} \in \mathbf{R}$ we have $\xi_1 y'(\varphi_1) + \cdots + \xi_{n_\lambda} y'(\varphi_{n_\lambda}) = 0$ on γ. Then

$$y'(\xi_1\varphi_1 + \cdots + \xi_{n_\lambda}\varphi_{n_\lambda}) = 0 \quad \text{on } \gamma, \tag{3.10}$$

and therefore, $y'(\xi_1\varphi_1 + \cdots + \xi_{n_\lambda}\varphi_{n_\lambda}) = 0$ on ω. This implies that

$$\frac{\partial y'(\xi_1\varphi_1 + \cdots + \xi_{n_\lambda}\varphi_{n_\lambda})}{\partial n_A(\Omega - \bar{\omega})} = 0 \quad \text{on } \gamma. \tag{3.11}$$

From (3.10) and (3.11), we get $y'(\xi_1\varphi_1 + \cdots + \xi_{n_\lambda}\varphi_{n_\lambda}) = 0$ on $\Omega - \bar{\omega}$, and therefore, $\xi_1\varphi_1 + \cdots + \xi_{n_\lambda}\varphi_{n_\lambda} = y'(\xi_1\varphi_1 + \cdots + \xi_{n_\lambda}\varphi_{n_\lambda}) = 0$ on Γ, or $\xi_1 = \cdots = \xi_{n_\lambda} = 0$. The second part of the statement can be proved using similar arguments. □

The following proposition proves the existence and uniqueness of the optimal control when the states of the system are the solutions of the Dirichlet problems.

PROPOSITION 3.2. *Let us consider a fixed U_λ. Then problems* (3.8) *and* (3.9) *have unique solutions. Consequently, if the boundary conditions of Dirichlet problems* (2.11) *lie in the finite dimensional space U_λ, then there exists a unique optimal control of problem* (3.3) *corresponding to either the Dirichlet problem* (2.1), (2.2) *or the Neumann problem* (2.1), (2.3).

Proof. For a given λ, let V_λ denote the subspace of $L^2(\gamma)$ generated by $\{y'(\varphi_i)_{|\gamma}\}_{1 \le i \le n_\lambda}$, where $\{\varphi_i\}_{1 \le i \le n_\lambda}$ is a basis of U_λ, and $y'(\varphi_i)$ is the solution of problem (3.2) with $v = \varphi_i$. Since the norms $|\xi_1\varphi_1 + \cdots + \xi_{n_\lambda}\varphi_{n_\lambda}|_{L^2(\Gamma)}$ in U_λ, and $|\xi_1 y'(\varphi_1) + \cdots + \xi_{n_\lambda} y'(\varphi_{n_\lambda})|_{L^2(\gamma)}$ in V_λ are equivalent to the norm $(\xi_1^2 + \cdots + \xi_{n_\lambda}^2)^{1/2}$, the above lemma then implies that there exist two positive constants c and C such that

$$c|v|_{L^2(\Gamma)} \le |y'(v)|_{L^2(\gamma)} \le C|v|_{L^2(\Gamma)} \text{ for any } v \in U_\lambda.$$

Consequently, from the Lax–Milgram lemma we get that (3.8) has a unique solution. A similar reasoning proves that (3.9) also has a unique solution. This time we use the norm equivalence

$$c|v|_{L^2(\Gamma)} \leq \left| \frac{\partial y'(v)}{\partial n_A(\Omega - \bar{\omega})} \right|_{H^{-1}(\gamma)} \leq C|v|_{L^2(\Gamma)} \text{ for any } v \in U_\lambda$$

in the Lax–Milgram lemma. □

The following theorem proves the controllability of the solutions of the Dirichlet and Neumann problems in ω by the solutions of the Dirichlet problems in Ω.

THEOREM 3.3. *Let $\{U_\lambda\}_\lambda$ be a family of finite dimensional spaces satisfying* (3.1). *We associate the solution y of the Dirichlet problem* (2.1), (2.2) *in ω with problem* (3.3), *in which the cost function is given by* (2.13). *Also, the solution y of the Neumann problem* (2.1), (2.3) *is associated with problem* (3.3), *in which the cost function is given by* (2.15). *In both cases, there exists a positive constant C, and for any given $\varepsilon > 0$ there exists U_{λ_ε} such that*

$$|y(u_{\lambda_\varepsilon})_{|\omega} - y|_{H^{1/2}(\omega)} < C\varepsilon,$$

where $u_{\lambda_\varepsilon} \in U_{\lambda_\varepsilon}$ is the optimal control of the corresponding problem (3.3) *with $\lambda = \lambda_\varepsilon$, and $y(u_{\lambda_\varepsilon})$ is the solution of problem* (2.11) *with $v = u_{\lambda_\varepsilon}$.*

Proof. Let us consider an $\varepsilon > 0$ and $y \in H^{1/2}(\omega)$ as the solution of problem (2.1), (2.2). From Theorem 2.3, there exists $v_\varepsilon \in L^2(\Gamma)$ such that $y(v_\varepsilon) \in H^{1/2}(\Omega)$, the solution of problem (2.11) with $v = v_\varepsilon$, satisfies $|y - y(v_\varepsilon)_{|\omega}|_{H^{1/2}(\omega)} < \varepsilon$. Consequently, there exists a constant C_1 such that

$$|g_\gamma - y(v_\varepsilon)|_{L^2(\gamma)} < C_1\varepsilon. \tag{3.12}$$

Since $\cup_\lambda U_\lambda$ is dense in $L^2(\Gamma)$, there exist λ_ε and $v_{\lambda_\varepsilon} \in U_{\lambda_\varepsilon}$ such that $|v_\varepsilon - v_{\lambda_\varepsilon}|_{L^2(\Gamma)} < \varepsilon$, and then there exists a positive constant C_2 such that

$$|y(v_\varepsilon) - y(v_{\lambda_\varepsilon})|_{L^2(\gamma)} < C_2\varepsilon. \tag{3.13}$$

From (3.12) and (3.13) we get

$$|g_\gamma - y(v_{\lambda_\varepsilon})|_{L^2(\gamma)} < C_3\varepsilon$$

and, consequently,

$$|g_\gamma - y(u_{\lambda_\varepsilon})|_{L^2(\gamma)} < C_4\varepsilon,$$

where $u_{\lambda_\varepsilon} \in L^2(\Gamma)$ is the unique optimal control of problem (3.3) on U_{λ_ε} with the cost function given by (2.13). Therefore,

$$|y(u_{\lambda_\varepsilon})_{|\omega} - y|_{H^{1/2}(\omega)} < C\varepsilon.$$

A similar reasoning can be made for the solution $y \in H^{1/2}(\omega)$ of problem (2.1), (2.3). □

Using the basis $\varphi_1, \ldots, \varphi_{n_\lambda}$ of the space U_λ, we define the matrix

$$\Pi_\lambda = ((y'(\varphi_i), y'(\varphi_j))_{L^2(\gamma)})_{1 \leq i,j \leq n_\lambda} \tag{3.14}$$

and the vector

$$l_\lambda = ((g_\gamma - y_f, y'(\varphi_i))_{L^2(\gamma)})_{1 \leq i \leq n_\lambda}. \tag{3.15}$$

Then problem (3.8) can be written as

$$\xi_\lambda = (\xi_{\lambda,1}, \dots, \xi_{\lambda,n_\lambda}) \in \mathbf{R}^{n_\lambda} : \Pi_\lambda \xi_\lambda = l_\lambda. \tag{3.16}$$

Consequently, using Theorem 3.3, the solution y of problem (2.1), (2.2) can be obtained within any prescribed error by setting the restriction to ω of

$$y(u_\lambda) = \xi_{\lambda,1} y'(\varphi_1) + \cdots + \xi_{\lambda,n_\lambda} y'(\varphi_{n_\lambda}) + y_f, \tag{3.17}$$

where $\xi_\lambda = (\xi_{\lambda,1}, \dots, \xi_{\lambda,n_\lambda})$ is the solution of algebraic system (3.16). Above, y_f is the solution of problem (3.6), and $y'(\varphi_i)$ are the solutions of problems (3.2) with $v = \varphi_i$, $i = 1, \dots, n_\lambda$.

An algebraic system (3.16) is also obtained in the case of problem (3.9). This time the matrix of the system is given by

$$\Pi_\lambda = \left(\left(\frac{\partial y'(\varphi_i)}{\partial n_A(\omega)}, \frac{\partial y'(\varphi_j)}{\partial n_A(\omega)} \right)_{H^{-1}(\gamma)} \right)_{1 \le i,j \le n_\lambda}, \tag{3.18}$$

and the free term is

$$l_\lambda = \left(\left(h_\gamma - \frac{\partial y_f}{\partial n_A(\omega)}, \frac{\partial y'(\varphi_i)}{\partial n_A(\omega)} \right)_{H^{-1}(\gamma)} \right)_{1 \le i \le n_\lambda}. \tag{3.19}$$

Therefore, using Theorem 3.3, the solution y of problem (2.1), (2.3) can be estimated by (3.17). Also, y_f is the solution of problem (3.6), and $y'(\varphi_i)$ are the solutions of problems (3.2) with $v = \varphi_i$, $i = 1, \dots, n_\lambda$.

The case of the controllability with finite dimensional optimal controls for states of the system given by the solution of a Neumann problem is treated in a similar way. As in the previous section, the space of the controls is $\mathcal{U}$, given in (2.19), and the state of the system $y(v) \in H^{1/2}(\Omega)$ is given by the solution of Neumann problem (2.20) for a $v \in H^{-1}(\Gamma)$.

Let $\{U_\lambda\}_\lambda$ be a family of finite dimensional subspaces of the space $H^{-1}(\Gamma)$ such that

$$\bigcup_\lambda U_\lambda \text{ is dense in } \mathcal{U} = H^{-1}(\Gamma). \tag{3.20}$$

This time, the function $y'(v) \in H^{1/2}(\Omega)$ appearing in (3.4), (3.5), (3.8), and (3.9) is the solution of the problem

$$\begin{array}{ll} Ay'(v) = 0 & \text{in } \Omega, \\ \frac{\partial y'(v)}{\partial n_A(\Omega)} = v & \text{on } \Gamma \end{array} \tag{3.21}$$

for a $v \in H^{-1}(\Gamma)$. Also, $y_f \in H^2(\Omega)$ appearing in (3.7), (3.8), and (3.9) is the solution of the problem

$$\begin{array}{ll} Ay_f = f & \text{in } \Omega, \\ \frac{\partial y_f}{\partial n_A(\Omega)} = 0 & \text{on } \Gamma. \end{array} \tag{3.22}$$

With these changes, Lemma 3.1 also holds in this case, and the proof of the following proposition is similar to that of Proposition 3.2.

PROPOSITION 3.4. *For a given U_λ, the problems* (3.8) *and* (3.9) *have unique solutions. Consequently, if the boundary conditions of Neumann problems* (2.20) *lie in the finite dimensional space U_λ, then there exists a unique optimal control of problem* (3.3)*, corresponding to either Dirichlet problem* (2.1), (2.2) *or Neumann problem* (2.1), (2.3).

A proof similar to that given for Theorem 3.3 can also be given for the following theorem.

THEOREM 3.5. *Let $\{U_\lambda\}_\lambda$ be a family of finite dimensional spaces satisfying* (3.20). *We associate the solution $y \in H^{1/2}(\omega)$ of problem* (2.1), (2.2) *with problem* (3.3), *in which the cost function is given by* (2.13). *Also, the solution y of problem* (2.1), (2.3) *is associated with problem* (3.3), *in which the cost function is given by* (2.15). *In both cases, there exists a positive constant C, and for any given $\varepsilon > 0$ there exists λ_ε such that*

$$|y(u_{\lambda_\varepsilon})_{|\omega} - y|_{H^{1/2}(\omega)} < C\varepsilon,$$

where $u_{\lambda_\varepsilon} \in U_{\lambda_\varepsilon}$ is the optimal control of the corresponding problem (3.3) *with $\lambda = \lambda_\varepsilon$, and $y(u_{\lambda_\varepsilon})$ is the solution of problem* (2.20) *with $v = u_{\lambda_\varepsilon}$.*

Evidently, in the case of the controllability with solutions of Neumann problem (2.20) we can also write algebraic systems (3.16) using a basis $\varphi_1, \ldots, \varphi_{n_\lambda}$ of a given subspace U_λ of the space $\mathcal{U} = H^{-1}(\Gamma)$. As in the case of the controllability with solutions of the Dirichlet problem (2.11), these algebraic systems have unique solutions.

Theorems 3.3 and 3.5 prove the convergence of the embedding method associated with the optimal boundary control. An error analysis would be desirable, but it would go beyond the scope of this paper.

Remark 3.1. We have defined y_f as a solution of problems (3.6) or (3.22) in order to have $y(v) = y'(v) + y_f$ or $\frac{\partial y(v)}{\partial n_A(\Omega)} = \frac{\partial y'(v)}{\partial n_A(\Omega)} + \frac{\partial y_f}{\partial n_A(\Omega)}$, respectively, on the boundary Γ. In fact, we can replace $y(v)$ by $y'(v) + y_f$ in the cost functions (2.13) and (2.15) with $y_f \in H^2(\Omega)$ satisfying only

$$Ay_f = f \text{ in } \Omega, \tag{3.23}$$

and the results obtained in this section still hold.

Indeed, the two sets $\{y(v) = y'(v) + y_f \in H^{1/2}(\Omega) : v \in L^2(\Gamma)\}$ corresponding to y_f given by (3.23) and (3.6), $y'(v)$ being the solution of (3.2), are identical to the set $\{y(v) \in H^{1/2}(\Omega) : v \in L^2(\Gamma)\}$, $y(v)$ being the solution of (2.11). Also, the two sets $\{y(v) = y'(v) + y_f \in H^{1/2}(\Omega) : v \in H^{-1}(\Gamma)\}$ corresponding to y_f given by (3.23) and (3.22), $y'(v)$ being the solution of (3.21), are identical to the set $\{y(v) \in H^{1/2}(\Omega) : v \in H^{-1}(\Gamma)\}$, $y(v)$ being the solution of (2.20).

4. Approximate observations in finite dimensional spaces. In solving problems (3.8), (3.9), we require an appropriate interpolation which makes use of the values of $y'(v)$ computed only at some points on the boundary γ. We show below that using these interpolations, i.e., observations in finite dimensional subspaces, we can obtain the approximate solutions of problems (2.1), (2.2) and (2.1), (2.3).

As in the previous sections, we first deal with the case when the states of the system are given by the Dirichlet problem (2.11). Let U_λ be a fixed finite dimensional subspace of $\mathcal{U} = L^2(\Gamma)$ with the basis $\varphi_1, \ldots, \varphi_{n_\lambda}$.

Let us assume that for problem (2.1), (2.2), we choose a family of finite dimensional spaces $\{H_\mu\}_\mu$ such that

$$\bigcup_\mu H_\mu \text{ is dense in } \mathcal{H} = L^2(\gamma). \tag{4.1}$$

Similarly, for problem (2.1), (2.3) we choose the finite dimensional spaces $\{H_\mu\}_\mu$ such that

$$\bigcup_\mu H_\mu \text{ is dense in } \mathcal{H} = H^{-1}(\gamma). \tag{4.2}$$

The subspace H_μ given in (4.1) and (4.2) is a subspace of $\mathcal{H}$ given in (2.12) and (2.14), respectively.

An appropriate choice of H_μ is made based on the problem to be solved as discussed above. For a given φ_i, $i = 1, \dots, n_\lambda$, we consider below the solution $y'(\varphi_i)$ of problem (3.2) corresponding to $v = \varphi_i$, and we approximate its trace on γ by $y'_{\mu,i}$. Also, the approximation of $\frac{\partial y'(\varphi_i)}{\partial n_A(\omega)}$ on γ is denoted by $\frac{\partial y'_{\mu,i}}{\partial n_A(\omega)}$.

Since the system (3.16) has a unique solution, the determinants of the matrices Π_λ given in (3.14) and (3.18) are nonzero. Consequently, if $|y'(\varphi_i) - y'_{\mu,i}|_{L^2(\gamma)}$ or $|\frac{\partial y'(\varphi_i)}{\partial n_A(\omega)} - \frac{\partial y'_{\mu,i}}{\partial n_A(\omega)}|_{H^{-1}(\gamma)}$ are small enough, then the matrices

$$\Pi_{\lambda\mu} = ((y'_{\mu,i}, y'_{\mu,j})_{L^2(\gamma)})_{1\le i,j\le n_\lambda} \tag{4.3}$$

and

$$\Pi_{\lambda\mu} = \left(\left(\frac{\partial y'_{\mu,i}}{\partial n_A(\omega)}, \frac{\partial y'_{\mu,j}}{\partial n_A(\omega)} \right)_{H^{-1}(\gamma)} \right)_{1\le i,j\le n_\lambda} \tag{4.4}$$

have nonzero determinants. In this case, each of the algebraic systems

$$\xi_{\lambda\mu} = (\xi_{\lambda\mu,1}, \dots, \xi_{\lambda\mu,n_\lambda}) \in \mathbf{R}^{n_\lambda} : \Pi_{\lambda\mu}\xi_{\lambda\mu} = l_{\lambda\mu} \tag{4.5}$$

has a unique solution. In this system, the free term is

$$l_{\lambda\mu} = ((g_{\gamma\mu} - y_{f\mu}, y'_{\mu,i})_{L^2(\gamma)})_{1\le i\le n_\lambda} \tag{4.6}$$

if the matrix $\Pi_{\lambda\mu}$ is given by (4.3) and

$$l_{\lambda\mu} = \left(\left(h_{\gamma\mu} - \frac{\partial y_{f\mu}}{\partial n_A(\omega)}, \frac{\partial y'_{\mu,i}}{\partial n_A(\omega)} \right)_{H^{-1}(\gamma)} \right)_{1\le i\le n_\lambda} \tag{4.7}$$

if the matrix $\Pi_{\lambda\mu}$ is given by (4.4). Above, we have denoted by $g_{\gamma\mu}$ and $h_{\gamma\mu}$ some approximations in H_μ of g_γ and h_γ, respectively. Also, $y_{f\mu}$ and $\frac{\partial y_{f\mu}}{\partial n_A(\omega)}$ are some approximations of y_f and $\frac{\partial y_f}{\partial n_A(\omega)}$ in the corresponding H_μ of $L^2(\gamma)$ and $H^{-1}(\gamma)$, respectively, with $y_f \in H^2(\Omega)$ satisfying (3.23).

The solution y of problems (2.1), (2.2) and (2.1), (2.3) can be approximated with the restriction to ω of

$$y(u_{\lambda\mu}) = \xi_{\lambda\mu,1}y'(\varphi_1) + \cdots + \xi_{\lambda\mu,n_\lambda}y'(\varphi_{n_\lambda}) + y_f, \tag{4.8}$$

where $\xi_\lambda = (\xi_{\lambda\mu,1}, \dots, \xi_{\lambda\mu,n_\lambda})$ is the solution of appropriate algebraic system (4.5).

For a vector, $\xi = (\xi_1, \dots, \xi_{n_\lambda})$, we use the norm $|\xi| = \max_{1\le i\le n_\lambda} |\xi_i|$, and the corresponding matrix norm is denoted by $||\cdot||$. From (3.17) and (4.8) we have

$$|y(u_\lambda) - y(u_{\lambda\mu})|_{H^{1/2}(\omega)} \le C_\lambda |\xi_\lambda - \xi_{\lambda\mu}|, \tag{4.9}$$

where C_λ depends only on the basis in U_λ. From

$$||\Pi_\lambda^{-1} - \Pi_{\lambda\mu}^{-1}|| \leq \frac{||\Pi_\lambda^{-1}||||\Pi_\lambda - \Pi_{\lambda\mu}||}{1/||\Pi_\lambda^{-1}|| - ||\Pi_\lambda - \Pi_{\lambda\mu}||}$$

and algebraic systems (3.16) and (4.5), we have $\xi_\lambda = \Pi_\lambda^{-1} l_\lambda$ and $\xi_{\lambda\mu} = \Pi_{\lambda\mu}^{-1} l_{\lambda\mu}$ and we get that there exists $C_\lambda > 0$, depending on the basis in U_λ, such that

$$|\xi_\lambda - \xi_{\lambda\mu}| \leq C_\lambda \left(||\Pi_\lambda - \Pi_{\lambda\mu}|| + |l_\lambda - l_{\lambda\mu}|\right). \tag{4.10}$$

In the case of matrices (3.14) and (4.3) and the free terms (3.15) and (4.6), we have

$$\begin{aligned} ||\Pi_\lambda - \Pi_{\lambda\mu}|| &\leq C_\lambda \max_{1\leq i\leq n_\lambda} |y'(\varphi_i) - y'_{\mu,i}|_{L^2(\gamma)}, \\ |l_\lambda - l_{\lambda\mu}| &\leq C_\lambda \left(|g_\gamma - g_{\gamma\mu}|_{L^2(\gamma)} \right. \\ &\qquad \left. + |y_f - y_{f\mu}|_{L^2(\gamma)}\right) + C \max_{1\leq i\leq n_\lambda} |y'(\varphi_i) - y'_{\mu,i}|_{L^2(\gamma)}. \end{aligned} \tag{4.11}$$

Instead, if we take matrices (3.18) and (4.4) and the free terms (3.19) and (4.7), then we get

$$\begin{aligned} ||\Pi_\lambda - \Pi_{\lambda\mu}|| &\leq C_\lambda \max_{1\leq i\leq n_\lambda} \left|\frac{\partial y'(\varphi_i)}{\partial n_A(\omega)} - \frac{\partial y'_{\mu,i}}{\partial n_A(\omega)}\right|_{H^{-1}(\gamma)}, \\ |l_\lambda - l_{\lambda\mu}| &\leq C_\lambda \left(|h_\gamma - h_{\gamma\mu}|_{H^{-1}(\gamma)} + \left|\frac{\partial y_f}{\partial n_A(\omega)} - \frac{\partial y_{f\mu}}{\partial n_A(\omega)}\right|_{H^{-1}(\gamma)} \right) \\ &\qquad + C \max_{1\leq i\leq n_\lambda} |\frac{\partial y'(\varphi_i)}{\partial n_A(\omega)} - \frac{\partial y'_{\mu,i}}{\partial n_A(\omega)}|_{H^{-1}(\gamma)}, \end{aligned} \tag{4.12}$$

where C is a constant and C_λ depends on the basis in U_λ.

For states of the system given by the Neumann problem (2.20), U_λ is a subspace of $\mathcal{U} = H^{-1}(\Gamma)$. The material presented above for the case of the Dirichlet problems in Ω is applicable to the case of the Neumann problems in Ω, except for the difference that this time $y'(\varphi_i)$ are the solutions of problems (3.21) with $v = \varphi_i$, $i = 1, \ldots, n_\lambda$.

In both cases (i.e., when the control is affected via Dirichlet and Neumann problems), we obtain the following theorem from Theorems 3.3 and 3.5 and (4.9)–(4.12).

THEOREM 4.1. *Let $\{U_\lambda\}_\lambda$ be a family of finite dimensional spaces, either satisfying* (3.1) *if we consider problem* (2.11)*, or satisfying* (3.20) *if we consider problem* (2.20)*. Also, we associate problem* (2.1), (2.2) *or* (2.1), (2.3) *with a family of spaces $\{H_\mu\}_\mu$ satisfying* (4.1) *or* (4.2)*, respectively. Then, for any $\varepsilon > 0$, there exists λ_ε such that the following hold.*

(i) *If the space H_μ is taken such that $|y'(\varphi_i) - y'_{\mu,i}|_{L^2(\gamma)}$, $i = 1, \ldots, n_{\lambda_\varepsilon}$, are small enough, y is the solution of problem* (2.1)–(2.2)*, and $y(u_{\lambda_\varepsilon\mu})$ is given by* (4.8)*, in which $\xi_{\lambda\mu}$ is the solution of the algebraic system* (4.5) *with the matrix given by* (4.3) *and the free term given by* (4.6)*, then the algebraic system* (4.5) *has a unique solution and*

$$\begin{aligned} &|y(u_{\lambda_\varepsilon\mu})_{|\omega} - y|_{H^{1/2}(\omega)} < C\varepsilon \\ &+ C_{\lambda_\varepsilon} \left(|g_\gamma - g_{\gamma\mu}|_{L^2(\gamma)} + |y_f - y_{f\mu}|_{L^2(\gamma)} + \max_{1\leq i\leq n_\lambda} |y'(\varphi_i) - y'_{\mu,i}|_{L^2(\gamma)} \right). \end{aligned}$$

(ii) *If the space H_μ is taken such that $|\frac{\partial y'(\varphi_i)}{\partial n_A(\omega)} - \frac{\partial y'_{\mu,i}}{\partial n_A(\omega)}|_{H^{-1}(\gamma)}$, $i = 1, \ldots, n_{\lambda_\varepsilon}$, are small enough, y is the solution of problem* (2.1)–(2.3)*, and $y(u_{\lambda_\varepsilon\mu})$ is given by* (4.8) *in*

which $\xi_{\lambda\mu}$ is the solution of the algebraic system (4.5) *with the matrix given by* (4.4) *and the free term given by* (4.7)*, then the algebraic system* (4.5) *has a unique solution and*

$$\begin{aligned}&|y(u_{\lambda_\varepsilon\mu})_{|\omega} - y|_{H^{1/2}(\omega)} < C\varepsilon\\ &+ C_{\lambda_\varepsilon}\left(|h_\gamma - h_{\gamma\mu}|_{H^{-1}(\gamma)} + \left|\tfrac{\partial y_f}{\partial n_A(\omega)} - \tfrac{\partial y_{f\mu}}{\partial n_A(\omega)}\right|_{H^{-1}(\gamma)}\right.\\ &\left.+ \max_{1\le i\le n_\lambda}\left|\tfrac{\partial y'(\varphi_i)}{\partial n_A(\omega)} - \tfrac{\partial y'_{\mu,i}}{\partial n_A(\omega)}\right|_{H^{-1}(\gamma)}\right),\end{aligned}$$

where C is a constant and C_{λ_ε} depends on the basis of U_{λ_ε}.

Remark 4.1. Since the matrices $\Pi_{\lambda\mu}$ given by (4.3) and (4.4) are assumed to be nonsingular, it follows that $\{y'_{\mu,i}\}_{i=1,\dots,n_\lambda}$ and $\{\frac{\partial y'_{\mu,i}}{\partial n_A(\omega)}\}_{i=1,\dots,n_\lambda}$ are some linearly independent sets in $L^2(\gamma)$ and $H^{-1}(\gamma)$, respectively. Consequently, if m_μ is the dimension of the corresponding subspace H_μ, then $n_\lambda \le m_\mu$.

5. Exterior problems. In this section, we consider the domain $\omega \subset \mathbf{R}^N$ of problems (2.1), (2.2) and (2.1), (2.3) as the complement of the closure of a bounded domain, and it lies on only one side of its boundary. The same assumptions are made on the domain Ω of problems (2.11) and (2.20), and, evidently, $\omega \subset \Omega$. In order to retain continuity and to prove that the solutions of the problems in ω can be approximated by the solutions of problems in Ω, we have to specify the spaces in which the problems have solutions and also their correspondence with the trace spaces.

Since the domain $\Omega - \bar{\omega}$ is bounded, Lions's controllability theorem does not need to be extended to unbounded domains. Moreover, we see that the boundaries γ and Γ of the domains ω and Ω are bounded, and, consequently, we can use finite open covers of them (as for the bounded domains) to define the traces.

In order to avoid the use of the fractional spaces of the spaces in ω and Ω, we simply remark that if the controls in the Lions controllability theorem are taken in $H^{1/2}(\Gamma)$ instead of $L^2(\Gamma)$, then a similar proof of it gives the following.

The set $\{\frac{\partial z_0(v)}{\partial n_A(\Omega-\bar{\omega})} \in H^{-1/2}(\gamma) : v \in H^{1/2}(\Gamma)\}$ is dense in $H^{-1/2}(\gamma)$, where $z_0(v) \in H^1(\Omega - \bar{\omega})$ is the solution of the problem

$$\begin{aligned}&Az_0(v) = 0 \quad &&\text{in } \Omega - \bar{\omega},\\ &z_0(v) = v &&\text{on } \Gamma,\\ &z_0(v) = 0 &&\text{on } \gamma.\end{aligned}$$

Now we associate to the operator A the symmetric bilinear form

$$a(y,z) = \sum_{i,j=1}^{N}\int_\Omega a_{ij}\frac{\partial y}{\partial x_i}\frac{\partial z}{\partial x_j} + \int_\Omega a_0 yz \quad \text{for} \quad y,\ z \in H^1(\Omega),$$

which is continuous on $H^1(\Omega)\times H^1(\Omega)$. Evidently, a is also continuous on $H^1(\omega)\times H^1(\omega)$. Now if $f \in L^2(\omega)$, taking the boundary data $g_\gamma \in H^{1/2}(\gamma)$ and $h_\gamma \in H^{-1/2}(\gamma)$, then problems (2.1), (2.2) and (2.1), (2.3) can be written in the variational form

$$\begin{aligned}&y \in H^1(\omega) : a(y,z) = \textstyle\int_\omega fz \quad \text{for any} \quad z \in H_0^1(\omega),\\ &y = g_\gamma \text{ on } \gamma,\end{aligned} \tag{5.1}$$

and

$$y \in H^1(\omega) : a(y,z) = \int_\omega fz + \int_\gamma h_\gamma z \quad \text{for any} \quad z \in H^1(\omega), \tag{5.2}$$

respectively. Similar equations can also be written for problems (2.11) and (2.20).

Therefore, if there exists a constant $c_0 > 0$ such that $a_0 \geq c_0$ in Ω, then the bilinear form a is $H^1(\Omega)$-elliptic, i.e., there exists a constant $\alpha > 0$ such that $\alpha|y|^2_{H^1(\Omega)} \leq a(y,y)$ for any $y \in H^1(\Omega)$. It follows from the Lax–Milgram lemma that problems (2.11) and (2.20) have unique weak solutions in $H^1(\Omega)$. Naturally, problems (2.1), (2.2) and (2.1), (2.3) in ω also have unique weak solutions given by the solutions of problems (5.1) and (5.2), respectively.

We know that there exist an isomorphism and homeomorphism of $H^1(\Omega)/H^1_0(\Omega)$ onto $H^{1/2}(\Gamma)$ (see Theorem 7.53, p. 216, in [1], or Theorem 5.5, p. 99, and Theorem 5.7, p. 103, in [30]), i.e., there are two constants $k_1,\ k_2 > 0$ such that we have the following.

- For any $y \in H^1(\Omega)$, there exists $v \in H^{1/2}(\Gamma)$ such that $y = v$ on Γ and $\mid v \mid_{H^{1/2}(\Gamma)} \leq k_1 \mid y \mid_{H^1(\Omega)}$.
- For any $v \in H^{1/2}(\Gamma)$, there exists $y \in H^1(\Omega)$ such that $y = v$ on Γ and $\mid y \mid_{H^1(\Omega)} \leq k_2 \mid v \mid_{H^{1/2}(\Gamma)}$.

Using this correspondence, we can easily prove the continuous dependence of the solutions on data. For instance, for problems (2.1), (2.2) and (2.1), (2.3) we have

$$|y|_{H^1(\omega)} \leq C\{|f|_{L^2(\omega)} + |g_\gamma|_{H^{1/2}(\gamma)}\}$$

and

$$|y|_{H^1(\omega)} \leq C\{|f|_{L^2(\omega)} + |h_\gamma|_{H^{-1/2}(\gamma)}\},$$

respectively.

Therefore, if there exists a constant $c_0 > 0$ such that $a_0 \geq c_0$ in Ω, then we can proceed in the same manner and obtain similar results for the exterior problems to those obtained in the previous sections for the interior problems. Evidently, in this case we take

$$\mathcal{U} = H^{1/2}(\Gamma) \tag{5.3}$$

as a space of the controls for problem (2.11), in place of that given in (2.10), and the space of controls for problem (2.20) is taken as

$$\mathcal{U} = H^{-1/2}(\Gamma), \tag{5.4}$$

in place of the space given in (2.19).

If $a_0 = 0$ in Ω, the domain being unbounded, then the problems might not have solutions in the classical Sobolev spaces (see [11]), and we have to introduce the weighted spaces which take into account the particular behavior of the solutions at infinity.

For domains in $\mathbf{R}^2$, we use the weighted spaces introduced in [24], [25], specifically,

$$W^1(\Omega) = \{v \in \mathcal{D}'(\Omega) : (1+r^2)^{-1/2}(1+\log\sqrt{1+r^2})^{-1}v \in L^2(\Omega),\ \nabla v \in (L^2(\Omega))^2\},$$

where $\mathcal{D}'(\Omega)$ is the space of the distributions on Ω, and r denotes the distance from the origin. The norm on $W^1(\Omega)$ is given by

$$| v |_{W^1(\Omega)} = \left(| (1+r^2)^{-1/2}(1+\log\sqrt{1+r^2})^{-1} v |^2_{L^2(\Omega)} + | \nabla v |^2_{(L^2(\Omega))^2} \right)^{1/2}.$$

For domains in $\mathbf{R}^N$, $N \geq 3$, appropriate spaces, introduced in [21] and used in [20], [31], are

$$W^1(\Omega) = \{ v \in \mathcal{D}'(\Omega) : (1+r^2)^{-1/2} v \in L^2(\Omega),\ \nabla v \in (L^2(\Omega))^N \}$$

with the norm

$$| v |_{W^1(\Omega)} = \left(| (1+r^2)^{-1/2} v |^2_{L^2(\Omega)} + | \nabla v |^2_{(L^2(\Omega))^N} \right)^{1/2}.$$

We remark that the space $H^1(\Omega)$ is continuously embedded in $W^1(\Omega)$, and the two spaces coincide for the bounded domains. We use $W^1_0(\Omega)$ to denote the closure of $\mathcal{D}(\Omega)$ in $W^1(\Omega)$.

Concerning the space of the traces of the functions in $W^1(\Omega)$, we notice that, the boundary Γ being bounded, these traces lie in $H^{1/2}(\Gamma)$. This fact immediately follows from considering a bounded domain $D \subset \Omega$ such that $\Gamma \subset D$ and from taking into account that $W^1(D)$ and $H^1(D)$ are identical.

Assuming that

$$(1+r^2)^{1/2}(1+\log\sqrt{1+r^2}) f \in L^2(\Omega) \text{ if } N = 2,$$

$$(1+r^2)^{1/2} f \in L^2(\Omega) \text{ if } N \geq 3,$$

and using the spaces W^1 in place of the spaces H^1, we can rewrite the problems (5.1) and (5.2) and also similar equations for problems (2.11) and (2.20).

For $N = 2$, the bilinear form $a(y,z)$ generates on $W^1_0(\Omega)$ an equivalent norm with that induced by $W^1(\Omega)$ (see [24]). Also, the bilinear form $a(y,z)$ generates on $W^1(\Omega)/\mathbf{R}$ a norm which is equivalent to the standard norm.

For $N \geq 3$, the previously introduced norm on $W^1(\mathbf{R}^N)$ is equivalent to that generated by the bilinear form $a(.,.)$ (see [21]). Now if we extend the functions in $W^1_0(\Omega)$ with zero in $\mathbf{R}^N - \Omega$, we get that the bilinear form $a(y,z)$ also generates on $W^1_0(\Omega)$ a norm equivalent to that induced by $W^1(\Omega)$. Moreover, using the fact that the domain Ω is the complement of a bounded set, it can be proved that the bilinear form $a(y,z)$ generates in $W^1(\Omega)$ a norm equivalent to the above introduced norm.

Therefore, we can conclude that, in the case of $a_0 = 0$ on Ω, the exterior problems have unique solutions in the spaces W^1 if $N \geq 3$. If $N = 2$, the Dirichlet problems have unique solutions in W^1, and the Neumann problems have unique solutions in $W^1/\mathbf{R}$.

Using the fact that the spaces $W^1(D)$ and $H^1(D)$ coincide on the bounded domains D, the continuous embedding of $H^1(\Omega)$ in $W^1(\Omega)$, and the homeomorphism and isomorphism between $H^{1/2}(\Gamma)$ and $H^1(\Omega)/H^1_0(\Omega)$, we can easily prove that there exist a homeomorphism and isomorphism between $H^{1/2}(\Gamma)$ and $W^1(\Omega)/W^1_0(\Omega)$. Consequently, we get the following continuous dependence on the data of the solution y of problem (2.1), (2.2):

$$|y|_{W^1(\omega)} \leq C\{ |(1+r^2)^{1/2}(1+\log\sqrt{1+r^2}) f|_{L^2(\omega)} + |g_\gamma|_{H^{1/2}(\gamma)} \} \qquad \text{if } N = 2,$$

and

$$|y|_{W^1(\omega)} \le C\{|(1+r^2)^{1/2} f|_{L^2(\omega)} + |g_\gamma|_{H^{1/2}(\gamma)}\} \qquad \text{if } N \ge 3.$$

For the problem (2.1), (2.3), we have

$$\inf_{s\in\mathbf{R}} |y+s|_{W^1(\omega)} \le C\{|(1+r^2)^{1/2}(1+\log\sqrt{1+r^2}) f|_{L^2(\omega)} + |h_\gamma|_{H^{-1/2}(\gamma)}\} \qquad \text{if } N = 2,$$

and

$$|y|_{W^1(\omega)} \le C\{|(1+r^2)^{1/2} f|_{L^2(\omega)} + |h_\gamma|_{H^{-1/2}(\gamma)}\} \qquad \text{if } N \ge 3.$$

Therefore, we can prove in a manner similar to the previous sections that when $a_0 = 0$ on Ω and $N \ge 3$, the solutions of the Dirichlet and Neumann problems in ω can be approximated with solutions of both the Dirichlet and the Neumann problems in Ω. Naturally, the controls are taken in the appropriate space (5.3) or (5.4). If $a_0 = 0$ on Ω and $N = 2$, the solutions of the Dirichlet problems in ω can be approximated with solutions of the Dirichlet problem in Ω. The Neumann problems do not have unique solutions.

Since $y(v)$ and g_γ lie in $H^{1/2}(\gamma)$ in the case of problem (2.1), (2.2), and $\frac{\partial y(v)}{\partial n_A(\omega)}$ and h_γ lie in $H^{-1/2}(\gamma)$ when we solve (2.1), (2.3), the natural choices for the space of observations are

$$\mathcal{H} = H^{1/2}(\gamma) \tag{5.5}$$

and

$$\mathcal{H} = H^{-1/2}(\gamma), \tag{5.6}$$

respectively. Even if the convergence is assured for these spaces, their norms are numerically estimated with much difficulty. However, noticing that the inclusions $H^{1/2}(\gamma) \subset L^2(\gamma) \subset H^{-1/2}(\gamma) \subset H^{-1}(\gamma)$ are continuous, we can take the spaces of observations, as in the case of the bounded domains, given in (2.12) and (2.14). We mentioned earlier the need to avoid the use of the fractional Sobolev spaces for unbounded domains because of the lack of work on this subject (to the best of our knowledge), especially concerning the continuous dependence of the solution on the data of the problem. In the next section, we give a numerical example where the space of the controls is taken as for the bounded domains and the obtained results are accurate.

6. Numerical results. In this section, we choose some specific U_λ and H_μ. Hence we drop the subscripts λ and μ. First, we summarize the results obtained in the previous sections on the algebraic system we need to solve to obtain solutions, within a prescribed error, of problems (2.1), (2.2) or (2.1), (2.3), using the solutions of problems (2.11) or (2.20).

We recall that if, for both the bounded and unbounded domains, there exists a constant $c_0 > 0$ such that the coefficient a_0 of the operator A satisfies $a_0 \ge c_0$ in Ω, then the solutions of problems (2.1), (2.2) or (2.1), (2.3) can be approximated by the solutions of both problems (2.11) and (2.20). If $a_0 = 0$ in Ω, then the solutions of problems (2.1), (2.2) can be approximated by the solutions of problems (2.11) for both the bounded and the unbounded domains, and if also $N \ge 3$, then by the

solutions of problems (2.20) for unbounded domains only. If $a_0 = 0$ in Ω with the domains unbounded, then the solutions of problems (2.1), (2.3) can be obtained from the solutions of problems (2.11) and also from the solutions of (2.20) if $N \geq 3$.

Actually, we have to solve the algebraic system (4.5), which is rewritten as

$$\xi \in \mathbf{R}^n : \Pi\xi = l. \tag{6.1}$$

Some remarks on the computing of the elements of the matrix Π and the free term l are made below.

- Depending on the problem in Ω, we choose the space of controls $\mathcal{U}$ and a finite dimensional subspace of it, $U \subset \mathcal{U}$. Let $\varphi_1, \ldots, \varphi_n$, $n \in \mathbf{N}$, be the basis of U, and let $y'(\varphi_i)$, $i = 1, \ldots, n$ be the corresponding solutions of problems (3.2) or (3.21) if the problem in Ω is (2.11) or (2.20), respectively.
- If the problem in ω is (2.1), (2.2), then we calculate the values of $y'(\varphi_i)$, $i = 1, \ldots, n$, at the mesh points on γ. For the problem (2.1), (2.3) we calculate the values of $\frac{\partial y'(\varphi_i)}{\partial n_a(\omega)}$, $i = 1, \ldots, n$, at the mesh points on γ.
- Using the computed values of $y'(\varphi_i)$ or $\frac{\partial y'(\varphi_i)}{\partial n_a(\omega)}$, $i = 1, \ldots, n$, at the mesh points on γ, we compute the elements of the matrix Π which are some inner products either in $\mathcal{H} = L^2(\gamma)$ when we solve the problem (2.1), (2.2) or in $\mathcal{H} = H^{-1}(\gamma)$ when we solve the problem (2.1), (2.3). The finite dimensional subspace $H \subset \mathcal{H}$ depends on the numerical integration method that we use. We remark that the matrix Π is symmetric and full.
- The elements of the free term l are also some inner products in the space of observations $\mathcal{H}$. We use a solution y_f of (3.23) and the boundary data of the problem in ω (i.e., g_γ or h_γ if the problem is (2.1), (2.2) or (2.3), (2.1), respectively) in these inner products.
- For problem (2.1), (2.2), the matrix Π and the free term l are given by (4.3) and (4.6), respectively. Also, for problem (2.1), (2.3) the matrix Π and the free term l are given in (4.4) and (4.7), respectively. In these equations, y'_i and $\frac{\partial y'_i}{\partial n_A(\omega)}$ are some approximations in H of $y'(\varphi_i)$ and $\frac{\partial y'(\varphi_i)}{\partial n_A(\omega)}$, respectively. These approximations arise from the use of numerical integration on γ and numerical values of $y'(\varphi_i)$ and $\frac{\partial y'(\varphi_i}{\partial n_A(\omega)}$ at the mesh points on γ. These values can be found either by evaluating an algebraic expression or by interpolation. Indeed, when the finite element method or any other method is used with a mesh over Ω which does not fit with the boundary γ, the values of the functions y_f and $y'(\varphi_i)$, $i = 1, \ldots, n$ at some mesh points in γ are found by interpolation.

Finally, if $\xi = (\xi_1, \ldots, \xi_n)$ is the solution of algebraic system (6.1) and y is the solution of the problem we solve, then its approximation is the restriction to ω of

$$\xi_1 y'(\varphi_1) + \cdots + \xi_n y'(\varphi_n) + y_f. \tag{6.2}$$

We recall that the matrices Π_λ given in (3.14) and (3.18) are nonsingular, and therefore, each of the problems (3.16) has a unique solution. Also, algebraic systems (6.1) have unique solutions if their matrices and free terms are good approximations in H of the matrix and the free term of the algebraic systems (3.16), respectively. Also, from Remark 4.1 we must take $n \leq m$, n being the dimension of U and m the dimension of H. However, as we recall from section 2, the problem in infinite dimensional space may not have a solution. Consequently, for very large n, we might

obtain algebraic systems (3.16) that are almost singular. These algebraic systems can be solved by an iterative method such as the conjugate gradient method. However, we applied the Gauss elimination method in order to find out whether the algebraic system is singular or nonsingular. This is done by checking the diagonal elements during the elimination phase.

In the following two subsections, we give some numerical examples for both interior and exterior problems in which the solutions of the problems in Ω are found either directly by a formula, or by a method using a mesh over Ω.

6.1. Interior problems.

Example 6.1. The first numerical test refers to the Dirichlet problem

$$\begin{aligned} -\Delta y &= f \text{ in } \omega, \\ y &= g_\gamma \text{ on } \gamma, \end{aligned} \tag{6.3}$$

where $\omega \subset \mathbf{R}^2$ is a square centered at the origin with sides parallel to the axes and of length of 2 units. The approximate solution of this problem is given by the solution of the Dirichlet problem

$$\begin{aligned} -\Delta y(v) &= f \text{ in } \Omega, \\ y(v) &= v \text{ on } \Gamma, \end{aligned} \tag{6.4}$$

in which the domain Ω is the disc centered at the origin with radius equal to 2. The solutions of the homogeneous Dirichlet problems in Ω are found by the Poisson formula

$$y(v)(z) = \frac{1}{2\pi r} \int_{|\zeta|=r} v(\zeta) \frac{r^2 - |z|^2}{|z-\zeta|^2} dS_\zeta. \tag{6.5}$$

The circle Γ is discretized with n equidistant points, and $U \subset \mathcal{U} \equiv L^2(\Gamma)$ is taken as the space of the piecewise constant functions. Naturally, an element φ_i in the basis of H is a function defined on Γ which takes the value 1 between the nodes i and $i+1$ and vanishes in the rest of Γ. The square γ is also discretized with m equidistant points, and $H \subset \mathcal{H} \equiv L^2(\gamma)$ is taken as the space of the continuous piecewise linear functions. Evidently, the inclusions in (3.1) and (4.1) are dense because the union of the spaces (over some sequence of mesh size approaching zero) of continuous piecewise linear or piecewise constant functions is dense in L^2.

The values of the integrals in the Poisson formula at the points on γ are calculated using the numerical integration with three nodes. The integrals in the inner products in $L^2(\gamma)$ are calculated using an exact formula when H is the space of the continuous piecewise linear functions. In particular, if we have on γ two continuous piecewise linear functions y_1 and y_2 such that

$$\begin{aligned} y_1(x) &= m_1^k (x - x_k) + y_1^k, \\ y_2(x) &= m_2^k (x - x_k) + y_2^k \end{aligned} \tag{6.6}$$

for $x \in [x_k, x_{k+1}]$, $k = 1 \ldots, m$, then

$$\int_\gamma y_1 y_2 = h \sum_{k=1}^{m} \left[y_1^k y_2^k + \frac{h^2}{3} m_1^k m_2^k + \frac{h}{2} (m_1^k y_2^k + m_2^k y_1^k) \right], \tag{6.7}$$

where $h = x_{k+1} - x_k$ is the mesh size on γ.

TABLE 6.1
Relative errors for the interior Dirichlet problem.

n	$\mathrm{err_d}$	$\mathrm{err_b}$
80	.36692E-07	.15956E-06
72	.46271E-08	.41101E-07
60	.14682E-09	.25103E-08
45	.12475E-08	.54357E-08
40	.64352E-12	.11638E-07
36	.67121E-12	.11648E-06
30	.12371E-05	.33923E-05
24	.39543E-12	.19851E-04
18	.10609E-03	.43901E-03
12	.29916E-10	.54208E-02
10	.94618E-02	.17096E-01

All computations below have been performed in fifteen digit arithmetics (double precision).

In the first example, we choose the exact solution to be $u(x_1, x_2) = x_1^2 + x_2^2$. Hence $g_\gamma(x_1, x_2) = x_1^2 + x_2^2$, and $f = -4$. We have taken $y_f = 2x_1^2$ as a solution of the inhomogeneous equation in Ω. It has been compared with the computed one at 19 equidistant points on a diagonal of the square: (-1.4,-1.4),. . .,(0,0),. . .,(1.4,1.4). Below $\mathrm{err_d}$ denotes the maximum of the relative errors between the exact and the computed solutions at these 19 considered points in the domain ω. A similar error only on the boundary γ is denoted by $\mathrm{err_b}$.

Table 6.1 shows errors $\mathrm{err_d}$ and $\mathrm{err_b}$ against n, the number of the equidistant points on Γ which is the dimension of the finite dimensional space U. Recall that Γ is boundary of the embedding domain Ω. All these computations use a mesh size of 0.1 on γ. It corresponds to $m = 80$, the number of equidistant points on γ, which is the dimension of the finite dimensional space H. The smallest diagonal element during the Gauss elimination method is of the order 10^{-17} for $n = 80$ and $n = 72$, and of the order 10^{-14} for $n = 60$. It is greater than 10^{-10} for $n = 10, \ldots, 45$. We should mention that in the cases when $n > 60$, where the last pivot is very small, we notice an increase in error. In all these cases the error $\mathrm{err_b}$, which can be calculated for any example even when the exact solution is not known, is a good indicator of the computational accuracy.

In the above example, the right-hand side f of the equation in ω is given by an exact algebraic formula, and it was extended in Ω by the same formula. Moreover, we have had for this simple example an exact solution y_f of the inhomogeneous equation in Ω, which could be exactly evaluated at the mesh points of the boundary γ of the domain ω. Also, the solutions of the homogeneous problems in Ω, given by the above Poisson formula, could be evaluated directly at these mesh points. In the following example we study the effect of various extensions of f in Ω on the computed solutions in ω. Therefore, in this example, the solution of the problem in Ω could be computed only at some nodes of a regular mesh over Ω, and their values at the mesh points on γ are calculated by interpolation.

Example 6.2. This example concerns the Dirichlet problem

$$\begin{aligned} &\Delta y - \sigma^2 y = f \text{ in } \omega, \\ &y = g_\gamma \text{ on } \gamma, \end{aligned} \tag{6.8}$$

where $\omega \subset \mathbf{R}^2$ is bounded by the straight lines $x_1 = -\pi/2$, $x_1 = \pi/2$, and $x_2 = -1.5$ and the curve $y = 0.5 + \cos(x + \pi/2)$. We approximate the solution of this problem by a solution of the Dirichlet problem

$$\begin{aligned} &\Delta y(v) - \sigma^2 y(v) = f \text{ in } \Omega, \\ &y(v) = v \text{ on } \Gamma, \end{aligned} \tag{6.9}$$

in which the domain Ω is the disc centered at the origin with the radius of 2.3 (see Figure 6.1 (a)). We have taken $\sigma^2 = 0.75$ in numerical computations.

We approximate the functions f and v by the discrete Fourier transforms

$$\begin{aligned} f(r,\theta) &= \sum_{k=-n/2}^{n/2-1} f_k(r) e^{ik\theta}, \\ v(\theta) &= \sum_{k=-n/2}^{n/2-1} v_k e^{ik\theta}. \end{aligned} \tag{6.10}$$

Then the solution of problem (6.9),

$$y(v) = y_f + y'(v), \tag{6.11}$$

can also be written as a discrete Fourier transform

$$\begin{aligned} y_f(r,\theta) &= \sum_{k=-n/2}^{n/2-1} y_k(r) e^{ik\theta}, \\ y'(v)(r,\theta) &= \sum_{k=-n/2}^{n/2-1} y'_k(r) e^{ik\theta}, \end{aligned} \tag{6.12}$$

where the Fourier coefficients $y_k(r)$ and $y'_k(r)$ are given by

$$\begin{aligned} y_k(r) = &- \int_0^r \rho K_k(\sigma r) I_k(\sigma\rho) f_k(\rho) d\rho - \int_r^R \rho I_k(\sigma r) K_k(\sigma\rho) f_k(\rho) d\rho \\ &+ \frac{I_k(\sigma r)}{I_k(\sigma R)} \int_0^R \rho K_k(\sigma R) I_k(\sigma\rho) f_k(\rho) d\rho, \\ y'_k(r) = &\frac{I_k(\sigma r)}{I_k(\sigma R)} v_k. \end{aligned} \tag{6.13}$$

Above, R is the radius of the disc, and I_k and K_k are the modified Bessel functions of the first and second kinds, respectively. We recall that $y'(v)$ and y_f in (6.11) are the solutions of problems (3.2) and (3.6), respectively. A fast algorithm is proposed in [4], which, using (6.13) and the fast Fourier transforms, evaluates y_f and $y'(v)$ in (6.12) at the nodes of a mesh on the disc Ω with n equidistant nodes in tangential direction and l equidistant nodes in the radial direction.

It is worth noting from (6.10) that the finite dimensional space of controls U is the space of real periodic functions defined on $[0, 2\pi]$ which can be written as a Fourier transform with the terms $-n/2, \ldots, 0, \ldots, n/2 - 1$. On the other hand, we have $\mathcal{U} = L^2(\Gamma) = L^2(0, 2\pi)$, and since the functions in $L^2(0, 2\pi)$ can be approximated by discrete Fourier transforms, we get that (3.1) holds with U as the above finite dimensional spaces. Since the controls v are real functions, it follows from (6.10)

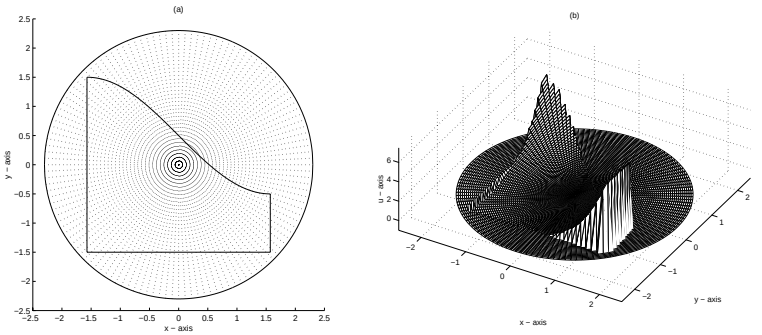


FIG. 6.1. (a) *Domains,* (b) *exact solution.*

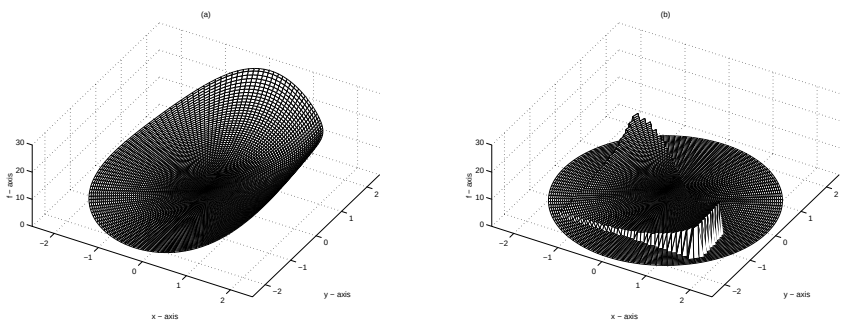


FIG. 6.2. *Extension of f by* (a) *the formula in the domain ω,* (b) *zero.*

that $v_i = \bar{v}_{-i}$ for $i = 1, \ldots, n/2 - 1$ and v_0 is real provided we choose $v_{-n/2} \in R$. Consequently, a basis of U is given by the functions: φ_0 which has the Fourier coefficient $v_0 = 1$, the other ones being zero, $\varphi_{-n/2}$ which has the Fourier coefficient $v_{-n/2} = 1$, the other ones being zero, and φ_j, $-n/2 + 1 \leq j \leq n/2 - 1$, $j \neq 0$, have the Fourier coefficients $v_j = 1 + i$, $v_{-j} = 1 - i$ with the rest being zero.

The boundary γ is discretized with m equidistant points, and, as in the previous example, H is taken to be the space of the piecewise linear functions. The integrals in the inner products in $L^2(\gamma)$ are calculated by the same formulae (6.7). We recall that the values of y_f and $y'(\varphi_)$ at the mesh points of the boundary γ were obtained by interpolation of function values at mesh points on Ω. Assuming that the point (r, θ) lies between the four mesh nodes (r_1, θ_1), (r_2, θ_1), (r_1, θ_2), (r_2, θ_2), we have linearly interpolated in radial direction first the values corresponding to (r_1, θ_1) and (r_2, θ_1), and then the values corresponding to (r_1, θ_2), (r_2, θ_2). Using the two obtained values, we have made a linear interpolation in the tangential direction.

For numerical purposes, we have taken $f(x_1, x_2) = (2 + x_1(1 - \sigma^2))e^{x_1} + (2 + x_2(1 - \sigma^2))e^{x_2}$ and $g_\gamma(x_1, x_2) = x_1 e^{x_1} + x_2 e^{x_2}$ in (6.8). Then problem (6.8) has the exact solution $y(x_1, x_2) = x_1 e^{x_1} + x_2 e^{x_2}$, which is shown in Figure 6.1 (b). In order to assess the effect of various extensions of the function f outside of ω on the numerical results, we have taken for this example only two types of extensions: (i) extending f using the above formula in ω; (ii) extending f by zero (see Figure 6.2).

Tables 6.2 through 6.5 show the arithmetic mean of the absolute errors between the exact and the computed solutions against various values of n (the number of the nodes in tangential direction, i.e., the number of nodes on Γ) and δ_r (the mesh size in radial direction), while keeping the number of mesh points on γ fixed at $m = 360$ for

TABLE 6.2
Errors on $\gamma - f$ extended with the formula in ω.

n/δ_r	0.1	0.05	0.02	0.01
8	0.15555E+00	0.15571E+00	0.15577E+00	0.15577E+00
16	0.25622E-01	0.25530E-01	0.25505E-01	0.25500E-01
32	0.58700E-02	0.55274E-02	0.55131E-02	0.55146E-02
64	0.26025E-02	0.13450E-02	0.12478E-02	0.12411E-02
128	0.12901E-02	0.56973E-03	0.36080E-03	0.35200E-03

TABLE 6.3
Errors in $\omega - f$ extended with the formula in ω.

n/δ_r	0.1	0.05	0.02	0.01
8	0.98198E-01	0.92264E-01	0.89501E-01	0.88875E-01
16	0.33058E-01	0.31403E-01	0.30967E-01	0.30862E-01
32	0.83707E-02	0.69124E-02	0.65851E-02	0.65232E-02
64	0.38456E-02	0.18422E-02	0.14402E-02	0.13976E-02
128	0.30019E-02	0.95631E-03	0.40010E-03	0.34864E-03

TABLE 6.4
Errors on $\gamma - f$ extended by zero.

n/δ_r	0.1	0.05	0.02	0.01
8	0.20670E+00	0.20546E+00	0.20347E+00	0.20331E+00
16	0.32825E-01	0.33941E-01	0.34529E-01	0.35906E-01
32	0.67604E-02	0.69137E-02	0.79452E-02	0.83573E-02
64	0.39507E-02	0.19624E-02	0.24754E-02	0.26836E-02
128	0.14346E-02	0.78505E-03	0.13167E-02	0.13784E-02

TABLE 6.5
Errors in $\omega - f$ extended by zero.

n/δ_r	0.1	0.05	0.02	0.01
8	0.15520E+00	0.15211E+00	0.15156E+00	0.15206E+00
16	0.30860E-01	0.27219E-01	0.25270E-01	0.25012E-01
32	0.72434E-02	0.54336E-02	0.52230E-02	0.51386E-02
64	0.40250E-02	0.19991E-02	0.15890E-02	0.15415E-02
128	0.33861E-02	0.15554E-02	0.10941E-02	0.10286E-02

all these computations. The results in Tables 6.2 and 6.3 have been obtained with the extension of f in Ω made with the formula in ω, and the results in Tables 6.4 and 6.5 have been obtained with the extension made by zero. In Tables 6.2 and 6.4, we show the errors computed on the boundary γ by taking the average over $m = 360$ boundary points. On the other hand, we show in Tables 6.3 and 6.5 the errors computed in the domain ω by taking the average over all mesh points in ω.

We notice in these tables that errors on the boundary γ are of the same order as in the domain ω, and the extension of the function f outside of ω with the formula in ω gives smaller errors than the extension by zero. It may be worth noting here that the errors for this example are higher than those for the previous example (see Table 6.1) because the values of y_f and $y(\varphi_i)$ on the boundary γ were found by interpolation in this example and by an exact algebraic expression in the previous example. Thus interpolation error is one of the possible sources of larger error in these tables for this example. Figure 6.3 shows absolute errors at the mesh nodes in the domain ω when $n = 128$ and $\delta_r = 0.01$ for the two extensions of f.

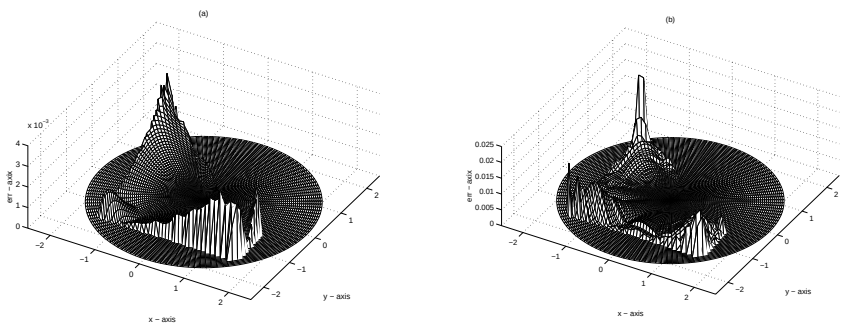


FIG. 6.3. *Errors in the domain when f is extended by* (a) *the formula in the domain* ω, (b) *zero.*

6.2. Exterior problems. Below, we show the performance of the method on two exterior problems.

Example 6.3. We solve the same problem as defined by (6.3) in Example 6.1 except that the domain ω is now the exterior of a square centered at the origin with sides parallel to the axes and of length of 2 units. For this problem, we consider exterior Dirichlet problem (6.4) with the embedding domain Ω as the exterior of a disc with its center at the origin and radius 0.99 unit.

Similar to Example 6.1, the solutions of the homogeneous Dirichlet problems in Ω are found by the Poisson formula

$$y(v)(z) = \frac{-1}{2\pi r} \int_{|\zeta|=r} v(\zeta) \frac{r^2 - |z|^2}{|z-\zeta|^2} dS_\zeta. \tag{6.14}$$

The spaces U, $\mathcal{U}$, H, and $\mathcal{H}$ are the same as in Example 6.1, and the integrals on the boundary γ use the same formula (6.7).

The problem in ω we have numerically solved has had $g_\gamma(x_1, x_2) = x_1 x_2$ and $f = 0$. Evidently, we take $y_f = 0$. In this case, we do not know the exact solution of the problem, but we recall from previous examples that the error on the boundary γ was very close to that in domain ω. Hence Table 6.6 shows the maximum relative errors between the exact prescribed data and the computed solutions on boundary γ against various values of n (the number of the nodes in tangential direction, i.e., number of nodes on Γ) while keeping the number of mesh points on γ fixed at $m = 120$ (corresponding to a mesh size of $1/15$ on γ) for all these computations.

We found that the smaller diagonal element during the Gauss elimination method is of the order 10^{-15} for $n = 120$ and of the order 10^{-14} for $n = 118$, and it is greater

TABLE 6.6
Errors obtained for the exterior Dirichlet problem.

n	$\mathrm{err_b}$
120	0.10995E-03
118	0.93472E-04
116	0.24253E-05
115	0.38082E-03
110	0.33003E-02
100	0.55797E-01
90	0.18828E+00
60	0.21087E+00
30	0.77558E+00

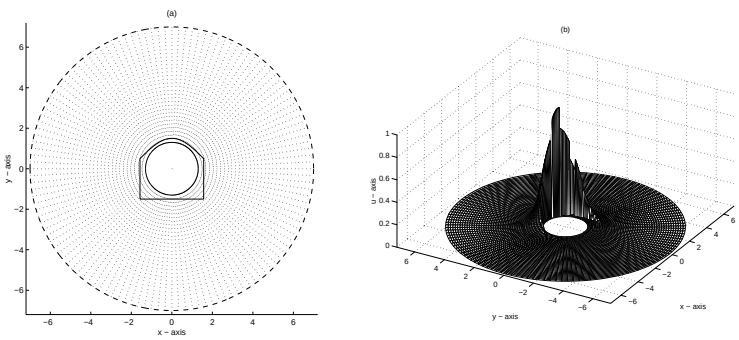


FIG. 6.4. (a) *Domains,* (b) *exact solution.*

than 10^{-12} for $n = 30, \ldots, 116$. We see in Table 6.6 that for $n > 116$, the error increases when the pivots in the Gauss elimination method become very small.

Example 6.4. Here we solve the same problem as defined by (6.8) in Example 6.2 except that the domain ω now is the open complement of the domain bounded by the straight lines $x_1 = -\pi/2$, $x_1 = \pi/2$, and $x_2 = -1.5$ and the curve $y = 0.5 + \cos(x)$. For this problem, the embedding domain Ω is taken to be the exterior of a disc with its center at the origin and radius 1.3 unit (see Figure 6.4, (a)).

We approximate the solution of this problem by a solution of the exterior Neumann problem

$$\begin{aligned} &\Delta y(v) - \sigma^2 y(v) = f \text{ in } \Omega, \\ &\tfrac{\partial y(v)}{\partial n_A(\Omega)} = v \text{ on } \Gamma, \end{aligned} \tag{6.15}$$

where Γ is the inner boundary of the embedding domain Ω. Similar to Example 6.2, we have taken $\sigma^2 = 0.75$ in numerical computations.

As before, functions f and v are approximated by the discrete Fourier transforms (6.10). Then the solution of problem (6.15) admits representation given by (6.11) and (6.12) except that the Fourier coefficients $y_k(r)$ and $y_k'(r)$ are now given by

$$\begin{aligned} y_k(r) = &- \int_R^r \rho K_k(\sigma r) I_k(\sigma\rho) f_k(\rho) d\rho - \int_r^\infty \rho I_k(\sigma r) K_k(\sigma\rho) f_k(\rho) d\rho \\ &- \frac{K_k(\sigma r)}{K_{k-1}(\sigma R) + K_{k+1}(\sigma R)} \int_R^\infty \rho [I_{k-1}(\sigma R) + I_{k+1}(\sigma R)] K_k(\sigma\rho) f_k(\rho) d\rho, \\ y_k'(r) = &\frac{K_k(\sigma r)}{K_{k-1}(\sigma R) + K_{k+1}(\sigma R)} \frac{2}{\sigma} v_k. \end{aligned} \tag{6.16}$$

Above, R is the radius of the disc whose complement is the domain Ω, and I_k and K_k are the modified Bessel functions of first and second kinds, respectively. In order to compute the solution of problem (6.15) at mesh points of the domain Ω with n equidistant nodes in the tangential direction and l equidistant nodes in the radial direction, we use the algorithm proposed in [4]. This algorithm uses (6.16) and the fast Fourier transforms to compute y_f and $y'(v)$ in (6.12). For numerical computations, the domain Ω is considered to be the annulus with the radii R and R_∞, where R_∞ is chosen very large so that its effect is minimal on the accuracy of the solutions.

The spaces U and H are the same as in Example 6.2. Also, the values of y_f and $y'(\varphi_)$ at the mesh points of the boundary γ were obtained by interpolating the values of the function at mesh points on Ω.

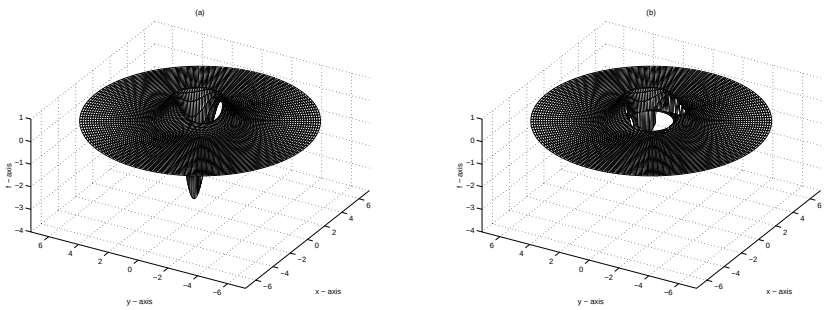


FIG. 6.5. *Extension of f by* (a) *the formula in the domain* ω, (b) *zero.*

For numerical purposes in this example, we have considered (6.8) with $f(x_1, x_2) = [4(x_1^2 + x_2^2 - x_1 - x_2) - 2 + \sigma^2]e^{-x_1^2 - x_2^2 + x_1 + x_2}$ and $g_\gamma(x_1, x_2) = e^{-x_1^2 - x_2^2 + x_1 + x_2}$. This problem has the exact solution $y(x_1, x_2) = e^{-x_1^2 - x_2^2 + x_1 + x_2}$ (it lies in $H^1(\omega)$ and satisfies the equation and the boundary conditions of problem (6.8)), which is plotted in Figure 6.4 (b).

Numerical computations show that $|y(r)| \leq 0.104E - 16$ for $r > 7$, where r is the distance of the point from the origin. Hence we have taken $R_\infty = 7$ in these computations. As in Example 6.2, we have extended f outside of ω in two different ways: (i) by the above formula, and (ii) by zero. These extensions are plotted in Figure 6.5. We have taken $m = 360$, the number of the mesh points on the boundary γ.

The error tables are similar to those in Example 6.2. Tables 6.7 and 6.8 correspond to the case when the extension of f in Ω is made with the formula in ω, and Tables 6.9 and 6.10 correspond to the extension made by zero. In Tables 6.7 and 6.9, we show the arithmetic mean of the absolute errors computed on the boundary γ by taking the average over $m = 360$ boundary points. On the other hand, we show in Tables 6.8 and 6.10 the errors computed in the domain ω by taking the average over all mesh points in ω. It is worth noting in these tables that, this time, the errors on the boundary γ are less than those in the domain ω, and the two extensions of f give solutions with

TABLE 6.7
Errors on $\gamma - f$ *extended with the formula in* ω.

n/δ_r	0.1	0.05	0.02	0.01
8	0.13247E-01	0.13231E-01	0.13233E-01	0.13233E-01
16	0.25712E-02	0.25628E-02	0.25500E-02	0.25496E-02
32	0.59286E-03	0.58076E-03	0.57869E-03	0.57859E-03
64	0.18186E-03	0.15977E-03	0.15536E-03	0.15462E-03
128	0.63343E-04	0.51301E-04	0.45571E-04	0.45775E-04

TABLE 6.8
Errors in $\omega - f$ *extended with the formula in* ω.

n/δ_r	0.1	0.05	0.02	0.01
8	0.29115E-02	0.28034E-02	0.26385E-02	0.26264E-02
16	0.11901E-02	0.10997E-02	0.10582E-02	0.10493E-02
32	0.66451E-03	0.62745E-03	0.61610E-03	0.61432E-03
64	0.56566E-03	0.54864E-03	0.54777E-03	0.54815E-03
128	0.60927E-03	0.53842E-03	0.53886E-03	0.53988E-03

TABLE 6.9
Errors on $\gamma - f$ extended by zero.

n/δ_r	0.1	0.05	0.02	0.01
8	0.13045E-01	0.13030E-01	0.13016E-01	0.13016E-01
16	0.26982E-02	0.26937E-02	0.27057E-02	0.26834E-02
32	0.61089E-03	0.61377E-03	0.63730E-03	0.64057E-03
64	0.17974E-03	0.15425E-03	0.16077E-03	0.16632E-03
128	0.63717E-04	0.52191E-04	0.53991E-04	0.57498E-04

TABLE 6.10
Errors in $\omega - f$ extended by zero.

n/δ_r	0.1	0.05	0.02	0.01
8	0.28435E-02	0.27346E-02	0.25635E-02	0.25483E-02
16	0.11929E-02	0.11024E-02	0.10533E-02	0.10487E-02
32	0.67268E-03	0.64000E-03	0.63015E-03	0.63105E-03
64	0.58007E-03	0.56682E-03	0.56861E-03	0.57090E-03
128	0.60658E-03	0.55784E-03	0.56152E-03	0.56343E-03

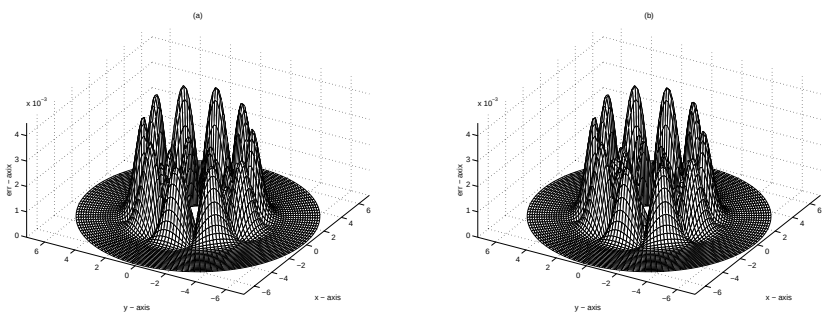


FIG. 6.6. *Errors in the domain when f is extended by* (a) *the formula in the domain ω,* (b) *zero.*

errors of the same order. In Figure 6.6, we have plotted the absolute error at the mesh nodes in the domain ω when $n = 128$ and $\delta_r = 0.01$ for these two extensions of f.

Acknowledgment. We sincerely thank the referees for their constructive criticism.

REFERENCES

[1] R. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] G. P. ASTRAKMANTSEV, *Methods of fictitious domains for a second order elliptic equation with natural boundary conditions*, USSR Computational Math. Math. Phys., 18 (1978), pp. 114–121.
[3] C. ATAMIAN, Q. V. DINH, R. GLOWINSKI, JIWEN HE, AND J. PÉRIAUX, *Control approach to fictitious-domain methods. Application to fluid dynamics and electro-magnetics*, in Proceedings of the Fourth International Symposium on Domain Decomposition Methods for Partial Differential Equations, R. Glowinski, Y. Kuznetsov, G. Meurant, J. Périaux, and O. B. Widlund, eds., SIAM, Philadelphia, 1991, pp. 275–309.
[4] L. BADEA AND P. DARIPA, *A Fast Algorithm for Two-Dimensional Elliptic Problems*, manuscript.
[5] C. BORGERS, *Domain embedding methods for the Stokes equations*, Numer. Math., 57 (1990), pp. 435–452.

[6] B. L. Buzbee, F. W. Dorr, J. A. George, and G. H. Golub, *The direct solution of the discrete Poisson equation on irregular regions*, SIAM J. Numer. Anal., 8 (1971), pp. 722–736.

[7] P. Daripa, *A fast algorithm to solve nonhomogeneous Cauchy–Riemann equations in the complex plane*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 1418–1432.

[8] P. Daripa and D. Mashat, *Singular integral transforms and fast numerical algorithms* I, Numer. Algorithms, 18 (1998), pp. 133–157.

[9] J. Daňková and J. Haslinger, *Numerical realization of a fictitious domain approach used in shape optimization.* I. *Distributed controls*, Appl. Math., 41 (1996), pp. 123–147.

[10] E. J. Dean, Q. V. Dinh, R. Glowinski, Jiwen He, T. W. Pan, and J. Périaux, *Least squares/domain imbedding methods for Neumann problems: Applications to fluid dynamics*, in Proceedings of the Fifth International Symposium on Domain Decomposition Methods for Partial Differential Equations, D. E. Keyes, T. F. Chan, G. Meurant, J. S. Scroggs, and R. G. Voigt, eds., SIAM, Philadelphia, 1991, pp. 451–475.

[11] J. Deny and J. L. Lions, *Les espaces du type Beppo-Levi*, Ann. Inst. Fourier (Grenoble), 5 (1953–1954), pp. 305–370.

[12] Q. V. Dinh, R. Glowinski, Jiwen He, T. W. Pan, and J. Périaux, *Lagrange multiplier approach to fictitious domain methods: Applications to fluid dynamics and electro-magnetics*, in Proceedings of the Fifth International Symposium on Domain Decomposition Methods for Partial Differential Equations, D. E. Keyes, T. F. Chan, G. Meurant, J. S. Scroggs, and R. G. Voigt, eds., SIAM, Philadelphia, 1991, pp. 151–194.

[13] M. Elghaoui and R. Pasquetti, *A spectral embedding method applied to the advection-diffusion equation*, J. Comput. Phys., 125 (1996), pp. 464–476.

[14] S. A. Finogenov and Y. A. Kuznetsov, *Two-stage fictitious component methods for solving the Dirichlet boundary value problem*, Soviet J. Numer. Anal. Math. Modelling, 3 (1988), pp. 301–323.

[15] V. Girault, R. Glowinski, and H. Lopez, *Error analysis of a finite element realization of a fictitious domain/domain decomposition method for elliptic problems*, East-West J. Numer. Math., 5 (1997), pp. 35–56.

[16] R. Glowinski and Y. Kuznetsov, *On the solution of the Dirichlet problem for linear elliptic operators by a distributed Lagrange multiplier method*, C.R. Acad. Sci. Paris Sér I Math., 327 (1998), pp. 693–698.

[17] R. Glowinski, T.-W. Pan, T. I. Hesla, D. D. Joseph, and J. Périaux, *A fictitious domain method with distributed Lagrange multipliers for the numerical simulation of a particulate flow*, in Domain Decomposition Methods 10 (Boulder, CO, 1997), Contemp. Math. 218, AMS, Providence, RI, 1998, pp. 121–137.

[18] R. Glowinski, T.-W. Pan, and J. Périaux, *Lagrange multiplier/fictitious domain method for the Dirichlet problem generalization to some flow problems*, Japan J. Indust. Appl. Math., 12 (1995), pp. 87–108.

[19] R. Glowinski, T.-W. Pan, and J. Périaux, *Fictitious domain/Lagrange multiplier methods for partial differential equations*, in Domain-Based Parallelism and Problem Decomposition Methods in Computational Science and Engineering, SIAM, Philadelphia, 1995, pp. 177–192.

[20] G. H. Guirguis, *On the coupling boundary integral and finite element methods for the exterior Stokes problem in* 3*D*, SIAM J. Numer. Anal., 24 (1987), pp. 310–322.

[21] A. Hanouzet, *Espaces de Sobolev avec poids. Application à un problème de Dirichlet dans un demi-espace*, Rend. Sem. Mat. Univ. Padova, 46 (1971), pp. 227–272.

[22] J. Haslinger, *Fictitious domain approaches in shape optimization*, in Computational Methods for Optimal Design and Control (Arlington, VA, 1997), Progr. Systems Control Theory 24, Birkhäuser Boston, Boston, 1998, pp. 237–248.

[23] J. Haslinger and A. Klarbring, *Fictitious domain/mixed finite element approach for a class of optimal shape design problems*, RAIRO Modél. Math. Anal. Numér., 29 (1995), pp. 435–450.

[24] M. N. Le Roux, *Equations intégrales pour le problème du potentiel électrique dans le plan*, C. R. Acad. Sci. Paris, t. 278, série A (1974), pp. 541–544.

[25] M. N. Le Roux, *Méthode d'éléments finis pour la résolution numérique de problèmes extérieurs en dimension* 2, RAIRO Anal. Numér., 11 (1977), pp. 27–60.

[26] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.

[27] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications*, Vol. I, II, III, Springer-Verlag, New York, 1973.

[28] G. I. Marchuk, Y. A. Kuznetsov, and A. M. Matsokin, *Fictitious domain and domain decomposition methods*, Soviet J. Numer. Anal. Math. Modelling, 1 (1986), pp. 3–35.

[29] R. A. E. Mäkinen, P. Neittaanmäki, and D. Tiba, *A boundary controllability approach in optimal shape design problems*, in Boundary Control and Boundary Variation (Sophia-Antipolis, 1990), Lecture Notes in Control Inform. Sci. 178, Springer-Verlag, Berlin, 1992, pp. 309–320.

[30] J. Nečas, *Les méthodes directes en théorie des équation elliptiques*, Editions de l'Academie Tschecoslovaque des Sciences, Prague, 1967.

[31] J. Nedelec and J. Planchard, *Une méthode variationnelle d'éléments finis pour la résolution numérique d'un problème extérieur dans* R^3, RAIRO Anal. Numér., 7 (1973), pp. 105–129.

[32] P. Neittaanmäki and D. Tiba, *On the approximation of the boundary control in two-phase Stefan-type problems*, Control Cybernet., 16 (1987), pp. 33–44.

[33] P. Neittaanmäki and D. Tiba, *An embedding of domains approach in free boundary problems and optimal design*, SIAM J. Control Optim., 33, (1995), pp. 1587–1602.

[34] D. P. O'Leary and O. Widlund, *Capacitance matrix methods for the Helmholtz equation on general three-dimensional regions*, Math. Comp., 3 (1979), pp. 849–879.

[35] W. Proskurowsky and O. B. Widlund, *On the numerical solution of Helmholtz equation by the capacitance matrix method*, Math. Comp., 30 (1979), pp. 433–468.

[36] D. P. Young, R. G. Melvin, M. B. Bieterman, F. T. Johnson, S. S. Samanth, and J. E. Bussolety, *A locally refined finite rectangular grid finite element method. Application to computational physics*, J. Comput. Physics, 92 (1991), pp. 1–66.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 2, pp. 450–474

LINEAR-QUADRATIC CONTROL OF BACKWARD STOCHASTIC DIFFERENTIAL EQUATIONS*

ANDREW E. B. LIM† AND XUN YU ZHOU‡

Abstract. This paper is concerned with optimal control of linear backward stochastic differential equations (BSDEs) with a quadratic cost criteria, or backward linear-quadratic (BLQ) control. The solution of this problem is obtained completely and explicitly by using an approach which is based primarily on the completion-of-squares technique. Two alternative, though equivalent, expressions for the optimal control are obtained. The first of these involves a pair of Riccati-type equations, an uncontrolled BSDE, and an uncontrolled forward stochastic differential equation (SDE), while the second is in terms of a Hamiltonian system. Contrary to the deterministic or stochastic forward case, the optimal control is no longer a feedback of the *current* state; rather, it is a feedback of the *entire* history of the state. A key step in our derivation is a proof of global solvability of the aforementioned Riccati equations. Although of independent interest, this issue has particular relevance to the BLQ problem since these Riccati equations play a central role in our solution. Last but not least, it is demonstrated that the optimal control obtained coincides with the solution of a certain *forward* linear-quadratic (LQ) problem. This, in turn, reveals the origin of the Riccati equations introduced.

Key words. backward stochastic differential equations (BSDEs), linear-quadratic (LQ) optimal control, Riccati equations, completion of squares

AMS subject classifications. 93E20, 49N10, 34A12

PII. S0363012900374737

1. Introduction.

A backward stochastic differential equation (BSDE) is an Ito stochastic differential equation (SDE) for which a *random terminal condition* on the state has been specified. The linear version of this type of equation was first introduced by Bismut [4] as the adjoint equation in the stochastic maximum principle (see also [3, 17, 20]). General nonlinear BSDEs, introduced independently by Pardoux and Peng [16] and Duffie and Epstein [9], have received considerable research attention in recent years due to their nice structure and wide applicability in a number of different areas, especially in mathematical finance (see, e.g., [7, 10, 11, 13, 15, 19]). For example, the Black–Scholes formula for options pricing can be recovered via a system of forward-backward stochastic differential equations (FBSDEs). In this case, the random terminal condition is related to the price of the underlying stock at a given terminal date. Unlike a (forward) SDE, the solution of a BSDE is a *pair* of adapted processes $(x(\cdot), z(\cdot))$. The additional term $z(\cdot)$ may be interpreted as a risk-adjustment factor and is required for the equation to have *adapted* solutions. This restriction of solutions to the class of *adapted processes* is necessary if the insights gained from the study of BSDEs are to be useful in applications. Adapted processes depend on *past* and *present* information but do *not* rely (clairvoyantly) on future knowledge. This is natural in virtually all applications; for example, the replicating portfolio for a contingent claim may depend at any particular time on past and present

*Received by the editors July 5, 2000; accepted for publication (in revised form) January 25, 2001; published electronically July 19, 2001.

http://www.siam.org/journals/sicon/40-2/37473.html

†Center for Applied Probability, Columbia University, New York, NY, 10027 (lim@ieor.columbia.edu).

‡Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong (xyzhou@se.cuhk.edu.hk). The research of this author was supported by the RGC earmarked grants CUHK 4435/99E.

stock prices but not, quite naturally, on future stock prices. For recent accounts on BSDE theory and applications, the reader is referred to the books [15, 19].

Since a BSDE is a well-defined dynamic system, it is very natural and appealing, first at the theoretical level, to consider the optimal control of the BSDE. As for applications, optimally controlled BSDEs promise to have a great potential. For example, an optimal control problem of a linear BSDE comes out in the process of solving a *forward* stochastic linear-quadratic (LQ) control problem in [6]. Moreover, controlled BSDEs are expected to have important applications in mathematical finance. For instance, a situation in which funds may be injected or withdrawn from the replication process of a contingent claim so as to achieve some other goal may be viewed quite naturally as an optimal BSDE control problem. However, the study on controlled BSDEs is quite lacking in literature. To our best knowledge there are only a few papers dealing with optimal control of BSDEs, including [18] and [8], which establish local and global maximum principles, respectively, and [11], in which a controlled BSDE with linear state drift is studied.

This paper is concerned with optimal control of a linear BSDE with a quadratic cost criteria, namely, a stochastic backward linear-quadratic (BLQ) problem. It is well known that LQ control is one of the most important classes of optimal control, and the solution of this problem has had a profound impact on many engineering applications. Stochastic forward LQ theory has been well established, especially with the recent development on the so-called *indefinite* stochastic LQ control [1, 5, 6, 14]. However, stochastic BLQ control remains an almost completely unexplored area. An attempt was made in [8], where a special stochastic BLQ problem without state cost was considered. An optimal control was derived, using the maximum principle obtained in the paper, under the assumption that a certain SDE admits a solution. This SDE, while it resembles the Riccati equation, is not exactly of Riccati type since it is *not* symmetric, and its solvability is hard to verify in general.

The main contribution of this paper is a complete solution of a general BLQ problem. As it turns out, the optimal control can no longer be expressed as a linear feedback of the *current* state as in the deterministic or stochastic forward case. Rather, it depends, in general, on the entire past history of the state pair $(x(\cdot), z(\cdot))$. It will be shown that this dependence is linear, and explicit formulas for the optimal control and the optimal cost in terms of a pair of Riccati equations, a Lyapunov equation, an uncontrolled BSDE, and an uncontrolled SDE are established. The basic idea is to first establish a lower bound to the optimal cost via the completion-of-squares technique and then to construct a control that achieves exactly this lower bound. A key part of our derivation is a proof of existence and uniqueness of solutions of the Riccati equations mentioned above. Although this issue is one which has independent interest, the proof of global solvability presented in this paper has direct relevance to the BLQ problem since these Riccati equations play a central role in our analysis.

It is interesting to remark that our original approach to solving the BLQ problem was inspired by [15, 12], where an (uncontrolled) BSDE is viewed as a *controlled forward* SDE. Extending this idea, we can show that the optimal control of the BLQ problem is the limit of a sequence of square integrable processes, obtained by solving a family of *forward* LQ problems. During this procedure, the key Riccati equations, along with other related equations, come out very naturally. What is more interesting is that once these equations are in place, one may forget about the forward formulation and limiting procedure, which is rather complicated, and instead use these equations *directly* along with the completion-of-squares technique to obtain the optimal control

for the original BLQ problem. Nevertheless, the forward formulation still represents an alternative and insightful approach to the backward control problem, and for this reason, an outline of this procedure is also presented in this paper.

The outline of this paper is as follows. In section 2, we formulate the BLQ problem. In section 3, we present the main result of the paper (with its proof deferred to section 5). In addition, we compare the solution of the stochastic BLQ problem with that of the deterministic case. A key ingredient in our analysis is the existence and uniqueness of solutions of certain Riccati equations, an issue which is addressed in section 4. A proof of the main result is carried out in section 5. In section 6, we explain, in a rather informal way, the origin of the key Riccati equations, and we present an alternative approach to the BLQ problem. In particular, we show that the optimal BLQ control, established in section 3, coincides with the limit of the solutions of a family of forward LQ problems. Finally, section 7 concludes the paper.

2. Problem formulation. We assume throughout that $(\Omega, \mathcal{F}, \{\mathcal{F}\}_{t\geq 0}, P)$ is a given and fixed complete filtered probability space and that $W(\cdot)$ is a scalar-valued Brownian motion on this space. (Our assumption that $W(\cdot)$ is scalar-valued is for the sake of simplicity. No essential difficulties are encountered when extending our analysis to the case of vector-valued Brownian motions). In addition, we assume that $\mathcal{F}_t$ is the augmentation of $\sigma\{W(s) \,|\, 0 \leq s \leq t\}$ by all the P-null sets of $\mathcal{F}$.

Throughout this paper, we denote the set of symmetric $n \times n$ matrices with real elements by S^n. If $M \in S^n$ is positive (semi)definite, we write $M > (\geq)\ 0$. Let X be a given Hilbert space. The set of X-valued continuous functions is denoted by $C(0, T; X)$. If $N(\cdot) \in C(0, T; S^n)$ and $N(t) > (\geq)\ 0$ for every $t \in [0, T]$, we say that $N(\cdot)$ is positive (semi)definite, which is denoted by $N(\cdot) > (\geq)\ 0$. Suppose $\eta : \Omega \to \mathbb{R}^n$ is an $\mathcal{F}_T$-random variable. We write $\eta \in L^2_{\mathcal{F}_T}(\Omega; \mathbb{R}^n)$ if η is square integrable (i.e., $E|\eta|^2 < \infty$). Consider now the case when $f : [0, T] \times \Omega \to \mathbb{R}^n$ is an $\{\mathcal{F}_t\}_{t\geq 0}$ adapted process. If $f(\cdot)$ is square integrable (i.e., $E \int_0^T |f(t)|^2\, dt < \infty$), we shall write $f(\cdot) \in L^2_{\mathcal{F}}(0, T; \mathbb{R}^n)$; if $f(\cdot)$ is uniformly bounded (i.e., ess $\sup_{(t,\, w)\in[0,\, T]\times\Omega} |f(t)| < \infty$), then $f(\cdot) \in L^\infty_{\mathcal{F}}(0, T; \mathbb{R}^n)$. If $f(\cdot)$ has (P-almost surely (a.s.)) continuous sample paths and $E \sup_{t\in[0,\, T]} |f(t)|^2 < \infty$, we write $f(\cdot) \in L^2_{\mathcal{F}}(0, T; C(0, T; \mathbb{R}^n))$; if ess $\sup_{w\in\Omega} \sup_{t\in[0,\, T]} |f(t)| < \infty$, then $f(\cdot) \in L^\infty_{\mathcal{F}}(0, T; C(0, T; \mathbb{R}^n))$. These definitions generalize in the obvious way to the case when $f(\cdot)$ is $\mathbb{R}^{n\times m}$—or S^n—valued. Finally, in cases where we are restricting ourselves to deterministic Borel measurable functions $f : [0, T] \to \mathbb{R}^n$, we shall drop the subscript $\mathcal{F}$ in the notation; for example, $L^\infty(0, T; \mathbb{R}^n)$.

Consider the BSDE

$$\left\{ \begin{array}{rcl} dx(t) &=& \{A(t)\, x(t) + B(t)\, u(t) + C(t) z(t)\}\ dt + z(t)\, dW(t), \\ x(T) &=& \xi, \end{array} \right. \tag{2.1}$$

where $u(\cdot)$ is the control process. The class of *admissible controls* for (2.1) is

$$\mathcal{U} = L^2_{\mathcal{F}}(0, T; \mathbb{R}^m). \tag{2.2}$$

Later, we shall state assumptions on the coefficients $A(\cdot)$, $B(\cdot)$, $C(\cdot)$, and the terminal condition ξ so as to guarantee the existence of a unique solution pair $(x(\cdot), z(\cdot)) \in L^2_{\mathcal{F}}(\Omega; C(0, T; \mathbb{R}^n)) \times L^2_{\mathcal{F}}(0, T; \mathbb{R}^n)$ of the BSDE (2.1) for every admissible control $u(\cdot) \in \mathcal{U}$. We refer to such a three-tuple $(x(\cdot), z(\cdot); u(\cdot))$ as an *admissible triple*. The

cost associated with an admissible triple $(x(\cdot),\, z(\cdot);\, u(\cdot))$ is given by

$$J(\xi; u(\cdot)) := E\,\frac{1}{2}\Bigg[x(0)'Hx(0) + \int_0^T \left(x(t)'Q(t)x(t) + z(t)'S(t)z(t) + u(t)'R(t)u(t)\right)\,dt\Bigg]. \tag{2.3}$$

The *BLQ control problem* can be stated as follows:

$$\begin{cases} \min J(\xi; u(\cdot)) \\ \text{subject to} \\ \quad u(\cdot) \in \mathcal{U}, \\ \quad (x(\cdot),\, z(\cdot);\, u(\cdot)) \text{ satisfies (2.1)}. \end{cases} \tag{2.4}$$

Throughout this paper, we shall assume the following:

Assumption (A1).

$$\begin{cases} A,\, C & \in & L^\infty(0,\, T;\, \mathbb{R}^{n\times n}), \\ B & \in & L^\infty(0,\, T;\, \mathbb{R}^{n\times m}), \\ Q,\, S & \in & L^\infty(0,\, T;\, S^n),\ Q,\, S \,\geq\, 0, \\ R & \in & L^\infty(0,\, T; S^m),\ R \,>\, 0, \\ H & \in & S^n,\ H \,\geq\, 0, \\ \xi & \in & L^2_{\mathcal{F}_T}(\Omega;\, \mathbb{R}^n). \end{cases}$$

In particular, Assumption (A1) is sufficient to guarantee the existence of a unique solution pair $(x(\cdot),\, z(\cdot)) \in L^2_{\mathcal{F}}(\Omega; C(0,\, T;\, \mathbb{R}^n)) \times L^2_{\mathcal{F}}(0,\, T;\, \mathbb{R}^n)$ of (2.1) for every admissible control $u(\cdot) \in \mathcal{U}$; see [19, Chapter 7].

3. Main result. Before we present the main result of the paper, which gives a complete solution to the above BLQ problem, let us see how one would solve the *deterministic* BLQ problem. This corresponds to $\xi \in \mathbb{R}^n$ being deterministic, $C = 0$, $S = 0$, and an admissible class $\mathcal{U}_d = L^2(0,\, T;\, \mathbb{R}^n)$. The other parameters satisfy (A1), while the cost and dynamics are given by

$$J_d(\xi; u(\cdot)) := \frac{1}{2}\,x(0)'Hx(0) + \frac{1}{2}\int_0^T \left(x(t)'Q(t)x(t) + u(t)'R(t)u(t)\right)\,dt,$$

$$\begin{cases} \dot{x}(t) = A(t)\,x(t) + B(t)\,u(t), \\ x(T) = \xi, \end{cases}$$

respectively. By reversing time,

$$\tau = T - t, \quad t \in [0,\, T],$$

we obtain an equivalent forward LQ problem that can be solved using a standard (Riccati) approach (see, e.g., [19, Chapter 6, section 2]). In particular, this gives us the following result.

Proposition 3.1 (deterministic BLQ problem). *The optimal cost and optimal feedback control for the deterministic BLQ problem are*

$$J_d^*(\xi) = \frac{1}{2}\xi'\,Z(T)\,\xi, \tag{3.1}$$

$$u(t) = R(t)^{-1}B(t)'Z(t)x(t), \tag{3.2}$$

respectively, where $Z(\cdot)$ is the unique solution of the Riccati equation

$$\begin{cases} \dot{Z}(t) + Z(t)A(t) + A(t)'Z(t) + Z(t)B(t)R(t)^{-1}B(t)'Z(t) - Q(t) = 0, \\ Z(0) = H, \end{cases} \tag{3.3}$$

and $x(\cdot)$ is the unique solution of the differential equation

$$\begin{cases} \dot{x}(t) = (A(t) + B(t)\, R(t)^{-1}B(t)'Z(t))\, x(t), \\ x(T) = \xi. \end{cases}$$

It is important to recognize that the above time reversal technique *cannot* be extended to the stochastic BLQ problem, (2.4), as it would destroy the adaptiveness which is essential in the model. In particular, a control obtained in this way will not, in general, be $\{\mathcal{F}_t\}_{t\geq 0}$-adapted and hence is not admissible.

It turns out that the solution to (2.4) is more involved. In the remainder of this section, we present two alternative expressions (which are later shown to be equivalent) for the solution of the optimal BLQ control. The first one is analogous to the solution to the deterministic BLQ problem just presented. It gives an explicit formula via a pair of Riccati equations, a Lyapunov equation, an uncontrolled BSDE, and an uncontrolled SDE.

First, consider the following Riccati-type equation:

$$\begin{cases} \dot{\Sigma}(t) - A(t)\Sigma(t) - \Sigma(t)A(t)' - \Sigma(t)Q(t)\Sigma(t) \\ \quad +B(t)R(t)^{-1}B(t)' + C(t)\Sigma(t)\left(S(t)\Sigma(t) + I\right)^{-1}C(t)' = 0, \\ \Sigma(T) = 0. \end{cases} \tag{3.4}$$

The existence and uniqueness of a solution to this equation will be addressed in section 4; see Theorem 4.5. Letting $\Sigma(\cdot)$ be the solution to (3.4), we define the following equations:

$$\begin{cases} \dot{Z}(t) + Z(t)A(t) + A(t)'Z(t) \\ \quad +Z(t)[B(t)R(t)^{-1}B(t)' \\ \quad +C(t)\Sigma(t)(I + S(t)\Sigma(t))^{-1}C(t)']Z(t) - Q(t) = 0, \\ Z(0) = H, \end{cases} \tag{3.5}$$

$$\begin{cases} \dot{N}(t) + N(t)(A(t) + \Sigma(t)Q(t)) + (A(t) + \Sigma(t)Q(t))'N(t) - Q(t) = 0, \\ N(0) = \frac{1}{2}\{H(I + \Sigma(0)H)^{-1} + (I + H\Sigma(0))^{-1}H\}, \end{cases} \tag{3.6}$$

$$\begin{cases} dh(t) = \{(A(t) + \Sigma(t)Q(t))h(t) + C(t)(I + \Sigma(t)S(t))^{-1}\,\eta(t)\}\,dt \\ \quad +\eta(t)\,dW(t), \\ h(T) = -\xi. \end{cases} \tag{3.7}$$

The first equation (3.5) is again a Riccati-type equation. It is a generalization of the Riccati equation (3.3) associated with the deterministic problem. The second equation is a Lyapunov equation, while the third is a linear BSDE. Based on the solutions $Z(\cdot)$ and $(h(\cdot), \eta(\cdot))$ to (3.5) and (3.7), respectively, we finally introduce

$$\begin{cases} dq(t) = \{-[A(t) + B(t)R(t)^{-1}B(t)'Z(t) + C(t)(I + \Sigma(t)S(t))^{-1}\Sigma(t)C(t)'Z(t)]'\, q(t) \\ \quad +Z(t)C(t)(I + \Sigma(t)S(t))^{-1}\eta(t)\}dt + \{(Z(t) - S(t))(I + \Sigma(t)S(t))^{-1}\eta(t) \\ \quad +(I + Z(t)\Sigma(t))(I + S(t)\Sigma(t))^{-1}C(t)'(I + Z(t)\Sigma(t))^{-1}(Z(t)h(t) - q(t))\}\,dW(t), \\ q(0) = 0. \end{cases} \tag{3.8}$$

The existence and uniqueness of the solutions of (3.5)–(3.8) will be discussed in section 4. It should be noted that (3.4), (3.6), (3.7), and (3.8) play no role in the solution of the deterministic BLQ problem.

THEOREM 3.2. *The BLQ problem* (2.4) *is uniquely solvable. Moreover, the control*

$$u(t) = R(t)^{-1}B(t)'(Z(t)x(t) + q(t)) \tag{3.9}$$

is optimal, where $Z(\cdot)$ *and* $q(\cdot)$ *are the solutions of* (3.5) *and* (3.8)*, respectively. The optimal state trajectory* $(x(\cdot), z(\cdot))$ *is the unique solution of the BSDE*

$$\begin{cases} dx(t) = \{(A(t) + B(t)\,R(t)^{-1}B(t)'Z(t))\,x(t) \\ \qquad\qquad +C(t)z(t) + B(t)\,R(t)^{-1}\,B(t)'\,q(t)\}\,dt + z(t)\,dW(t), \\ x(T) = \xi, \end{cases} \tag{3.10}$$

and the optimal cost is

$$J^*(\xi) := E\frac{1}{2}\Bigg\{\xi'N(T)\xi + \int_0^T \{\eta(t)'[(S(t)\Sigma(t) + I)^{-1}S(t) - N(t)]\eta(t) \\ -2\eta(t)'(I + S(t)\Sigma(t))^{-1}C(t)'N(t)h(t)\}\,dt\Bigg\}, \tag{3.11}$$

where $N(\cdot)$ *is the unique solution of* (3.6).

Remark 3.1. If we compare the two optimal controls, (3.2) and (3.9), for the deterministic and stochastic BLQ problems, respectively, we see that the latter involves an additional random nonhomogeneous term $q(\cdot)$. This addition disqualifies (3.9) from a feedback control of the *current* state, contrary to the deterministic BLQ (see Proposition 3.1) or stochastic forward LQ (see [5]) cases. The reason is because $q(\cdot)$ depends on $(h(\cdot), \eta(\cdot))$, which in turn depends on ξ, the terminal condition of part of the state variable, $x(\cdot)$. This is one of the major distinctive features of the stochastic BLQ problem. On the other hand, when ξ is nonrandom, $C = 0$ and $S = 0$, the optimal control (3.9) reduces to the solution (3.2) of the deterministic problem. In this case, it is easy to see (by the uniqueness of the solutions of (3.7)) that $\eta(t) \equiv 0$. This implies, in turn, that $q(t) \equiv 0$, and hence the optimal control (3.9) agrees with the solution (3.2) of the deterministic problem. In addition, since

$$N(t) = \frac{1}{2}[Z(t)(I + \Sigma(t)Z(t))^{-1} + (I + Z(t)\Sigma(t))^{-1}Z(t)]$$

(see Proposition 4.8), it follows that $N(T) = Z(T)$ and the optimal cost (3.11) reduces to (3.1) for the deterministic problem. Through the above comparison, we can also see that the fundamental difference between the solutions to the deterministic and stochastic BLQ problems lies in the introduction of (3.4).

Although for the stochastic BLQ problem the optimal control is no longer a feedback of the current state, it is indeed a linear state feedback of the *entire past history* of the state process $(x(\cdot), z(\cdot))$. This conclusion is a consequence of the second form of the optimal control we will present, which is in terms of the Hamiltonian system:

$$\begin{cases} dx(t) = \{A(t)x(t) - B(t)R(t)^{-1}B(t)'y(t) + C(t)z(t)\}dt + z(t)dW(t), \\ x(T) = \xi, \end{cases} \tag{3.12}$$

$$\begin{cases} dy(t) = \{-A(t)'y(t) - Q(t)x(t)\}dt + \{-C(t)'y(t) - S(t)z(t)\}dW(t), \\ y(0) = -Hx(0). \end{cases} \tag{3.13}$$

Notice that the combination of (3.12)–(3.13) does not qualify as a conventional FBSDE as defined in, say, [19, 15]. The subtle difference is that the forward and backward variables in (3.12)–(3.13) are directly related at the *initial* time, while those in the FBSDE are related at the *terminal* time. Moreover, one cannot transform between these two types of equations by reversing the time, due to the required adaptiveness. In what follows, we shall refer to any three-tuple of processes

$$(x(\cdot),\, z(\cdot),\, y(\cdot)) \in L^2_{\mathcal{F}}(\Omega;\, C(0,\, T;\, \mathbb{R}^n)) \times L^2_{\mathcal{F}}(0,\, T;\, \mathbb{R}^n) \times L^2_{\mathcal{F}}(\Omega;\, C(0,\, T;\, \mathbb{R}^n)),$$

which satisfies (3.12)–(3.13) as a *solution of the Hamiltonian system* (3.12)–(3.13).

THEOREM 3.3. *The Hamiltonian system* (3.12)–(3.13) *has a unique solution* $(x(\cdot),\, z(\cdot),\, y(\cdot))$. *Moreover, the BLQ problem* (2.4) *is uniquely solvable with the optimal control*

$$u(t) = -R(t)^{-1}B(t)'y(t), \tag{3.14}$$

and $(x(\cdot),\, z(\cdot))$ *as the corresponding optimal state process. The optimal cost is* (3.11).

Remark 3.2. If (3.14) is optimal, then (3.12)–(3.13) are exactly the corresponding state equation and adjoint equation; see [8]. This is the reason why we call (3.12)–(3.13) the Hamiltonian system.

Theorem 3.3 shows that the optimal control is linear in the process $y(\cdot)$. The following simple result further reveals that the optimal control is a linear feedback of the past and current values of the state process $(x(\cdot),\, z(\cdot))$.

PROPOSITION 3.4. *Let* $y(\cdot)$ *be the process obtained from the Hamiltonian system* (3.12)–(3.13). *Then*

$$\begin{aligned} y(t) = \Phi(t)\Bigg\{ & -Hx(0) + \int_0^t \Phi(s)^{-1}[Q(s)x(s) + C(s)'S(s)z(s)]\, ds \\ & - \int_0^t \Phi(s)^{-1}S(s)z(s)dW(s)\Bigg\}, \end{aligned}$$

where $\Phi(\cdot)$ *is the unique solution of the matrix SDE*

$$\begin{cases} d\Phi(t) &= -A(t)'\Phi(t)dt - C(t)'\Phi(t)dW(t), \\ \Phi(0) &= I. \end{cases}$$

Proof. This is an immediate consequence of the variation-of-constant formula; see [19, p. 47, Theorem 6.14]. □

Proofs of Theorems 3.2 and 3.3 are deferred to section 5.

4. Riccati equations. Before proving the main result formulated in the previous section, in this section we first study the existence and uniqueness of solutions to (3.4)–(3.8), mainly focusing on the Riccati equations (3.4) and (3.5).

To start, let us first consider the two equations

$$\begin{cases} \dot{\Sigma}(t) - A(t)\Sigma(t) - \Sigma(t)A(t)' - \Sigma(t)Q(t)\Sigma(t) \\ \quad +B(t)R(t)^{-1}B(t)' + C(t)\Sigma(t)(S(t)\Sigma(t) + I)^{-1}C(t)' = 0, \\ \Sigma(T) = M, \end{cases} \tag{4.1}$$

$$\begin{cases} \dot{P}(t) + P(t)A(t) + A(t)'P(t) \\ \quad -P(t)(B(t)R(t)^{-1}B(t)' + C(t)(S(t) + P(t))^{-1}C(t)')P(t) + Q(t) = 0, \\ P(T) = M^{-1}, \end{cases} \tag{4.2}$$

where M is a given symmetric $n \times n$ matrix in (4.1) and a nonsingular symmetric $n \times n$ matrix in (4.2). It will be seen from what follows that (4.2) is introduced as a means of dealing with the solvability of (4.1).

PROPOSITION 4.1. *Let M be a symmetric $n \times n$ matrix. If the Riccati equation* (4.1) *is solvable, then the solution is unique.*

Proof. Suppose that $\Sigma_1(\cdot)$, $\Sigma_2(\cdot) \in C(0, T; S^n)$ are two solutions of (4.1). Since $\Sigma_1(\cdot)$ and $\Sigma_2(\cdot)$ are continuous, it follows that $\Delta(\cdot) := \Sigma_1(\cdot) - \Sigma_2(\cdot)$ is uniformly bounded. It is easy to show that $\Delta(\cdot)$ is a solution of the equation

$$\begin{cases} \dot{\Delta}(t) = (A(t) + \Sigma_1(t)\,Q(t))\,\Delta(t) + \Delta(t)\,(A(t) + \Sigma_1(t)\,Q(t))' - \Delta(t)Q(t)\Delta(t) \\ \quad -C(t)[I - \Sigma_2(t)(S(t)\Sigma_2(t) + I)^{-1}S(t)]\Delta(t)(S(t)\Sigma_1(t) + I)^{-1}C(t)', \\ \Delta(T) = 0. \end{cases}$$

Integrating both sides of this equation from t to T, it follows from the uniform boundedness of $\Delta(\cdot)$ and all the coefficients that there is a constant $0 < K < \infty$ such that

$$\|\Delta(t)\| \le K \int_t^T \|\Delta(s)\|\,ds.$$

Hence, by Gronwall's inequality, it follows that $\Sigma_1(t) - \Sigma_2(t) = 0$ for all $t \in [0, T]$. □

Next we prove the existence of solutions to (4.1). We first consider the case when $S = 0$. In this case, the Riccati equations (4.1) and (4.2) become

$$\begin{cases} \dot{\Sigma}(t) - A(t)\Sigma(t) - \Sigma(t)A(t)' + C(t)\Sigma(t)C(t)' \\ \qquad -\Sigma(t)Q(t)\Sigma(t) + B(t)R(t)^{-1}B(t)' = 0, \\ \Sigma(T) = M, \end{cases} \tag{4.3}$$

$$\begin{cases} \dot{P}(t) + P(t)A(t) + A(t)'P(t) + Q(t) \\ \qquad -P(t)(B(t)R(t)^{-1}B(t)' + C(t)P(t)^{-1}C(t)')P(t) = 0, \\ P(T) = M^{-1}. \end{cases} \tag{4.4}$$

PROPOSITION 4.2. *Let $M \ge 0$ be a given symmetric $n \times n$ matrix. Then the Riccati equation* (4.3) *is uniquely solvable. Moreover,*

(i) *if $M > 0$, then the solution $\Sigma(\cdot) > 0$, and*

(ii) *if $M \ge 0$, then the solution $\Sigma(\cdot) \ge 0$.*

Proof. Case 1*: $M > 0$.* Consider first the case when $Q(t) > 0$ for almost every (a.e.) $t \in [0, T]$. Then (4.3) is a standard Riccati equation (arising in deterministic LQ control) and is uniquely solvable with the solution $\Sigma(\cdot) > 0$ (see, e.g., [2, 19]). Suppose now that we have only $Q(\cdot) \ge 0$. Define $Q_i := Q + (1/i)\,I$ for $i \in \mathbb{Z}^+$. Let $\Sigma_i(\cdot)$ be the unique positive definite solution of (4.3) when Q is replaced by Q_i. Note first that $\Sigma_i(\cdot)$ is uniformly bounded. To see this, consider the Lyapunov equation

$$\begin{cases} \dot{\bar{\Sigma}}(t) = A(t)\bar{\Sigma}(t) + \bar{\Sigma}(t)A(t)' - C(t)\bar{\Sigma}(t)C(t)' - B(t)R(t)^{-1}B(t)', \\ \bar{\Sigma}(T) = M. \end{cases} \tag{4.5}$$

Since (4.5) is a linear ordinary differential equation (ODE) with bounded coefficients, it follows that it has a unique solution $\bar{\Sigma}(\cdot)$ which is uniformly bounded. For any

$i \in \mathbb{Z}^+$, let $\bar{\Delta}_i(\cdot) := \bar{\Sigma}(\cdot) - \Sigma_i(\cdot)$. It is easy to show that $\bar{\Delta}_i(\cdot)$ is a solution of the Riccati equation

$$\begin{cases} \dot{\bar{\Delta}}_i(t) = A_i(t)\bar{\Delta}_i(t) + \bar{\Delta}_i(t)A_i(t)' - C(t)\bar{\Delta}_i(t)C(t)' \\ \qquad\qquad + \bar{\Delta}_i(t)Q_i(t)\bar{\Delta}_i(t) - \bar{\Sigma}(t)Q_i(t)\bar{\Sigma}(t), \\ \bar{\Delta}(T) = 0, \end{cases} \tag{4.6}$$

where $A_i(t) := A(t) + \Sigma_i(t)Q_i(t)$. This is again a standard Riccati equation which has a unique solution $\bar{\Delta}_i(\cdot) \geq 0$. Therefore, $0 \leq \Sigma_i(\cdot) \leq \bar{\Sigma}(\cdot)$, so $\Sigma_i(\cdot)$ is uniformly bounded, as claimed. Next, observe that $\Sigma_i(\cdot)$ is nondecreasing in i. To see this, suppose that $j < i$. Then $\Delta(\cdot) := \Sigma_i(\cdot) - \Sigma_j(\cdot)$ is the unique solution of the Riccati equation

$$\begin{cases} \dot{\Delta}(t) = \bar{A}(t)\Delta(t) + \Delta(t)\bar{A}(t)' - C(t)\Delta(t)C(t)' + \Delta(t)Q_i(t)\Delta(t) - (\frac{1}{j} - \frac{1}{i})\Sigma_j(t)\,\Sigma_j(t), \\ \Delta(T) = 0, \end{cases}$$

where $\bar{A} := A + \Sigma_j Q_i$. As before, $\Delta(\cdot)$ is positive semidefinite, and hence $\Sigma_i(\cdot) \geq \Sigma_j(\cdot)$. Since $\{\Sigma_i(\cdot)\}_{i\geq 1}$ is a nondecreasing, uniformly bounded sequence of functions, it follows that there is a function $\Sigma(\cdot)$ (which is not necessarily continuous) such that $\Sigma_i(t) \uparrow \Sigma(t)$ for every $t \in [0, T]$ as $i \uparrow \infty$. Therefore, $\Sigma(t)$ is symmetric, and $\Sigma(t) > 0$ for every $t \in [0, T]$. Finally, we show that $\Sigma(\cdot)$ is continuous and is a solution of (4.3). Observe first that by virtue of (4.3), the relation

$$\begin{aligned} \Sigma_i(t) = M - \int_t^T & (A(s)\Sigma_i(s) + \Sigma_i(s)A(s)' - C(s)\Sigma_i(s)C(s)' \\ & + \Sigma_i(s)Q_i(s)\Sigma_i(s) - B(s)R(s)^{-1}B(s)')\,ds \end{aligned}$$

holds. Since $\Sigma_i(t) \uparrow \Sigma(t)$ for every $t \in [0, T]$ as $i \uparrow \infty$, it follows from the bounded convergence theorem that

$$\begin{aligned} \Sigma(t) = M - \int_t^T & (A(s)\Sigma(s) + \Sigma(s)A(s)' - C(s)\Sigma(s)C(s)' \\ & + \Sigma(s)Q(s)\Sigma(s) - B(s)R(s)^{-1}B(s')\,ds. \end{aligned}$$

Therefore, $\Sigma(\cdot) \in C(0, T; S^n)$ is a solution of (4.3), and $\Sigma(\cdot) > 0$. Uniqueness follows from Proposition 4.1.

Case 2: $M \geq 0$. In this case, the one difference, when applying the argument above, is that $\Sigma(t) \geq 0$ instead of $\Sigma(t) > 0$ for all $t \in [0, T]$. □

PROPOSITION 4.3. *Let $M > 0$ be a given symmetric $n \times n$ matrix. Then the Riccati equation* (4.4) *is uniquely solvable with the solution $P(\cdot) > 0$.*

Proof. We begin by proving existence. By Proposition 4.2, (4.3) is uniquely solvable with the solution $\Sigma(\cdot) > 0$. It follows that $\Sigma(\cdot)^{-1}$ is well defined, symmetric, and positive definite. By evaluating $\frac{d}{dt}(\Sigma(t)\Sigma(t)^{-1}) = 0$, it can be shown that $P(\cdot) := \Sigma(\cdot)^{-1} \in C(0, T, S^n)$ is a solution of (4.4).

To prove uniqueness, let $P_i(\cdot)$, $i = 1, 2$, be solutions of (4.4). Since (4.4) involves $P_i(\cdot)^{-1}$, $P_i(t)$ is invertible for every $t \in [0, T]$, and $P_i(\cdot)^{-1}$ is differentiable. Since $P_i(\cdot)^{-1}$ is a solution of (4.3), the uniqueness property of (4.3) implies that $P_1(\cdot) = P_2(\cdot)$. □

Now we proceed to the general case when $S \geq 0$. We begin by proving global solvability of the Riccati equation (4.2) when $M > 0$. The following notions, introduced in [5], play an important role in our analysis. Let

$$\hat{\mathcal{K}} := \{K \in L^\infty(0,\, T;\, S^n) \,|\, K(t),\, K(t)^{-1} > 0, \\ \text{a.e. } t \in [0,\, T],\ \text{and } K(\cdot)^{-1} \in L^\infty(0,\, T;\, S^n)\}.$$

For every $K \in \hat{\mathcal{K}}$, the Riccati equation

$$\begin{cases} \dot{P}(t) + P(t)A(t) + A(t)'P(t) \\ \quad -P(t)(B(t)R(t)^{-1}B(t)' + C(t)K(t)^{-1}C(t)')P(t) + Q(t) = 0, \\ P(T) = M^{-1} \end{cases} \tag{4.7}$$

is a standard Riccati equation which is uniquely solvable, with the solution $P(\cdot) > 0$. Therefore, the mapping $\psi : \hat{\mathcal{K}} \to C(0,\, T;\, S^n)$, where $P = \psi(K)$ is the solution of (4.7) associated with K, is well defined. A sufficient condition for unique solvability of the Riccati equation (4.2) is the existence of $K \in \hat{\mathcal{K}}$ such that

$$S + \psi(K) \geq K; \tag{4.8}$$

see [5, Theorem 4.6]. Hence we have the following result.

THEOREM 4.4. *Let $M > 0$ be a given symmetric $n \times n$ matrix. Then the Riccati equation* (4.2) *has a unique solution $P(\cdot) \in C(0,\, T;\, S^n)$. Moreover, $P(\cdot) > 0$.*

Proof. Let $\bar{P}(\cdot) \in C(0,\, T;\, S^n)$ denote the solution of the Riccati equation (4.4). It follows from Proposition 4.3 that $\bar{P} \in \hat{\mathcal{K}}$ and $\bar{P} = \psi(\bar{P})$. Moreover, since $S \geq 0$, it is clear that (4.8) is satisfied with $K = \bar{P}$, and hence (4.2) is uniquely solvable (with solution $P(\cdot) \in C(0,\, T;\, S^n)$). To see that $P(t) \geq \bar{P}(t) > 0$ for all $t \in [0,\, T]$, observe that $x'P(t)x$ is the optimal cost associated with the optimal control problem [5]

$$\begin{cases} \min_{u(\cdot),\, v(\cdot)} E\int_t^T \{x(s)'Q(s)x(s) + u(s)'R(s)u(s) + v(s)'S(s)v(s)\}\, ds + x(T)'M^{-1}x(T) \\ \text{subject to} \\ \quad dx(s) = \{A(s)x(s) + B(s)u(s) + C(s)v(s)\}\, ds + v(s)\, dW(s), \quad s \in [t, T], \\ \quad x(t) = x, \\ \quad (u(\cdot),\, v(\cdot)) \in L^2_{\mathcal{F}}(0,\, T;\, \mathbb{R}^m) \times L^2_{\mathcal{F}}(0,\, T;\, \mathbb{R}^n), \end{cases} \tag{4.9}$$

while $x'\bar{P}(t)x$ is the optimal cost associated with

$$\begin{cases} \min_{u(\cdot),\, v(\cdot)} E\int_t^T \{x(s)'Q(s)x(s) + u(s)'R(s)u(s)\}\, ds + x(T)'M^{-1}x(T) \\ \text{subject to} \\ \quad dx(s) = \{A(s)x(s) + B(s)u(s) + C(s)v(s)\}\, ds + v(s)\, dW(s), \quad s \in [t, T], \\ \quad x(t) = x, \\ \quad (u(\cdot),\, v(\cdot)) \in L^2_{\mathcal{F}}(0,\, T;\, \mathbb{R}^m) \times L^2_{\mathcal{F}}(0,\, T;\, \mathbb{R}^n). \end{cases} \tag{4.10}$$

Since $S \geq 0$, it follows that $x'P(t)x \geq x'\bar{P}(t)x$ for all $(t,\, x) \in [0,\, T] \times \mathbb{R}^n$, from which the result follows. □

Remark 4.1. A special case of Theorem 4.4 is proved in [12, Theorem 4.2] under an additional assumption (inequality (4.3) in [12]).

THEOREM 4.5. *Let $M \geq 0$ be a given symmetric $n \times n$ matrix. Then the Riccati equation* (4.1) *is uniquely solvable. Moreover,*

(i) *if $M > 0$, then the solution $\Sigma(\cdot) > 0$, and*

(ii) *if $M \geq 0$, then the solution $\Sigma(\cdot) \geq 0$.*

Proof. We consider the issue of existence for the cases $M > 0$ and $M \geq 0$ separately. Uniqueness follows immediately from Proposition 4.1.

Case 1: $M > 0$. Let $P(\cdot)$ denote the solution of the Riccati equation (4.2). By Theorem 4.4, it follows that $P(t) > 0$ for all $t \in [0, T]$. Therefore, $P(t)^{-1}$ is well defined. Using the fact that $\frac{d}{dt}(P(t)^{-1}P(t)) = 0$ and $(S(t) + P(t))^{-1} = P(t)^{-1}(S(t)P(t)^{-1} + I)^{-1}$, it can be shown that $\Sigma(\cdot) := P(\cdot)^{-1}$ is a solution of (4.1). Clearly, $\Sigma(t) > 0$ for all $t \in [0, T]$.

Case 2: $M \geq 0$. Let $M_i := M + (1/i)I$ and $\Sigma_i(\cdot)$, $P_i(\cdot)$ denote the solutions of (4.1) and (4.2), respectively, corresponding to $M_i > 0$. Then $\Sigma_i(\cdot) = P_i(\cdot)^{-1} > 0$. Since $x'P_i(t)x$ is the optimal cost for the optimal control problem

$$\begin{cases} \min_{u(\cdot),\, v(\cdot)} E\int_t^T \{x(s)'Q(s)x(s) + u(s)'R(s)u(s) + v(s)'S(s)v(s)\}\, ds + x(T)'M_i^{-1}x(T) \\ \text{subject to} \\ \quad dx(s) = \{A(s)x(s) + B(s)u(s) + C(s)v(s)\}\, ds + v(s)\, dW(s), \quad s \in [t, T], \\ \quad x(t) = x, \\ \quad (u(\cdot),\, v(\cdot)) \in L^2_{\mathcal{F}}(0,\, T;\, \mathbb{R}^m) \times L^2_{\mathcal{F}}(0,\, T;\, \mathbb{R}^n), \end{cases}$$

it follows that $0 < P_i(t) \leq P_j(t)$ for all $i < j$, and hence $0 < \Sigma_j(t) \leq \Sigma_i(t)$. Therefore, $\Sigma_i(t)$ is a monotonically decreasing sequence that is bounded below and hence converges; that is, $\Sigma_i(t) \downarrow \Sigma(t) \geq 0$ for all $t \in [0, T]$. On the other hand,

$$\begin{aligned} \Sigma_i(t) = M + \frac{1}{i}I - \int_t^T \{&A(s)\Sigma_i(s) + \Sigma_i(s)A(s)' \\ &- B(s)R(s)^{-1}B(s)' - C(s)\Sigma_i(s)[S(s)\Sigma_i(s) + I]^{-1}C(s)' + \Sigma_i(s)Q(s)\Sigma_i(s)\}\, ds. \end{aligned}$$

Hence it follows from the bounded convergence theorem that

$$\begin{aligned} \Sigma(t) = M - \int_t^T \{&A(s)\Sigma(s) + \Sigma(s)A(s)' \\ &- B(s)R(s)^{-1}B(s)' - C(s)\Sigma(s)[S(s)\Sigma(s) + I]^{-1}C(s)' + \Sigma(s)Q(s)\Sigma(s)\}\, ds, \end{aligned}$$

so $\Sigma(\cdot)$ is a solution of (4.1). □

The above theorem implies, in particular, that (3.4) is uniquely solvable. Now we are in the position to prove the unique solvability of the Riccati equation (3.5).

COROLLARY 4.6. *Let* $\Sigma(\cdot)$ *denote the solution of* (3.4). *Then the Riccati equation* (3.5) *is uniquely solvable. Moreover,*

(i) *if* $H > 0$, *then the solution* $Z(\cdot) > 0$, *and*

(ii) *if* $H \geq 0$, *then the solution* $Z(\cdot) \geq 0$.

Proof. By making the time reversing transformation

$$\tau = T - t, \quad t \in [0, T],$$

the Riccati equation (3.5) is equivalent to

$$\begin{cases} \dot{Z}(t) - Z(t)A(t) - A(t)'Z(t) - Z(t)\left[B(t)R(t)^{-1}B(t)' \right. \\ \qquad\qquad \left. + C(t)\Sigma(t)(I + S(t)\Sigma(t))^{-1}C(t)'\right] Z(t) + Q(t) = 0, \\ Z(T) = H. \end{cases} \tag{4.11}$$

We now show that (4.11) is a special case of the Riccati equation (4.1). To see this, let $\Sigma_i(\cdot)$ and $P_i(\cdot)$ denote the solutions of (4.1) and (4.2) when $M = (1/i)I$. Since

$$\Sigma(t)(I + S(t)\Sigma(t))^{-1} = \lim_{i \uparrow \infty} \Sigma_i(t)(I + S(t)\Sigma_i(t))^{-1} = \lim_{i \uparrow \infty}(S(t) + P_i(t))^{-1},$$

which implies, in particular, that $B(t)R(t)^{-1}B(t)' + C(t)\Sigma(t)(I + S(t)\Sigma(t))^{-1}C(t)'$ is symmetric, and

$$Q(t) = (Q(t)^{\frac{1}{2}})\, I^{-1}\, (Q(t)^{\frac{1}{2}})',$$

it follows that (4.11) is an equation of the form (4.1). Therefore, Theorem 4.5 applies, and (3.5) is uniquely solvable. □

Now we have proved the unique solvability of the two Riccati equations (3.4) and (3.5). The unique solvability of (3.7) and (3.8), with the solutions $(h(\cdot), \eta(\cdot)) \in L^2_{\mathcal{F}}(\Omega; C(0, T; \mathbb{R}^n)) \times L^2_{\mathcal{F}}(0, T; \mathbb{R}^n)$ and $q(\cdot) \in L^2_{\mathcal{F}}(\Omega; C(0, T; \mathbb{R}^n))$, is evident as they are linear BSDE/SDE with bounded linear coefficients and square integrable nonhomogeneous terms; see, e.g., [19, Theorem 2.2, p. 349] and [19, Theorem 6.14, p. 47]. Finally, the unique solvability of the Lyapunov equation (3.6) is well known.

Next let us study the asymptotic behavior of some equations with respect to the terminal condition of those equations, which is important in proving the main results, Theorems 3.2 and 3.3. Let $M > 0$ be a symmetric $n \times n$ matrix. Consider the following equations parameterized by M:

$$\text{(4.12)} \quad \begin{cases} \dot{P}(t) + P(t)A(t) + A(t)'P(t) \\ \quad -P(t)[B(t)R(t)^{-1}B(t)' + C(t)(S(t) + P(t))^{-1}C(t)']P(t) + Q(t) = 0, \\ P(T) = M^{-1}, \end{cases}$$

$$\text{(4.13)} \quad \begin{cases} \dot{\Sigma}(t) - A(t)\Sigma(t) - \Sigma(t)A(t)' - \Sigma(t)Q(t)\Sigma(t) \\ \quad +B(t)R(t)^{-1}B(t)' + C(t)\Sigma(t)(S(t)\Sigma(t) + I)^{-1}C(t)' = 0, \\ \Sigma(T) = M, \end{cases}$$

$$\text{(4.14)} \quad \begin{cases} dh(t) = \{(A(t) + \Sigma(t)Q(t))h(t) + C(t)(I + \Sigma(t)S(t))^{-1}\, \eta(t)\}\, dt \\ \quad +\eta(t)\, dW(t), \\ h(T) = -\xi. \end{cases}$$

Notice that the $\Sigma(\cdot)$ appearing on the right-hand side of (4.14) is the solution to (4.13) which depends on M; hence (4.14) and (3.7) are *different*.

PROPOSITION 4.7. *Let M_i $(i \in \mathbb{Z}^+)$ and M be symmetric, $n \times n$, positive semidefinite matrices. Let $\Sigma_i(\cdot)$, $(h_i(\cdot), \eta_i(\cdot))$ and $\Sigma(\cdot)$, $(h(\cdot), \eta(\cdot))$ be solutions of* (4.13)–(4.14)*, corresponding to M_i and M, respectively. If $M_i \to M$, then $\Sigma_i(\cdot) \to \Sigma(\cdot)$, uniformly on $[0, T]$, and $(h_i(\cdot), \eta_i(\cdot)) \to (h(\cdot), \eta(\cdot))$ in $L^2_{\mathcal{F}}(\Omega; C(0, T; \mathbb{R}^n)) \times L^2_{\mathcal{F}}(0, T; \mathbb{R}^n)$, as $i \uparrow \infty$.*

Proof. Since every convergent sequence is bounded, there exists $0 < \bar{M} \in S^n$ such that $M_i \le \bar{M}$ and $M \le \bar{M}$. Therefore, $\Sigma(\cdot) \le \bar{\Sigma}(\cdot)$ and $\Sigma_i(\cdot) \le \bar{\Sigma}(\cdot)$, where $\bar{\Sigma}(\cdot)$ is the solution of (4.1) corresponding to $\bar{M}$. It follows that if $\Delta_i(\cdot) := \Sigma_i(\cdot) - \Sigma(\cdot)$, then $\|\Delta_i(t)\| \le C$ for every $t \in [0, T]$ uniformly in i, where $C < \infty$ is a constant independent of i. As in the proof of Proposition 4.1, it can be shown that

$$\Delta_i(t) := (M_i - M) - \int_t^T \{[A(s) + \Sigma(s)Q(s)]\Delta_i(s) + \Delta_i(s)[A(s) + \Sigma(s)Q(s)]'$$
$$+\Delta_i(s)Q(s)\Delta_i(s) - C(s)[I - \Sigma_i(s)(S(s)\Sigma_i(s) + I)^{-1}S(s)]\Delta_i(s)(S(s)\Sigma(s) + I)^{-1}C(s)'\}\, ds.$$

Since $\|\Delta_i(t)\| \le C$, uniformly in i, it follows that

$$\|\Delta_i(t)\| \le \|M_i - M\| + K \int_t^T \|\Delta_i(s)\|\, ds,$$

where $K < \infty$ is a constant independent of i. From Gronwall's inequality, it follows that $\|\Delta_i(t)\| \le \|M_i - M\| e^{KT}$, so $\Delta_i(\cdot) \to 0$ as $i \uparrow \infty$, uniformly on $[0, T]$.

To show convergence of $(h_i(\cdot), \eta_i(\cdot))$ to $(h(\cdot), \eta(\cdot))$, observe that

$$\begin{cases} d(h(t) - h_i(t)) = [(A(t) + \Sigma(t)Q(t))(h(t) - h_i(t)) \\ \qquad\qquad + C(t)(I + \Sigma(t)S(t))^{-1}(\eta(t) - \eta_i(t)) \\ \qquad\qquad + (\Sigma(t) - \Sigma_i(t))Q(t)h_i(t) \\ \qquad\qquad - C(t)(I + \Sigma(t)S(t))^{-1}(\Sigma(t) - \Sigma_i(t))S(t)(I + \Sigma_i(t)S(t))^{-1}\eta_i(t)]\, dt \\ \qquad\qquad + (\eta(t) - \eta_i(t))dW(t), \\ h(T) - h_i(T) = 0. \end{cases}$$

Since $(h(\cdot) - h_i(\cdot), \eta(\cdot) - \eta_i(\cdot))$ is the (unique) solution of a linear BSDE, it follows from [19, Theorem 2.2, p. 349] that

$$\begin{aligned} &E \sup_{t\in[0,\,T]} |h(t) - h_i(t)|^2 + E \int_0^T |\eta(t) - \eta_i(t)|^2\, dt \\ &\le K_1 E \int_0^T |(\Sigma(t) - \Sigma_i(t))Q(t)h_i(t) \\ &\quad -C(t)(I + \Sigma(t)S(t))^{-1}(\Sigma(t) - \Sigma_i(t))S(t)(I + \Sigma_i(t)S(t))^{-1}\eta_i(t)|^2\, dt, \\ &\le K_2 \|\Sigma(\cdot) - \Sigma_i(\cdot)\|^2 E \int_0^T (|h_i(t)|^2 + |\eta_i(t)|^2)\, dt \end{aligned}$$

for some constants K_1, $K_2 < \infty$ (which are independent of i). Finally, since $\Sigma_i(\cdot)$ is uniformly bounded in i, it can be shown (following the proof of [19, Theorem 2.2, p. 349]) that $E\int_0^T |h_i(t)|^2\, dt$ and $E\int_0^T |\eta_i(t)|^2\, dt$ are bounded, uniformly in i. Our result follows from the fact that $\Sigma_i(\cdot) \to \Sigma(\cdot)$ as $i \uparrow \infty$. □

Before we conclude this section, we present a representation result for the solution of the Lyapunov equation (3.6).

PROPOSITION 4.8. *The solution of the Lyapunov equation* (3.6) *is*

$$N(t) = \frac{1}{2}[Z(t)(I + \Sigma(t)Z(t))^{-1} + (I + Z(t)\Sigma(t))^{-1}Z(t)]. \tag{4.15}$$

Proof. Case 1: $H > 0$. In this case $Z(t) > 0$ for every $t \in [0, T]$; see Corollary 4.6. Therefore, $Z(t)^{-1}$ is well defined and (4.15) is equivalent to:

$$N(t) = (Z(t)^{-1} + \Sigma(t))^{-1}. \tag{4.16}$$

Thus it suffices to show that the right-hand side of (4.16) is a solution of (3.6). By evaluating $\frac{d}{dt}\{Z(t)Z(t)^{-1}\} = 0$, it is easy to show that

$$\begin{cases} \frac{d}{dt}\{Z(t)^{-1}\} = A(t)Z(t)^{-1} + Z(t)^{-1}A(t)' \\ \quad +B(t)R(t)^{-1}B(t)' + C(t)\Sigma(t)(I + S(t)\Sigma(t))^{-1}C(t)' - Z(t)^{-1}Q(t)Z(t)^{-1}, \\ Z(0)^{-1} = H^{-1}. \end{cases}$$

Therefore, $N(t)^{-1} = Z(t)^{-1} + \Sigma(t)$ is the solution of the ODE:

$$\begin{cases} \dfrac{d}{dt}\{N(t)^{-1}\} = (A(t) + \Sigma(t)Q(t))N(t)^{-1} + N(t)^{-1}(A(t) + \Sigma(t)Q(t))' \\ \qquad\qquad - N(t)^{-1}Q(t)N(t)^{-1}, \\ N(0)^{-1} = H^{-1} + \Sigma(0). \end{cases}$$

Finally, by evaluating $\frac{d}{dt}\{N(t)N(t)^{-1}\} = 0$, it is easy to show that the right-hand side of (4.16) (and hence that of (4.15)) is a solution of (3.6).

Case 2: $H \geq 0$. Let $Z(\cdot)$, $Z_i(\cdot)$ $(i \in \mathbb{Z}^+)$ be the solutions of (3.5) corresponding to H and $H_i := H + (1/i)I > 0$, respectively, where the $\Sigma(\cdot)$ in the coefficients of (3.5) is the solution to (3.4). Let

$$N_i(t) = \frac{1}{2}[Z_i(t)(I + \Sigma(t)Z_i(t))^{-1} + (I + Z_i(t)\Sigma(t))^{-1}Z_i(t)]. \tag{4.17}$$

It follows from Case 1 that $N_i(\cdot)$ is the unique solution of the Lyapunov equation

$$\begin{cases} \dot{N}_i(t) + N_i(t)(A(t) + \Sigma(t)Q(t)) + (A(t) + \Sigma(t)Q(t))'N_i(t) - Q(t) = 0, \\ N_i(0) = \frac{1}{2}\{H_i(I + \Sigma(0)H_i)^{-1} + (I + H_i\Sigma(0))^{-1}H_i\}. \end{cases}$$

On the other hand, we know from the continuity of solutions of linear ODEs with respect to initial conditions that $N_i(\cdot) \to N(\cdot)$, where $N(\cdot)$ is the solution of the ODE (3.6). Therefore, to prove that $N(\cdot)$ has the representation (4.15), we need only show that $Z_i(\cdot) \to Z(\cdot)$ since this will imply that the right-hand side of (4.17) converges to the right-hand side of (4.15). However, it follows from the fact that (3.5) is a special case of (3.4) (see the proof of Corollary 4.6) and the convergence properties of (3.4) (Proposition 4.7) that $Z_i(\cdot) \to Z(\cdot)$ as $i \uparrow \infty$. This proves our result. □

5. Proofs of Theorems 3.2 and 3.3. In this section we give proofs of the main results of the paper, Theorems 3.2 and 3.3. The basic idea is first to find a lower bound of the cost function (2.3) (see Lemma 5.1), and then to identify a control which achieves *exactly* this lower bound (see Proposition 5.3).

To obtain a lower bound of (2.3), we use the completion-of-squares technique. Consider (4.12)–(4.14) parameterized by $M > 0$. It has been shown in section 4 that these three equations have unique solutions $P_M(\cdot) > 0, \Sigma_M(\cdot) > 0$, and $(h_M(\cdot), \eta_M(\cdot)) \in L^2_{\mathcal{F}}(\Omega; C(0, T; \mathbb{R}^n)) \times L^2_{\mathcal{F}}(0, T; \mathbb{R}^n)$, respectively, for every $M > 0$, and $\Sigma_M(t) = P_M(t)^{-1}$ for all $t \in [0, T]$.

Let $(x(\cdot), z(\cdot))$ be the solution of the BSDE (2.1) corresponding to a control $u(\cdot) \in \mathcal{U}$. Applying Ito's formula to $(x(t) + h_M(t))'P_M(t)(x(t) + h_M(t))$, we obtain

$$\begin{aligned} &d\{(x + h_M)'P_M(x + h_M)\} \\ &= \{(x + h_M)'\,(P_MBR^{-1}B'P_M + P_MC(S + P_M)^{-1}C'P_M - Q)\,(x + h_M) \\ &\quad +2(x + h_M)'\,P_M\,(\Sigma_MQh_M + C(\Sigma_MS + I)^{-1}\eta_M - C\eta_M) \\ &\quad +(z + \eta_M)'P_M(z + \eta_M) + 2(z + \eta_M)'C'P_M(x + h_M) + 2u'B'P_M(x + h_M)\}\,dt \\ &\quad +\{\cdots\}\,dW. \end{aligned}$$

Integrating both sides with respect to t and taking expectations, we arrive at

$$\begin{aligned} 0 &= (x(0) + h_M(0))'P_M(0)(x(0) + h_M(0)) \\ &\quad +E\int_0^T \{(x + h_M)'(P_MBR^{-1}B'P_M + P_MC(S + P_M)^{-1}C'P_M - Q)(x + h_M) \\ &\quad +2(x + h_M)'\,P_M\,(\Sigma_MQh_M + C(\Sigma_MS + I)^{-1}\eta_M - C\eta_M) \\ &\quad +(z + \eta_M)'P_M(z + \eta_M) + 2(z + \eta_M)'C'P_M(x + h_M) + 2u'B'P_M(x + h_M)\}\,dt. \end{aligned} \tag{5.1}$$

Adding (5.1) to the right-hand side of (2.3), we obtain (after some manipulation)

$$\begin{aligned}J(\xi;u(\cdot)) = {}& \frac{1}{2}[x(0)+(\Sigma_M(0)H+I)^{-1}h_M(0)]'[H+P_M(0)][x(0)+(\Sigma_M(0)H+I)^{-1}h_M(0)]\\ &+\frac{1}{2}h_M(0)'(H\Sigma_M(0)+I)^{-1}Hh_M(0)+E\int_0^T\{h_M'Qh_M+\eta_M'(S\Sigma_M+I)^{-1}S\eta_M\}\,dt\\ &+E\frac{1}{2}\int_0^T\{[u+R^{-1}B'P_M(x+h_M)]'R[u+R^{-1}B'P_M(x+h_M)]\\ &+[z+(I+\Sigma_MS)^{-1}\eta_M+\Sigma_M(I+S\Sigma_M)^{-1}C'P_M(x+h_M)]'\\ &\times(S+P_M)[z+(I+\Sigma_MS)^{-1}\eta_M+\Sigma_M(I+S\Sigma_M)^{-1}C'P_M(x+h_M)]\}\,dt.\end{aligned}\tag{5.2}$$

In deriving this expression, we have used the fact that $\Sigma_M(t)=P_M(t)^{-1}$ together with the following simple relations:

$$\begin{aligned}&[H+P_M(0)]^{-1}P_M(0)=[\Sigma_M(0)H+I]^{-1},\\ &P_M(0)-P_M(0)(H+P_M(0))^{-1}P_M(0)=[H\Sigma_M(0)+I]^{-1}H,\\ &z+\eta_M+(I+\Sigma_MS)^{-1}\Sigma_M(C'P_M(x+h_M)-S\eta_M)\\ &\quad=z+(I+\Sigma_MS)^{-1}\eta_M+\Sigma_M(I+S\Sigma_M)^{-1}C'P_M(x+h_M).\end{aligned}$$

Since $H+P_M(0)>0$, $R>0$, and $S+P_M>0$, it follows from (5.2) that

$$\begin{aligned}J(\xi;u(\cdot))\geq{}&h_M(0)'[H\Sigma_M(0)+I]^{-1}Hh_M(0)\\ &+E\int_0^T\{h_M'Qh_M+\eta_M'(S\Sigma_M+I)^{-1}S\eta_M\}dt\end{aligned}\tag{5.3}$$

for any $u(\cdot)\in\mathcal{U}$. Note that the right-hand side of (5.3) depends on $\Sigma_M(\cdot)$ and $(h_M(\cdot),\eta_M(\cdot))$ (but does not depend on $P_M(\cdot)$). Therefore, it is well defined even when $M=0$. Thus we have the following result.

LEMMA 5.1. *We have*

$$\begin{aligned}J(\xi;u(\cdot))\geq{}&h(0)'[H\Sigma(0)+I]^{-1}Hh(0)\\ &+E\int_0^T\{h'Qh+\eta'(S\Sigma+I)^{-1}S\eta\}dt\ \forall u(\cdot)\in\mathcal{U},\end{aligned}\tag{5.4}$$

where $\Sigma(\cdot)$ *and* $(h(\cdot),\eta(\cdot))$ *are the solutions of* (3.4) *and* (3.7)*, respectively.*

Proof. Letting $M\to 0$ in (5.3) and appealing to Proposition 4.7, we obtain the result. □

Lemma 5.1 provides a lower bound on the cost function (2.3). Now we are to find a control that achieves this lower bound. To this end, recall the Hamiltonian system (3.12)–(3.13).

PROPOSITION 5.2. *The Hamiltonian system* (3.12)–(3.13) *has a unique solution* $(x(\cdot),z(\cdot),y(\cdot))$*. Moreover, the following relations are satisfied:*

$$\begin{cases}x(t)=\Sigma(t)y(t)-h(t),\\ z(t)=-\Sigma(t)(S(t)\Sigma(t)+I)^{-1}C(t)'y(t)-(\Sigma(t)S(t)+I)^{-1}\eta(t),\\ x(0)=-(\Sigma(0)H+I)^{-1}h(0),\end{cases}\tag{5.5}$$

where $\Sigma(\cdot)$ *and* $(h(\cdot),\eta(\cdot))$ *are the solutions of* (3.4) *and* (3.7)*, respectively.*

Proof. We begin by proving existence. Consider the following SDE:

$$\text{(5.6)}\quad \begin{cases} d\bar{y}(t) = [-(A(t)+\Sigma(t)Q(t))'\bar{y}(t) + Q(t)h(t)]\,dt \\ \quad\quad +[-(I+S(t)\Sigma(t))^{-1}C(t)'\bar{y}(t) + S(t)(I+\Sigma(t)S(t))^{-1}\eta(t)]\,dW(t), \\ \bar{y}(0) = H(\Sigma(0)H+I)^{-1}h(0). \end{cases}$$

Since (5.6) is a linear SDE with bounded coefficients and square integrable nonhomogeneous terms, it follows that it has a unique solution $\bar{y}(\cdot)$. On the other hand, we can define

$$\text{(5.7)}\qquad \bar{x}(t) := \Sigma(t)\bar{y}(t) - h(t).$$

By applying Ito's formula to (5.7), we obtain

$$\text{(5.8)}\quad \begin{cases} d\bar{x}(t) = [A(t)\bar{x}(t) - B(t)R(t)^{-1}B(t)'\bar{y}(t) \\ \quad\quad +C(t)(-(\Sigma(t)S(t)+I)^{-1}\eta(t) - \Sigma(t)(I+S(t)\Sigma(t))^{-1}C(t)'\bar{y}(t))]\,dt \\ \quad\quad +[-(\Sigma(t)S(t)+I)^{-1}\eta(t) - \Sigma(t)(I+S(t)\Sigma(t))^{-1}C(t)'\bar{y}(t)]\,dW(t), \\ \bar{x}(0) = -(\Sigma(0)H+I)^{-1}h(0). \end{cases}$$

Substituting

$$\text{(5.9)}\qquad \bar{z}(t) := -(\Sigma(t)S(t)+I)^{-1}\eta(t) - \Sigma(t)(I+S(t)\Sigma(t))^{-1}C(t)'\bar{y}(t)$$

into (5.8) and noting (from (5.7)) that $\bar{x}(T) = \xi$, it follows that

$$\text{(5.10)}\quad \begin{cases} d\bar{x}(t) = [A(t)\bar{x}(t) - B(t)R(t)^{-1}B(t)'\bar{y}(t) + C(t)\bar{z}(t)]\,dt + \bar{z}(t)\,dW(t), \\ \bar{x}(T) = \xi. \end{cases}$$

On the other hand, it follows from (5.9) that

$$\begin{aligned} &-C(t)'\bar{y}(t) - S(t)\bar{z}(t) \\ &= S(t)(\Sigma(t)S(t)+I)^{-1}\eta(t) \\ &\quad +[S(t)\Sigma(t) - (S(t)\Sigma(t)+I)](S(t)\Sigma(t)+I)^{-1}C(t)'\bar{y}(t) \\ &= S(t)(\Sigma(t)S(t)+I)^{-1}\eta(t) - (S(t)\Sigma(t)+I)^{-1}C(t)'\bar{y}(t). \end{aligned} \qquad \text{(5.11)}$$

Finally, substituting (5.7) and (5.11) into (5.6) and noting (from (5.8)) the initial value of $\bar{x}(0)$, it follows that $\bar{y}(t)$ is a solution of the differential equation

$$\text{(5.12)}\quad \begin{cases} d\bar{y}(t) = \{-A(t)'\bar{y}(t) - Q(t)\bar{x}(t)\}dt + \{-C(t)'\bar{y}(t) - S(t)\bar{z}(t)\}dW(t), \\ \bar{y}(0) = -H\bar{x}(0). \end{cases}$$

That is, $(\bar{x}(\cdot), \bar{z}(\cdot), \bar{y}(\cdot))$ is a solution of the system of equations (5.10), (5.12) and hence a solution of the Hamiltonian system (3.12)–(3.13). In addition, by virtue of (5.7), (5.8), and (5.9), the relations (5.5) are satisfied.

To prove uniqueness, suppose that $(x_1(\cdot), z_1(\cdot), y_1(\cdot))$ and $(x_2(\cdot), z_2(\cdot), y_2(\cdot))$ are solutions of (3.12)–(3.13). It follows that $(x(\cdot), z(\cdot), y(\cdot)) := (x_1(\cdot) - x_2(\cdot), z_1(\cdot) - z_2(\cdot), y_1(\cdot) - y_2(\cdot))$ is a solution of the Hamiltonian system

$$\text{(5.13)}\quad \begin{cases} dx(t) = \{A(t)x(t) - B(t)R(t)^{-1}B(t)'y(t) + C(t)z(t)\}dt + z(t)dW(t), \\ x(T) = 0, \end{cases}$$

$$\text{(5.14)}\quad \begin{cases} dy(t) = \{-A(t)'y(t) - Q(t)x(t)\}dt + \{-C(t)'y(t) - S(t)z(t)\}dW(t), \\ y(0) = -Hx(0). \end{cases}$$

By Ito's formula, we have

$$d\{x(t)'y(t)\} = \{ - x(t)'Q(t)x(t) - y(t)'B(t)R(t)^{-1}B(t)'y(t) - z(t)'S(t)z(t)\}dt + \{\cdots\}dW(t).$$

Integrating both sides from 0 to T and taking expectations, we obtain

$$x(0)'Hx(0) = -E\int_0^T (x(t)'Q(t)x(t) + y(t)'B(t)R(t)^{-1}B(t)'y(t) + z(t)'S(t)z(t))dt.$$

Since H, $Q(\cdot)$, $R(\cdot)$, and $S(\cdot)$ are all positive semidefinite (see (A1)), it follows that

$$E\int_0^T (x(t)'Q(t)x(t) + y(t)'B(t)R(t)^{-1}B(t)'y(t) + z(t)'S(t)z(t))dt = 0.$$

Finally, since $R(\cdot) > 0$, it follows that

$$B(t)'y(t) = 0, \quad \text{a.e. } t \in [0, T],\ P\text{-a.s..}$$

Substituting this into (5.13), it follows that $(x(\cdot), z(\cdot))$ is the solution of the linear BSDE:

$$\begin{cases} dx(t) = \{A(t)x(t) + C(t)z(t)\}dt + z(t)dW(t), \\ x(T) = 0, \end{cases}$$

and the uniqueness of solutions for linear BSDEs implies that $(x(\cdot), z(\cdot)) \equiv 0$. Substituting $(x(\cdot), z(\cdot)) \equiv 0$ into (5.14), it follows that $y(\cdot)$ is a solution of linear SDE

$$\begin{cases} dy(t) = -A(t)'y(t)dt - C(t)'y(t)dW(t), \\ y(0) = -Hx(0). \end{cases}$$

Hence it follows from the uniqueness again that $y(\cdot) \equiv 0$. This proves our result. □

Remark 5.1.We have shown, in fact, that $(x(\cdot), z(\cdot), y(\cdot))$ is the solution of the Hamiltonian system (3.12)–(3.13) if and only if $y(\cdot)$ is the solution of (5.6), $x(\cdot)$ is the solution of (5.8), and $z(\cdot)$ satisfies (5.9). This means that we may use the Hamiltonian system (3.12)–(3.13) or (5.6), (5.8), and (5.9) interchangeably to describe the processes $y(\cdot)$, $x(\cdot)$, and $z(\cdot)$. This is an important observation which simplifies much of our subsequent analysis.

PROPOSITION 5.3. *Let $(x(\cdot), z(\cdot), y(\cdot))$ be the solution of the Hamiltonian system* (3.12)–(3.13)*, and let $u(\cdot)$ be given by*

$$u(t) = -R(t)^{-1}B(t)'y(t). \tag{5.15}$$

Then $(x(\cdot), z(\cdot))$ is the solution of the BSDE (2.1) *corresponding to* (5.15) *and*

$$\begin{aligned} J(\xi; u(\cdot)) =& h(0)'[H\Sigma(0) + I]^{-1}Hh(0) \\ &+ E\int_0^T \{h'Qh + \eta'(S\Sigma + I)^{-1}S\eta\}\, dt \end{aligned} \tag{5.16}$$

is the associated cost.

Proof. Let $(x(\cdot), z(\cdot), y(\cdot))$ be the solution of the Hamiltonian system (3.12)–(3.13), and let $u(\cdot)$ be given by (5.15). It follows from Remark 5.1 that $y(\cdot)$ is also the unique solution of the SDE (5.6). Regarding $y(\cdot)$ in this way, it follows that $(x(\cdot), z(\cdot))$ (as determined from (3.12)–(3.13)) is also the solution of BSDE (2.1) when $u(\cdot)$ is given by (5.15).

To determine the cost associated with the control (5.15), we shall use the fact that $y(\cdot)$ is also the unique solution of the linear SDE (5.6). By Ito's formula, it can be shown that

$$\begin{aligned} d\{y'\,\Sigma\, y\} &= \{-y'[\Sigma Q\Sigma + BR^{-1}B' + C(\Sigma S + I)^{-1}\Sigma S\Sigma(S\Sigma + I)^{-1}C']y \\ &\quad -2\eta'(S\Sigma + I)^{-1}S\Sigma(S\Sigma + I)^{-1}C'y + 2y'\Sigma Qh \\ &\quad +\eta'(S\Sigma + I)^{-1}S\Sigma S(\Sigma S + I)^{-1}\eta\}\,dt + \{\cdots\}\,dW. \end{aligned}$$

Therefore,

$$\begin{aligned} &E\int_0^T \{y'[\Sigma Q\Sigma + BR^{-1}B' + C(\Sigma S + I)^{-1}\Sigma S\Sigma(S\Sigma + I)^{-1}C']\,y \\ &\quad +2\eta'(S\Sigma + I)^{-1}S\Sigma(S\Sigma + I)^{-1}C'y - 2y'\Sigma Qh\}\,dt \\ &= h(0)'(I + H\Sigma(0))^{-1}H\Sigma(0)H(I + \Sigma(0)H)^{-1}h(0) \\ &\quad +E\int_0^T \eta'(S\Sigma + I)^{-1}S\Sigma S(\Sigma S + I)^{-1}\eta\,dt. \end{aligned} \tag{5.17}$$

On the other hand, the cost associated with the control (5.15) can be obtained by substituting (5.5) and (5.15) into (2.3). By doing this, we obtain

$$\begin{aligned} J(\xi; u(\cdot)) &= h(0)'(I + H\Sigma(0))^{-1}H(I + \Sigma(0)H)^{-1}h(0) \\ &\quad +E\int_0^T \{\eta'(I + S\Sigma)^{-1}S(I + \Sigma S)^{-1}\eta + h'Qh \\ &\quad +y'[\Sigma Q\Sigma + BR^{-1}B' + C(\Sigma S + I)^{-1}\Sigma S\Sigma(S\Sigma + I)^{-1}C']\,y \\ &\quad +2\eta'(S\Sigma + I)^{-1}S\Sigma(S\Sigma + I)^{-1}C'y - 2y'\Sigma Qh\}\,dt. \end{aligned}$$

Hence it follows from (5.17) that

$$\begin{aligned} J(\xi; u(\cdot)) &= h(0)'(I + H\Sigma(0))^{-1}(H + H\Sigma(0)H)(I + \Sigma(0)H)^{-1}h(0) \\ &\quad +E\int_0^T \{h'Qh + \eta'(S\Sigma + I)^{-1}(S + S\Sigma S)(I + \Sigma S)^{-1}\eta\}\,dt \\ &= h(0)'[H\Sigma(0) + I]^{-1}Hh(0) + E\int_0^T \{h'Qh + \eta'(S\Sigma + I)^{-1}S\eta\}\,dt, \end{aligned}$$

which is precisely the expression (5.16). □

Proof of Theorem 3.3. The unique solvability of the Hamiltonian system (3.12)–(3.13) has been proved in Proposition 5.2. The optimality of (5.15) follows from the fact that the cost (5.16) associated with the control (5.15) is equal to a lower bound to the optimal cost; see Lemma 5.1. The expression (3.11) for the optimal cost can be obtained by applying Ito's formula to $h(t)'N(t)h(t)$. By doing this, we obtain the

relation

$$\begin{aligned} &h(0)'H(I+\Sigma(0)H)^{-1}h(0)+E\int_0^T\{h'Qh+\eta'(S\Sigma+I)^{-1}S\eta\}\,dt\\ &\quad=E\xi'N(T)\xi+E\int_0^T\{\eta(t)'[(S(t)\Sigma(t)+I)^{-1}S(t)-N(t)]\eta(t)\\ &\qquad-2\eta(t)'(I+S(t)\Sigma(t))^{-1}C(t)'N(t)h(t)\}\,dt. \end{aligned} \tag{5.18}$$

This yields the optimal cost (3.11). Finally, we are able to conclude that the control (5.15) is unique because the BLQ problem (2.4) is a (strictly) convex optimization problem. The set of admissible triples $(x(\cdot), z(\cdot), u(\cdot))$ associated with (2.1) is a convex set, and the cost (2.3) is a strictly convex function on this set. □

Now we proceed to prove Theorem 3.2. First we have the following lemma.

LEMMA 5.4. *Let $(x(\cdot), z(\cdot), y(\cdot))$ be the solution of the Hamiltonian system* (3.12)–(3.13)*, and let $q(\cdot)$ be the solution of the SDE* (3.8)*. Then*

$$y(t)=-Z(t)x(t)-q(t). \tag{5.19}$$

Proof. Let $(x(\cdot), z(\cdot), y(\cdot))$ denote the solution of the Hamiltonian system (3.12)–(3.13). We have already shown that $x(\cdot)=\bar{x}(\cdot)$, where $\bar{x}(\cdot)$ is the solution of the SDE (5.8); see also Remark 5.1. Therefore, we can prove (5.19) by showing that

$$y(t)=-Z(t)\bar{x}(t)-q(t). \tag{5.20}$$

Let $y(\cdot)$, $\bar{x}(\cdot)$, and $Z(\cdot)$ be solutions of (5.6), (5.8), and (3.5), respectively. Also, let us assume for the time being that $q(\cdot)$ is the unique solution of the SDE

$$\begin{cases} dq(t)=\{-[A(t)+B(t)R(t)^{-1}B(t)'Z(t)\\ \qquad\qquad +C(t)(I+\Sigma(t)S(t))^{-1}\Sigma(t)C(t)'Z(t)]'\,q(t)\\ \qquad +Z(t)C(t)(I+\Sigma(t)S(t))^{-1}\eta(t)\}dt+\{(Z(t)-S(t))(I+\Sigma(t)S(t))^{-1}\eta(t)\\ \qquad +(I+Z(t)\Sigma(t))(I+S(t)\Sigma(t))^{-1}C(t)'y(t)\}\,dW(t),\\ q(0)=0. \end{cases} \tag{5.21}$$

(Note that (3.8) and (5.21) differ only in their diffusion terms. It will be shown later that (3.8) and (5.21) have the same solution. In the meantime, however, it will be easier to work with (5.21)). Finally, by virtue of (5.5), $y(\cdot)$ is also the unique solution of the SDE

$$\begin{cases} dy(t)=\{-A(t)'y(t)-Q(t)\bar{x}(t)\}\,dt\\ \qquad +\{S(t)(I+\Sigma(t)S(t))^{-1}\eta(t)-(I+S(t)\Sigma(t))^{-1}C(t)'y(t)\}\,dW(t),\\ y(0)=H(\Sigma(0)H+I)^{-1}h(0). \end{cases} \tag{5.22}$$

With $q(\cdot)$ denoting the solution of (5.21), it is easy to show (using Ito's formula) that

$$\begin{cases} d\{y(t)+Z(t)\bar{x}(t)+q(t)\}\\ \qquad =[A(t)+B(t)R(t)^{-1}B(t)'Z(t)+C(t)(I+\Sigma(t)S(t))^{-1}\Sigma(t)C(t)'Z(t)]'\\ \qquad \times(y(t)+Z(t)\bar{x}(t)+q(t))\,dt,\\ y(0)+Z(0)\bar{x}(0)+q(0)=0. \end{cases} \tag{5.23}$$

Hence it follows from the uniqueness of solutions of linear SDEs that

$$y(t) + Z(t)x(t) + q(t) = y(t) + Z(t)\bar{x}(t) + q(t) = 0 \quad \forall\, t \in [0,\, T],\ P\text{-a.s.}, \tag{5.24}$$

where $q(\cdot)$ is the solution of (5.21). Finally, substituting (5.7) into (5.24), it is easy to show that

$$y(t) = (I + Z(t)\Sigma(t))^{-1}(Z(t)h(t) - q(t)). \tag{5.25}$$

Substituting (5.25) into (5.21), it follows that $q(\cdot)$ is the unique solution of both (5.21) and (3.8). □

Proof of Theorem 3.2. It follows immediately from Proposition 5.3 and Lemma 5.4. □

Finally, it is important to recognize that the expressions for the optimal control, as presented in Theorems 3.3 and 3.2, are equivalent expressions of the same process; that is, this does not contradict the uniqueness of optimal controls for (2.4).

6. Origin of idea: Forward formulation. In section 5, we obtained the solution of the BLQ problem (2.4) by showing that the control (3.14) or (3.9) achieves a lower bound to the cost function. In showing this result, (3.4)–(3.8), especially the Riccati equations (3.4) and (3.5), play a crucial role. In other words, once these equations are in place, then the whole derivation, albeit quite tedious, is essentially in the same spirit as the completion-of-squares technique commonly used in tackling forward LQ problems. However, the reader may be puzzled about how these (rather complicated) equations were obtained in the first place. This section serves to unfold the origin of those equations by presenting an alternative and intuitively appealing approach to the BLQ problem (2.4). The idea is basically inspired by [15, 12], where an (uncontrolled) BSDE is regarded as a *controlled forward* SDE. Here we go one step further to show that the BLQ problem can also be viewed as a (constrained) forward LQ problem, and that the solution (5.15) of the BLQ problem and the relationships (5.5) coincide with the limiting solution of a sequence of unconstrained forward LQ problems. In this process, the Riccati equations (3.4) and (3.5), along with other related equations, come out very naturally. It should be noted that our aim in this section is to highlight the origin of (3.4)–(3.8) as well as (5.5), and hence the material in this section will be presented in an informal way. For this reason, certain convergence results required in this derivation, for example, are taken for granted, although they can be verified rigorously using standard techniques from stochastic analysis, the details of which are left to the interested reader. (As a matter of fact, the first version of this paper *was* written rigorously using the forward formulation, but then we went for the current version finding that the presentation would be much simpler once all the necessary equations were identified.) Finally, for the sake of notational convenience, we shall assume throughout this section that $S = 0$. The extension to the case $S \geq 0$ can be obtained in a similar way.

Forward LQ problem.

Consider the following SDE:

$$\begin{cases} dx(t) &= \{A(t)x(t) + B(t)u(t) + C(t)v(t)\}dt + v(t)\, dW(t), \\ x(0) &= x^0. \end{cases} \tag{6.1}$$

We assume throughout that $x^0 \in \mathbb{R}^n$ and $(u(\cdot),\, v(\cdot)) \in \bar{\mathcal{U}}$, where

$$\bar{\mathcal{U}} = L^2_{\mathcal{F}}(0,\, T;\, \mathbb{R}^m) \times L^2_{\mathcal{F}}(0,\, T;\, \mathbb{R}^n).$$

For every $i \in \mathbb{Z}^+$ let

$$\begin{aligned} &J(x^0,\, u(\cdot),\, v(\cdot);\, i) \\ &:= E\frac{1}{2}\{x^{0\prime} H x^0 + \int_0^T (x(t)'Q(t)x(t) + u(t)'R(t)u(t))\, dt + i\,|x(T) - \xi|^2\}. \end{aligned} \tag{6.2}$$

The family of LQ problems, parameterized by i, is defined by

$$\begin{cases} \min_{x^0,\,(u(\cdot),\,v(\cdot))} J(x^0,\, u(\cdot),\, v(\cdot);\, i) \\ \text{subject to} \\ \quad (u(\cdot),\, v(\cdot)) \in \bar{\mathcal{U}},\ x^0 \in \mathbb{R}^n, \\ \quad (x^0,\, x(\cdot),\, u(\cdot),\, v(\cdot)) \text{ satisfies (6.1).} \end{cases} \tag{6.3}$$

Comparing (6.3) with the BLQ problem (2.4), it is clear that the control $v(\cdot)$ replaces the process $z(\cdot)$ in the BSDE, while the terminal condition $x(T) = \xi$ in (2.4) is replaced by a penalty term in the cost of the forward problem (6.3). One fundamental difference between (2.4) and (6.3) should be recognized. In the BLQ problem (2.4), the initial condition $x(0)$ and the process $z(\cdot)$ are part of the state process $(x(\cdot),\, z(\cdot))$; that is, once $u(\cdot)$ has been chosen, the pair $(x(\cdot),\, z(\cdot))$ (and hence $x(0)$) is uniquely determined. On the other hand, the pair $(u(\cdot),\, v(\cdot))$ and the initial condition $x(0)$ are decision variables in the forward problem (6.3). This additional degree of freedom is possible because the forward problem (6.3) does not involve a terminal condition on the state $x(\cdot)$. We shall show that the optimal solution of the BLQ problem (2.4), as stated in Theorems 3.3 and 3.2, can be obtained by solving the forward problem (6.3) and letting $i \uparrow \infty$.

Completion of squares. The solution of the forward problem (6.3) can be obtained by using a completion-of-squares approach via the Riccati equation studied in [5]. In particular, let $P_i(\cdot)$ and $(h_i(\cdot),\, \eta_i(\cdot))$ be the unique solutions of the following equations:

$$\begin{cases} \dot{P}_i(t) + P_i(t)A(t) + A(t)'P_i(t) \\ \quad -P_i(t)[B(t)R(t)^{-1}B(t)' + C(t)P_i(t)^{-1}C(t)']P_i(t) + Q(t) = 0, \\ P_i(T) = i\, I, \end{cases} \tag{6.4}$$

$$\begin{cases} dh_i(t) = \{(A(t) + P_i(t)^{-1}Q(t))h_i(t) + C(t)\,\eta_i(t)\}\, dt + \eta_i(t)\, dW(t), \\ h_i(T) = -\xi. \end{cases} \tag{6.5}$$

Note that (6.5) is introduced to cope with the linear term $E\{\, i\xi x(T)\}$ in the terminal cost part of (6.2). Evaluating $\Sigma_i(t) := P_i(t)^{-1}$, it turns out that $\Sigma_i(\cdot)$ is a solution of the Riccati equation

$$\begin{cases} \dot{\Sigma}_i(t) = A(t)\Sigma_i(t) + \Sigma_i(t)A(t)' - C(t)\Sigma_i(t)C(t)' \\ \qquad\qquad + \Sigma_i(t)Q(t)\Sigma_i(t) - B(t)R(t)^{-1}B(t)', \\ \Sigma_i(T) = \frac{1}{i} I. \end{cases} \tag{6.6}$$

(The above explains the origin of the key equations (4.1), (4.2), and (4.14).) Applying Ito's formula to $(x(t) + h_i(t))'P_i(t)(x(t) + h_i(t))$, it can be shown that

$$\begin{aligned} 0 &= (x(0) + h_i(0))'P_i(0)(x(0) + h_i(0)) \\ &\quad + E\int_0^T \{(x + h_i)'(P_iBR^{-1}B'P_i + P_iCP_i^{-1}C'P_i + Q)(x + h_i) + 2(x + h_i)'\Sigma_i Q h_i \\ &\quad + (z + \eta_i)'P_i(z + \eta_i) + 2(z + \eta_i)'C'P_i(x + h_i) + 2u'B'P_i(x + h_i)\}\, dt. \end{aligned}$$

Adding this to the right-hand side of the cost (6.2), we obtain (after some manipulation) that

$$\begin{aligned} J(x^0,\,u(\cdot),\,v(\cdot);\,i) &= \frac{1}{2}h_i(0)'(H\Sigma_i(0)+I)^{-1}Hh_i(0) + E\frac{1}{2}\int_0^T h_i(t)'Q(t)h_i(t)\,dt \\ &+\frac{1}{2}[x^0+(I+\Sigma_i(0)H)^{-1}h_i(0)]'[H+P_i(0)][x^0+(I+\Sigma_i(0)H)^{-1}h_i(0)] \\ &+E\frac{1}{2}\int_0^T [(u+R^{-1}B'P_i(x+h_i))'R(u+R^{-1}B'P_i(x+h_i)) \\ &+(v+\Sigma_iC'P_i(x+h_i)+\eta_i)'P_i(v+\Sigma_iC'P_i(x+h_i)+\eta_i)]\,dt. \end{aligned} \tag{6.7}$$

It is interesting to observe that the expression (6.7) for the cost of the forward LQ problem is similar to the expression (5.2) for the backwards LQ cost. Nevertheless, they are fundamentally different in that $(u(\cdot),\,v(\cdot))$ and $x(0)$ are free to be chosen in (6.7), while $x(0)$ and $z(\cdot)$ are uniquely determined in (5.2) once $u(\cdot)$ has been chosen. Clearly, the optimal cost for the forward LQ problem (6.3) is

$$J_i^*(\xi) = \frac{1}{2}h_i(0)'(H\Sigma_i(0)+I)^{-1}Hh_i(0) + E\frac{1}{2}\int_0^T h_i(t)'Q(t)h_i(t)\,dt, \tag{6.8}$$

which is obtained when

$$\begin{cases} u_i(t) = -R(t)^{-1}B(t)'P_i(t)(x_i(t)+h_i(t)), \\ v_i(t) = -\Sigma_i(t)C(t)'P_i(t)(x_i(t)+h_i(t)) - \eta_i(t), \\ x_i^0 = -(I+\Sigma_i(0)H)^{-1}h_i(0), \end{cases} \tag{6.9}$$

where $x_i(\cdot) \in L^2_{\mathcal{F}}(\Omega;\,C(0,\,T;\,\mathbb{R}^n))$, the optimal state trajectory, is the unique solution of the SDE

$$\begin{cases} dx_i(t) &= \{A(t)x_i(t) - B(t)R(t)^{-1}B(t)'P_i(t)(x_i(t)+h_i(t)) \\ &\quad -C(t)P_i(t)^{-1}C(t)'P_i(t)(x_i(t)+h_i(t)) - C(t)\eta_i(t)\}\,dt \\ &\quad -\{P_i(t)^{-1}C(t)'P_i(t)(x_i(t)+h_i(t)) + \eta_i(t)\}\,dW(t), \\ x_i(0) &= x_i^0. \end{cases} \tag{6.10}$$

Limiting solution: $i \uparrow \infty$. Let $y_i(\cdot)$ be defined by the relation

$$y_i(t) := P_i(t)(x_i(t)+h_i(t)). \tag{6.11}$$

It follows that

$$x_i(t) = \Sigma_i(t)y_i(t) - h_i(t), \tag{6.12}$$

where $x_i(\cdot)$ is the solution of (6.10). It is easy to show that

$$\begin{cases} dx_i(t) = \{A(t)x_i(t) - B(t)R(t)^{-1}B(t)'y_i(t) \\ \qquad\qquad + C(t)(-\Sigma_i(t)C(t)'y_i(t) - \eta_i(t))\}\,dt \\ \qquad\qquad +\{-\Sigma_i(t)C(t)'y_i(t) - \eta_i(t)\}\,dW(t), \\ x_i(0) = x_i^0 \end{cases} \tag{6.13}$$

$$\begin{cases} dy_i(t) = \{-(A(t)+\Sigma_i(t)Q(t))'y_i(t) + Q(t)h_i(t)\}\,dt \\ \qquad\qquad - C(t)'y_i(t)\,dW(t), \\ y_i(0) = H(\Sigma_i(0)H+I)^{-1}h_i(0). \end{cases} \tag{6.14}$$

Substituting (6.11) into (6.6), (6.8)–(6.10) and letting $i \uparrow \infty$, we obtain

$$\begin{cases} u(t) = -R(t)^{-1}B(t)'y(t), \\ x(t) = \Sigma(t)y(t) - h(t), \\ v(t) = -\Sigma(t)C(t)'y(t) - \eta(t), \\ x^0 = -(I + \Sigma(0)H)^{-1}h(0) \end{cases} \tag{6.15}$$

and

$$\begin{aligned} J^*(\xi) &= \frac{1}{2}h(0)'(H\Sigma(0) + I)^{-1}Hh(0) + E\frac{1}{2}\int_0^T h(t)'Q(t)h(t)\,dt \\ &= E\frac{1}{2}\left\{\xi' N(T)\xi - \int_0^T (\eta' N\eta + 2h'NC\eta)\,dt\right\}, \end{aligned} \tag{6.16}$$

where

$$\begin{cases} \dot{\Sigma}(t) = A(t)\Sigma(t) + \Sigma(t)A(t)' - C(t)\Sigma(t)C(t)' \\ \qquad\qquad + \Sigma(t)Q(t)\Sigma(t) - B(t)R(t)^{-1}B(t)', \\ \Sigma(T) = 0, \end{cases} \tag{6.17}$$

$$\begin{cases} \dot{N}(t) + N(t)(A(t) + \Sigma(t)Q(t)) + (A(t) + \Sigma(t)Q(t))'N(t) - Q(t) = 0, \\ N(0) = \frac{1}{2}\{H(I + \Sigma(0)H)^{-1} + (I + H\Sigma(0))^{-1}H\}, \end{cases} \tag{6.18}$$

$$\begin{cases} dh(t) = \{(A(t) + \Sigma(t)Q(t))h(t) + C(t)\,\eta(t)\}\,dt + \eta(t)\,dW(t), \\ h(T) = -\xi, \end{cases} \tag{6.19}$$

$$\begin{cases} dx(t) = \{A(t)x(t) - B(t)R(t)^{-1}B(t)'y(t) \\ \qquad\qquad + C(t)(-\Sigma(t)C(t)'y(t) - \eta(t))\}\,dt \\ \qquad\qquad + \{-\Sigma(t)C(t)'y(t) - \eta(t)\}\,dW(t), \\ x(0) = x^0, \end{cases} \tag{6.20}$$

$$\begin{cases} dy(t) = \{-(A(t) + \Sigma(t)Q(t))'y(t) + Q(t)h(t)\}\,dt - C(t)'y(t)\,dW(t), \\ y(0) = H(\Sigma(0)H + I)^{-1}h(0), \end{cases} \tag{6.21}$$

the second equality in (6.16) being obtained by using the identity (5.18). The Hamiltonian system (3.12)–(3.13) is obtained by substituting (6.15) into (6.20)–(6.21), together with the observation that

$$x(T) = \Sigma(T)y(T) - h(T) = \xi.$$

The optimal control (3.14), the optimal cost (3.11), and the relations (5.5) are recovered in (6.15)–(6.16). Hence the solution of the optimal BLQ control problem (2.4) as outlined in Theorem 3.3 coincides with the limiting solution of a family (6.3) of forward LQ problems. Theorem 3.2 can be recovered simply by applying the transformation as outlined in Lemma 5.4.

On the other hand, it is not surprising that the forward approach recovers the solution of the BLQ problem. In particular, it is clear that if

$$J(x^0, u(\cdot), v(\cdot)) = E\frac{1}{2}\left\{x^{0\prime}Hx^0 + \int_0^T (x(t)'Q(t)x(t) + u(t)'R(t)u(t))\,dt\right\}, \tag{6.22}$$

then the problem

$$\text{(6.23)} \qquad \begin{cases} J^* := \min_{x^0,\,(u(\cdot),\,v(\cdot))} J(x^0,\,u(\cdot),\,v(\cdot)) \\ \text{subject to} \\ \quad E\frac{1}{2}|x(T)-\xi|^2 = 0, \\ \quad (u(\cdot),\,v(\cdot)) \in \bar{\mathcal{U}},\ x^0 \in \mathbb{R}^n, \\ \quad (x^0,\,x(\cdot),\,u(\cdot),\,v(\cdot)) \text{ is admissible for (6.1)} \end{cases}$$

is equivalent to the BLQ problem (2.4). Moreover, the solution of (6.23) can be obtained by using a penalty function approach which coincides precisely with the unconstrained problem (6.3). This provides an alternative approach to (2.4), which, as mentioned, was our original idea for solving the BLQ problem. The details are left to the interested readers.

7. Conclusion. In this paper, the optimal control for the BLQ control problem is derived explicitly in terms of a pair of Riccati equations, a forward SDE, and a BSDE. Moreover, this optimal control coincides with the solution of a constrained forward LQ problem and is the limiting solution of a family of unconstrained forward LQ problems. A key part of our derivation is a proof of the existence and uniqueness of solutions of the Riccati equations. Although of independent interest, this proof of global solvability has direct relevance to the BLQ problem since the Riccati equations play a central role in the analysis.

An outstanding open problem to study is the BLQ problem where all the coefficients are random. In this case, the Riccati equations (3.4) and (3.5) both become (nonlinear) BSDEs (rather than ODEs as in this paper), the solvability of which is very challenging to prove.

Acknowledgments. The authors would like to thank the two anonymous referees for their careful reading and constructive comments that led to an improved version of the paper.

REFERENCES

[1] M. Ait Rami and X. Y. Zhou, *Linear matrix inequalities, Riccati equations, and indefinite stochastic linear quadratic control*, IEEE Trans. Automat. Control, 45 (2000), pp. 1131–1143.

[2] B. D. O. Anderson and J. B. Moore, *Optimal Control—Linear Quadratic Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[3] A. Bensoussan, *Lectures on stochastic control. Part* I, in Nonlinear Filtering and Stochastic Control (Cortona, 1981), Lecture Notes in Math. 972, Springer-Verlag, Berlin, New York, 1982, pp. 1–39.

[4] J.-M. Bismut, *An introductory approach to duality in optimal stochastic control*, SIAM Rev., 20 (1978), pp. 62–78.

[5] S. Chen, X. J. Li, and X. Y. Zhou, *Stochastic linear quadratic regulators with indefinite control weight costs*, SIAM J. Control Optim., 36 (1998), pp. 1685–1702.

[6] S. Chen and X. Y. Zhou, *Stochastic linear quadratic regulators with indefinite control weight costs.* II, SIAM J. Control Optim., 39 (2000), pp. 1065–1081.

[7] J. Cvitanić and J. Ma, *Hedging options for a large investor and forward-backward SDEs*, Ann. Appl. Probab., 6 (1996), pp. 370–398.

[8] N. G. Dokuchaev and X. Y. Zhou, *Stochastic control problems with terminal contingent conditions*, J. Math. Anal. Appl., 238 (1999), pp. 143–165.

[9] D. Duffie and L. Epstein, *Stochastic differential utility*, Econometrica, 60 (1992), pp. 353–394.

[10] D. Duffie, J. Ma, and J. Yong, *Black's consol rate conjecture*, Ann. Appl. Probab., 5 (1995), pp. 356–382.

[11] N. El Karoui, S. Peng, and M. C. Quenez, *Backward stochastic differential equations in finance*, Math. Finance, 7 (1997), pp. 1–71.
[12] M. Kohlmann and X. Y. Zhou, *Relationship between backward stochastic differential equations and stochastic controls: A linear-quadratic approach*, SIAM J. Control Optim., 38 (2000), pp. 1392–1407.
[13] A. E. B. Lim and X. Y. Zhou, *Mean-variance portfolio selection with random parameters*, Math. Oper. Res., to appear.
[14] A. E. B. Lim and X. Y. Zhou, *Stochastic optimal LQR control with integral quadratic constraints and indefinite control weights*, IEEE Trans. Automat. Control, 44 (1999), pp. 359–369.
[15] J. Ma and J. Yong, *Forward-Backward Stochastic Differential Equations and Their Applications*, Lecture Notes in Math. 1702, Springer-Verlag, New York, 1999.
[16] E. Pardoux and S. Peng, *Adapted solution of backward stochastic differential equation*, Systems Control Lett., 14 (1990), pp. 55–61.
[17] S. Peng, *A general stochastic maximum principle for optimal control problems*, SIAM J. Control Optim., 28 (1990), pp. 966–979.
[18] S. Peng, *Backward stochastic differential equations and applications to optimal control*, Appl. Math. Optim., 27 (1993), pp. 125–144.
[19] J. Yong and X. Y. Zhou, *Stochastic Controls: Hamiltonian Systems and HJB Equations*, Springer-Verlag, New York, 1999.
[20] X. Y. Zhou, *A unified treatment of maximum principle and dynamic programming in stochastic controls*, Stochastics Stochastics Rep., 36 (1991), pp. 137–161.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 2, pp. 475–495

FLATNESS OF HEAVY CHAIN SYSTEMS*

NICOLAS PETIT† AND PIERRE ROUCHON†

Abstract. In this paper the *flatness* [M. Fliess, J. Lévine, P. Martin, and P. Rouchon, *Internat. J. Control*, 61 (1995), pp. 1327–1361, M. Fliess, J. Lévine, P. Martin, and P. Rouchon, *IEEE Trans. Automat. Control*, 44 (1999), pp. 922–937] of heavy chain systems, i.e., trolleys carrying a fixed length heavy chain that may carry a load, is addressed in the partial derivatives equations framework. We parameterize the system trajectories by the trajectories of its free end and solve the motion planning problem, namely, steering from one state to another state. When considered as a finite set of small pendulums, these systems were shown to be flat [R. M. Murray, in *Proceedings of the IFAC World Congress*, San Francisco, CA, 1996, pp. 395–400]. Our study is an extension to the infinite dimensional case.

Under small angle approximations, these heavy chain systems are described by a one-dimensional (1D) partial differential wave equation. Dealing with this infinite dimensional description, we show how to get the explicit parameterization of the chain trajectory using (distributed and punctual) advances and delays of its free end.

This parameterization results from symbolic computations. Replacing the time derivative by the Laplace variable s yields a second order differential equation in the spatial variable where s is a parameter. Its fundamental solution is, for each point considered along the chain, an entire function of s of exponential type. Moreover, for each, we show that, thanks to the Liouville transformation, this solution satisfies, modulo explicitly computable exponentials of s, the assumptions of the Paley–Wiener theorem. This solution is, in fact, the transfer function from the flat output (the position of the free end of the system) to the whole state of the system. Using an inverse Laplace transform, we end up with an explicit motion planning formula involving both distributed and punctual advances and delays operators.

Key words. wave equation, delay systems, flatness, motion planning

AMS subject classification. 99C20

PII. S0363012900368636

Introduction. The notion of *flatness* [3, 4] has proven to be relevant in many problems where motion planning problems have been solved [10, 5]. The existence of a *flat output* is the key to explicit formulas that can be implemented as open-loop controllers. Many systems of engineering interest are flat. So far the dynamics under consideration have been nonlinear ordinary differential equations, constant of varying delay equations, or even partial differential equations. In these cases the open-loop controller expression involved algebraic computations, punctual advances and delays [11, 6, 12], distributed advance and delay operators [12, 5, 14, 16], composition of functions [15], etc. In this paper we use both distributed and punctual advances and delays operators.

The heavy chain systems under consideration in this paper are defined by a trolley carrying a fixed length heavy chain to which a load may be attached. The dynamics are studied in a fixed vertical plane. When approximated as a finite set of small pendulums, such heavy chain systems were shown to be flat (see [13]). Their trajectories can be explicitly parameterized by the trajectories of their free ends. These parameterizations involve numerous derivatives (twice as many as the number of pendulums). When this number goes to infinity, the derivative order goes to infinity as

*Received by the editors February 24, 2000; accepted for publication (in revised form) February 26, 2001; published electronically July 19, 2001.
http://www.siam.org/journals/sicon/40-2/36863.html

†Centre Automatique et Systèmes, École des Mines de Paris, 60, bd. Saint-Michel, 75272, Paris Cedex 6, France (petit@cas.ensmp.fr, rouchon@cas.ensmp.fr).

well, yielding series expansions. This makes these relations difficult to handle and to use in practice.

In order to overcome these difficulties, we consider infinite dimensional descriptions of heavy chain systems. Around the stable vertical steady-state and under the small angle assumption, the dynamics are described by second order ordinary differential equations (dynamics of the load at position $y(t)$) coupled with one-dimensional (1D) wave equations (dynamics of the chain $X(x,t)$), where wave speed depends on x, the spatial variable along the chain length.

This combined ordinary and partial differential equation description turns out to be a significant shortcut to an explicit motion planning formula. Instead of an infinite number of derivatives, the explicit parameterization of the trajectories involves a small number of both distributed and punctual advances and delays. The controllability of such hybrid systems could be analyzed via Hilbert's uniqueness method [8, 9], as done in [7]. The work presented here is also a constructive proof of the controllability of these systems in the sense that it provides the open-loop control for steering the system from any given state to any other state. In a real application it should be used as a feedforward term complemented by a closed-loop controller using the energy method as proposed in [2].

In the case of a single homogeneous heavy chain as depicted in Figure 1.1 (see section 1 for details), our explicit parameterization shows that the general solution of

$$\frac{\partial}{\partial x}\left(gx\frac{\partial X}{\partial x}\right) - \frac{\partial^2 X}{\partial t^2} = 0$$

is given by the integral

$$X(x,t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} y(t + 2\sqrt{x/g}\sin\theta)\ d\theta, \tag{0.1}$$

where $t \mapsto y(t)$ is any smooth-enough time function: $X(0,t) = y(t)$ corresponds then to the free end position; the control $u(t) = X(L,t)$ is the trolley position.

For the general cases, we show here that relationships similar to (0.1) exist. They are expressed by (2.2) and (3.2). The structure is similar, but the moving averages involve weights (i.e., kernels) depending on the mass distribution. More precisely, given any mass distribution along the chain and any punctual mass at $x = 0$, we prove that there is a one-to-one correspondence between the trajectory of the load $t \mapsto y(t) = X(0,t)$ and the trajectory of the whole system (namely, the cable and the trolley): $t \mapsto X(x,t)$ and $t \mapsto u(t) = X(L,t)$. This correspondence yields the explicit parameterization of the trajectories: $X(x,\cdot) = \mathcal{A}_x y$, where $\{\mathcal{A}_x\}$ is a set of operators including time derivations, advances, and delays. In other words, $(x,t) \mapsto (\mathcal{A}_x y)(t)$ verifies the system equations for any smooth function $t \mapsto y(t)$. For each x, the operator $\mathcal{A}_x$ admits compact support. Thus it is possible to steer the system from any initial point to any other point in finite time.

This parameterization results from symbolic computations. Replacing the time derivative by the Laplace variable s yields a second order differential equation in x with s as a parameter. For each x, its fundamental solution A_x is an entire function of s of exponential type. Furthermore, for each x we show, thanks to the Liouville transformation, that $s \mapsto A_x(s)$ satisfies the assumptions of the Paley–Wiener theorem, modulo explicitly computable exponentials of s.

The paper is organized as follows.

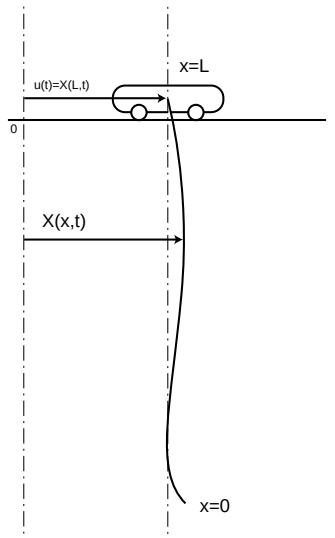


FIG. 1.1. *The homogeneous chain without any load.*

1. In section 1 we consider the case of a homogeneous chain without any load. Although it is the easiest case by far, it is explanatory, and it helps in understanding the meaning and control interest of our results.
2. In section 2 we address the case of an inhomogeneous chain without any load. The problem of the singularity at $x = 0$ of the second order differential equation receives special treatment. We prove the flatness of this system by Theorem 1.
3. In section 3 we solve the general problem of an inhomogeneous chain carrying a punctual load. By contrast with the previous case, the corresponding second order differential is not singular. Flatness of this system is proven by Theorem 2.

1. The homogeneous chain without any load. The computations are simple and explicit and summarize the goal of this paper.

Consider a heavy chain in stable position as depicted in Figure 1.1. Under the small angle approximation it is ruled by the dynamics[1]

$$\begin{cases} \dfrac{\partial}{\partial x}\left(gx\dfrac{\partial X}{\partial x}\right) - \dfrac{\partial^2 X}{\partial t^2} = 0, \\ X(L,t) = u(t), \end{cases} \tag{1.1}$$

where $x \in [0, L]$, $t \in \mathbb{R}$, $X(x,t) - X(L,t)$ is the deviation profile, g is the gravitational acceleration, and the control u is the trolley position.

Thanks to the classical mapping $y = 2\sqrt{\frac{x}{g}}$, we get

$$y\frac{\partial^2 X}{\partial y^2}(y,t) + \frac{\partial X}{\partial y}(y,t) - y\frac{\partial^2 X}{\partial t^2}(y,t) = 0.$$

[1]This model was used in the historical work of D. Bernoulli on a heavy chain system where the zero-order Bessel functions appear for the first time; see [18, pp. 3–4].

Use Laplace transform of X with respect to the variable t (denoted by $\hat{X}$ and with zero initial conditions, i.e., $X(.,0)=0$ and $\frac{\partial X}{\partial t}(.,0)=0$) to get

$$y\frac{\partial^2 \hat{X}}{\partial y^2}(y,s)+\frac{\partial \hat{X}}{\partial y}(y,s)-ys^2\hat{X}(y,s)=0.$$

Less classically, the mapping $z=\imath s y$ gives

$$z\frac{\partial^2 \hat{X}}{\partial z^2}(z,s)+\frac{\partial \hat{X}}{\partial z}(z,s)+z\hat{X}(z,s)=0. \tag{1.2}$$

This is a Bessel equation. Its solution writes in terms of J_0 and Y_0 the zero-order Bessel functions. Using the inverse mapping $z=2\imath s\sqrt{\frac{x}{g}}$, we get

$$\hat{X}(x,s)=A\ J_0(2\imath s\sqrt{x/g})+B\ Y_0(2\imath s\sqrt{x/g}).$$

Since we are looking for a bounded solution at $x=0$, we have $B=0$. Then

$$\hat{X}(x,s)=J_0(2\imath s\sqrt{x/g})\hat{X}(0,s), \tag{1.3}$$

where we can recognize the Clifford function $\mathcal{C}_\prime$ (see [1, p. 358]). Using Poisson's integral representation of J_0 [1, formula 9.1.18],

$$J_0(z)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\exp(\imath z\sin\theta)\ d\theta,$$

we have

$$J_0(2\imath s\sqrt{x/g})=\frac{1}{2\pi}\int_{-\pi}^{\pi}\exp(2s\sqrt{x/g}\sin\theta)\ d\theta.$$

In terms of Laplace transforms, this last expression is a combination of delay operators. Turning (1.3) back into the time-domain, we get

$$X(x,t)=\frac{1}{2\pi}\int_{-\pi}^{\pi}y(t+2\sqrt{x/g}\sin\theta)\ d\theta \tag{1.4}$$

with $y(t)=X(0,t)$.

Relation (1.4) means that there is a one-to-one correspondence between the (smooth) solutions of (1.1) and the (smooth) functions $t\mapsto y(t)$. For each solution of (1.1), set $y(t)=X(0,t)$. For each function $t\mapsto y(t)$, set X by (1.4) and u as

$$u(t)=\frac{1}{2\pi}\int_{-\pi}^{\pi}y(t+2\sqrt{L/g}\sin\theta)\ d\theta \tag{1.5}$$

to obtain a solution of (1.1).

Finding $t\mapsto u(t)$, steering the system from the steady-state $X\equiv 0$ at $t=0$ to the other one $X\equiv D$ at $t=T$ becomes obvious. Our analysis shows that T must be larger than 2Δ, where $\Delta=2\sqrt{L/g}$ is the travelling time of a wave between $x=L$ and $x=0$. It consists only in finding $t\mapsto y(t)$ that is equal to 0 for $t\leq\Delta$ and to D for $t>T-\Delta$ and in computing u via (1.5).

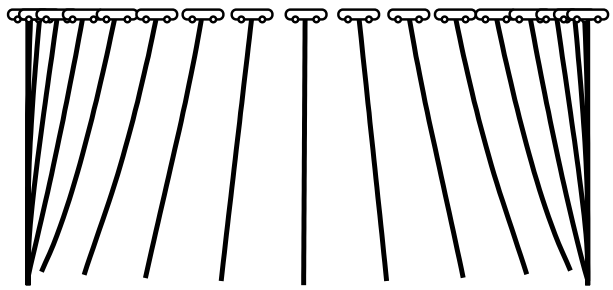


FIG. 1.2. *Steering from* 0 *to* $3L/2$ *in finite time* $T = 4\Delta$. *Regularly time-spaced positions of the heavy chain system are represented. The Matlab simulation code can be obtained from the second author via email.*

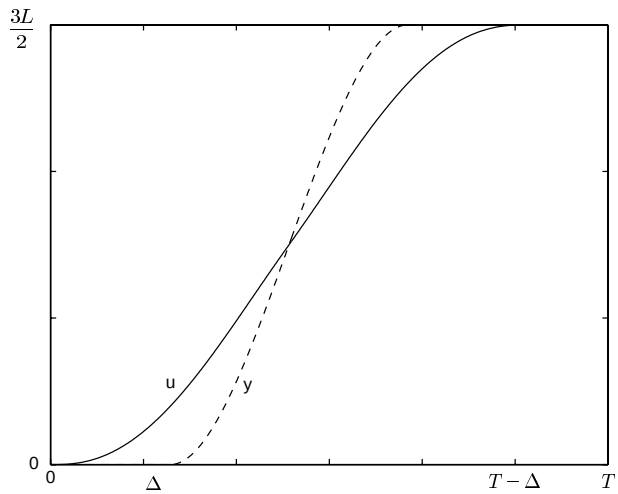


FIG. 1.3. *The steering control, trolley position* u, *and the "flat output," the free end* y.

Figure 1.2 illustrates computations based on (1.4) with

$$y(t) = \begin{cases} 0 \text{ if } t < \Delta, \\ \frac{3L}{2}\left(\frac{t-\Delta}{T-2\Delta}\right)^2\left(3-2\left(\frac{t-\Delta}{T-2\Delta}\right)\right) \text{ if } \Delta \le t \le T-\Delta, \\ \frac{3L}{2} \text{ if } t > T-\Delta, \end{cases}$$

where the chosen transfer time T equals 4Δ. For $t \le 0$ the chain is vertical at position 0. For $t \ge T$ the chain is vertical at position $D = 3L/2$.

Plots of Figure 1.3 show the control $[0,T] \ni t \mapsto u(t)$ required for such motion. Notice that the support of $\dot{u}$ is $[0,T]$, while the support of $\dot{y}$ is $[\Delta, T-\Delta]$. To be consistent with the small angle approximation, the horizontal acceleration of the end point $\ddot{y}$ must be much smaller than g. In our computations the maximum of $|\ddot{y}|$ is chosen rather large, $9g/16$. This is just for tutorial reasons. In practice, a reasonable transition time is $T = 5\Delta$ yielding $|\ddot{y}| \le g/4$.

2. The inhomogeneous (i.e., variable section) chain without any load. Formula (1.4) can be extended to a heavy chain with variable section and carrying no load (see Figure 2.1). Such an extension deserves special consideration because of the singularity of the partial differential system at $x = 0$.

Such a system is governed by the equations

$$\left\{\begin{array}{c} \dfrac{\partial}{\partial x}\left(\tau(x)\dfrac{\partial X}{\partial x}\right) - \dfrac{\tau'(x)}{g}\dfrac{\partial^2 X}{\partial t^2} = 0, \\ X(L,t) = u(t), \end{array}\right. \tag{2.1}$$

where $x \in [0, L]$, $t \in \mathbb{R}$, and u is the control. The tension of the chain is $\tau(x)$ with $\tau(0) = 0$ and $\tau(x) = gx + \mathcal{O}(x^2)$, while $\tau'(x)/g > 0$ is the mass distribution along the chain. Furthermore, we assume that there exists $a > 0$ such that $\tau(x) \geq ax \geq 0$.

THEOREM 1. *Consider* (2.1) *with* $[O, L] \ni x \mapsto \tau(x)$ *a smooth increasing function with* $\tau(0) = 0$ *and* $\tau' > 0$. *There is a one-to-one correspondence between the solutions* $[0, L] \times \mathbb{R} \ni (x, t) \mapsto (X(x, t), u(t))$ *that are* C^3 *in* t *and the* C^3 *functions* $\mathbb{R} \ni t \mapsto y(t)$ *via the formulas*

(2.2)

$$\begin{aligned} X(x,t) =& \frac{L^{1/4}\sqrt{g}}{2\pi^{3/2}(\tau(x)\tau'(x))^{1/4}}\sqrt{G(2\sqrt{\tau(x)/g})}\ \int_{-\pi}^{\pi} y\left(t + KG(2\sqrt{\tau(x)/g})\sin\theta\right) d\theta \\ &+ \frac{1}{(\tau(x)\tau'(x)/g)^{1/4}}\int_{-2\sqrt{\frac{\tau(x)}{ag}}}^{2\sqrt{\frac{\tau(x)}{ag}}} \mathcal{K}(G(2\sqrt{\tau(x)/g}), \xi)\ \dot{y}\,(t+\xi)\ d\xi, \\ u(t) =& X(L,t) \end{aligned}$$

with

$$y(t) = X(0,t),$$

where the constant K *and the functions* G *and* $\mathcal{K}$ *are defined by the function* τ *via formulas* (2.15) *and* (2.29).

The proof of this result is organized as follows.

1. A simple time-scaling simplifies the system. We shift from X to Y.
2. Symbolic computations where time derivatives are replaced by the Laplace variable s are performed.
3. The solution $Y(x, s)$ is factorized as $Y(x, s) = Y(0, s)A(x, s)$. A partial differential system is derived for $A(x, s)$.
4. A Liouville transformation is performed.
5. In these new coordinates the preceding transformed equation is compared to an equation that we have already solved in section 1, namely, the equation of a single homogeneous chain. We denote by $D(x, s)$ the difference between these two solutions.
6. $D(x, s)$ is proven to be an entire function of s and of exponential type.
7. A careful study of the Volterra equation satisfied by $D(x, s)$ shows that, for each x, the restriction to $D(x, s)/s$ to the imaginary axis is in L^2.
8. Thanks to the Paley–Wiener theorem, we prove that, for each x, $D(x, s)/s$ can be represented as a compact sum (discrete and continuous) of exponentials in s.
9. Gathering all the terms of $A(x, s)$, we get an expression involving the Bessel function J_0 (the solution for a homogeneous chain) and exponentials in s multiplied by s. This gives (2.2).

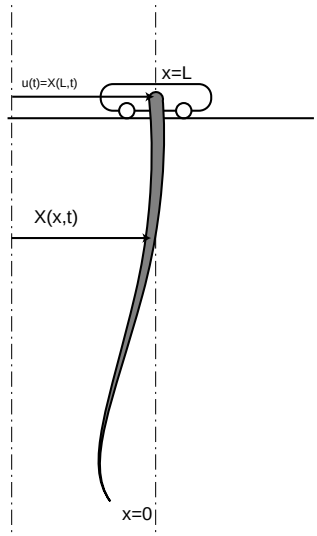


FIG. 2.1. *The inhomogeneous chain without any load.*

Proof. Simple change of coordinates Let[2] $Y(x,t) = X\left(\tau(x)/g, t\right)$.

[2]One may easily show the following result: if Y satisfies

$$\frac{\partial}{\partial x}\left(x\tau' \circ \tau^{-1}(gx)\frac{\partial Y}{\partial x}\right) - \frac{\partial^2 Y}{\partial t^2} = 0, \tag{2.3}$$

then $X(x,t) = Y(\tau(x)/g, t)$ satisfies

$$\frac{\partial}{\partial x}\left(\tau(x)\frac{\partial X}{\partial x}\right) - \frac{\tau'(x)}{g}\frac{\partial^2 X}{\partial t^2} = 0. \tag{2.4}$$

To show this, denote $\circ$ the composition operator with respect to the first variable. Thus $X = Y \circ (\tau/g)$. Then

$$\frac{\partial}{\partial x}\left(\tau\frac{\partial X}{\partial x}\right) = \frac{\partial}{\partial x}\left(\tau\tau'/g\frac{\partial Y}{\partial x} \circ (\tau/g)\right). \tag{2.5}$$

On the other hand, a factorization of (2.3) gives

$$\begin{aligned}\frac{\partial^2 Y}{\partial t^2} &= \frac{\partial}{\partial x}\left(\left(\tau/g\tau'\frac{\partial Y}{\partial x} \circ (\tau/g)\right) \circ \tau^{-1}(gx)\right)\\ &= \frac{\partial}{\partial x}\left(\tau^{-1}(gx)\right)\frac{\partial}{\partial x}\left(\tau\tau'/g\frac{\partial Y}{\partial x} \circ (\tau/g)\right) \circ \tau^{-1}(gx).\end{aligned}$$

So by using (2.5)

$$\frac{\partial}{\partial x}\left(\tau^{-1}(gx)\right)\frac{\partial}{\partial x}\left(\tau\frac{\partial X}{\partial x}\right) \circ \tau^{-1}(gx) = \frac{\partial^2 Y}{\partial t^2}.$$

Yet

$$\frac{\partial}{\partial x}\left(\tau^{-1}(gx)\right) = \frac{g}{\tau' \circ \tau^{-1}(gx)},$$

so

$$\frac{\partial}{\partial x}\left(\tau\frac{\partial X}{\partial x}\right) \circ \tau^{-1}(gx) = \frac{1}{g}\tau' \circ \tau^{-1}(gx)\frac{\partial^2 Y}{\partial t^2},$$

or, equivalently,

$$\frac{\partial}{\partial x}\left(\tau\frac{\partial X}{\partial x}\right) = \frac{\tau'}{g}\frac{\partial^2 Y}{\partial t^2} \circ (\tau/g) = \frac{\tau'}{g}\frac{\partial^2 X}{\partial t^2},$$

which gives the conclusion.

Now (2.1) gives

$$\frac{\partial}{\partial x}\left(\tau_1(x)\frac{\partial Y}{\partial x}\right)-\frac{\partial^2 Y}{\partial t^2}=0, \tag{2.6}$$

where $\tau_1(x)=x\tau'(\tau^{-1}(gx))$.

Symbolic computations. Replacing the time derivation by s gives

$$\frac{\partial}{\partial x}\left(\tau_1(x)\frac{\partial Y}{\partial x}\right)-s^2Y=0. \tag{2.7}$$

Factorization. It is very easy to check that $Y(x,s)=Y(0,s)A(x,s)$ is the solution of (2.7), provided that $A(x,s)$ is solution of the following partial differential system:

$$\begin{cases}\dfrac{\partial}{\partial x}\left(\tau_1(x)\dfrac{\partial A}{\partial x}\right)-s^2A=0,\\ \quad A(0,s)=1.\end{cases} \tag{2.8}$$

Existence of a solution. System (2.8) admits a smooth solution that is an entire function of exponential type in s. This solution reads

$$A(x,s)=\sum_{i\geq 0}\frac{s^{2i}}{i!}f_i(x), \tag{2.9}$$

where

$$\begin{cases}\quad f_0=1,\\ f_i(x)=\displaystyle\int_0^x\frac{1}{\tau_1(l)}\int_0^l if_{i-1}(s)ds\ dl.\end{cases} \tag{2.10}$$

It is very easy to check that, formally, $\sum_{i\geq 0}\frac{s^{2i}}{i!}f_i(x)$ is solution of (2.8): since

$$\frac{\partial}{\partial x}\left(\tau_1(x)\frac{\partial}{\partial x}f_i(x)\right)=if_{i-1}(x),$$

we can write

$$\begin{cases}\dfrac{\partial}{\partial x}\left(\tau_1(x)\dfrac{\partial}{\partial x}\displaystyle\sum_{i\geq 0}\frac{s^{2i}}{i!}f_i(x)\right)=s^2\sum_{i\geq 0}\frac{s^{2i}}{i!}f_i(x),\\ \quad\displaystyle\sum_{i\geq 0}\frac{s^{2i}}{i!}f_i(0)=f_0(0)=1.\end{cases} \tag{2.11}$$

Now let us address the convergence by proving that for all i

$$|f_i(x)|\leq\frac{1}{i!}\left(\frac{x}{a}\right)^i. \tag{2.12}$$

Suppose that (2.12) is true for a given i. (It is obviously the case for $i=0$.) Let us inductively prove that it is also true for $i+1$. From (2.10) we get

$$|f_{i+1}(x)|\leq\int_0^x\frac{l^{i+1}}{\tau_1(l)a^i i!}dl.$$

Yet $\tau' \geq a$, so $\tau_1(x) \geq ax \geq 0$, and then

$$|f_{i+1}(x)| \leq \int_0^x \frac{l^i}{a^{i+1} i!} dl \leq \frac{1}{(i+1)!} \left(\frac{x}{a}\right)^{i+1},$$

which is (2.12) at rank $i+1$.

So, gathering (2.9) and (2.12) and using $\frac{1}{(i!)^2} \leq \frac{2^{2i}}{(2i)!}$, we get

$$A(x,s) \leq \sum_{i\geq 0} \frac{s^{2i} x^i}{(i!)^2 a^i} \leq \sum_{i\geq 0} \frac{s^{2i} 2^{2i} x^i}{(2i)! a^i} \leq \exp\left(2s\sqrt{\frac{x}{a}}\right). \tag{2.13}$$

This proves that, for each x, $s \mapsto A(x,s)$ is an entire function of s of exponential type.

Liouville transformation. The Liouville transformation

$$(x, A) \mapsto (z, u)$$

(see, e.g., [19, p. 110]) turns equations of the form

$$\frac{d}{dx}\left(p(x)\frac{dA}{dx}\right) + (\lambda r(x) - q(x))\, A = 0$$

with $p(x) > 0$ into

$$\frac{d^2u}{dz^2} + (\rho^2 - h(z))u = 0,$$

where ρ is depending only on λ and can be considered as a parameter.

Here

$$p(x) = \tau_1(x), \quad \lambda = -s^2, \quad r(x) = 1, \quad q(x) = 0, \quad x \in [0, L],$$

and the transformation is defined for each $x > 0$. Nevertheless, it can be extended to $x = 0$ because around 0, $\tau_1(x) \approx gx$ with $g > 0$. It turns (2.8) into

$$\frac{d^2u}{dz^2} - K^2 s^2 u = \bar{h}(z) u \tag{2.14}$$

with

$$z = \frac{1}{K}\int_0^x \sqrt{\frac{1}{\tau_1}} \equiv G(2\sqrt{x}), \quad K = \frac{1}{\pi}\int_0^L \sqrt{\frac{1}{\tau_1}}, \tag{2.15}$$

$$u(z,s) = (\tau_1(x))^{1/4} A(x,s), \tag{2.16}$$

$$\bar{h}(z) = \frac{F''(z)}{F(z)} \quad \text{with } F(z) \equiv (\tau_1(x))^{1/4}. \tag{2.17}$$

Notice that since $\tau_1(x) \geq ax$ with $a > 0$, $\int_0^x 1/\tau_1$ is a smooth function of $\sqrt{x}$, and thus G is well defined and invertible. Similar arguments imply that $\bar{h}$ is, in fact, a function of z^2. Thus $\bar{h}(z) = h(z^2)$, and we have the following Laurent series around 0:

$$\bar{h}(z) = h(z^2) = \frac{-1}{4z^2} + \mathcal{O}(1).$$

Comparison to a simpler solution. We know from [1, formula 9.1.49, p. 362] that

$$u_0(z,s) = (Lg)^{1/4}\sqrt{\frac{z}{\pi}}\, J_0(iKsz) \tag{2.18}$$

satisfies

$$\frac{d^2u_0}{dz^2} - K^2s^2u_0 = \left(\frac{-1}{4z^2}\right)u_0. \tag{2.19}$$

According to the Laurent series of $\bar{h}$, we compare the solutions of (2.14), namely, $u(z,s)$, and (2.19), namely, $u_0(z,s)$. Let $D(z,s) = u(z,s) - u_0(z,s)$. We deduce from (2.14) and (2.19) that

$$\frac{d^2D}{dz^2} - K^2s^2D = \left(h(z^2) + \frac{1}{4z^2}\right)u_0 + h(z^2)D. \tag{2.20}$$

Since $z = G(2\sqrt{x})$ with G smooth and invertible, we have from (2.9) and (2.16)

$$u(z,s) = (Lg)^{1/4}\sqrt{\frac{z}{\pi}} + \mathcal{O}(z^{5/2}).$$

Then it is easy to check that for each s, D is a C^1 function of z around 0 with $D(0,s) = 0$ and $D'(0,s) = 0$. Equation (2.20) can be turned into the following integral equation (see [19, p. 111]):

$$\begin{aligned} D(z,s) =& \frac{1}{Ks}\int_0^z \sinh(Ks(z-t))\left(h(z^2) + \frac{1}{4t^2}\right)u_0(t,s)dt \\ &+ \frac{1}{Ks}\int_0^z \sinh(Ks(z-t))h(t^2)D(t,s)dt. \end{aligned} \tag{2.21}$$

Proving that $\mathbb{C} \ni s \mapsto D(z,s)$ is an entire function of exponential type. We already know that $A(x,s)$ and thus $u(z,s)$ (by (2.16)) are entire functions of exponential type in s. On the other hand, for each z, $s \mapsto u_0(z,s)$ is also an entire function of s of exponential type as J_0 is. This gives the conclusion.

Proving that $i\mathbb{R} \ni s \mapsto D(z,s)/s$ belongs to L^2. For each z, we need only an estimation of $D(z,iw)$ as w tends to ∞. For the sake of simplicity, we consider here $w \mapsto D(z,iw)$ for $w > 0$ large enough. The case $w < 0$ is similar. Classically (see, for instance, [19, p. 112]), let $M(z,w) = \sup_{0\le\zeta\le z}|D(\zeta,iw)|$. Using (2.21), we will get an estimation of $M(z,w)$. This gives

$$KwM(z,w) \le I_1(z,w) + I_2(z,w) \tag{2.22}$$

with

$$\begin{aligned} I_1(z,w) &= \int_0^z \left|h(t^2) + \frac{1}{4t^2}\right| |u_0(t,iw)|\, dt, \\ I_2(z,w) &= \int_0^z \left|h(t^2)\right| |D(t,iw)|\, dt. \end{aligned}$$

We know that

$$0 \le z \le \pi, \quad |u_0(t,iw)| \le (Lg)^{1/4}$$

since J_0 is bounded by 1 on the real axis. We know also that $h(t^2)+1/4t^2$ is bounded on $[0,\pi]$. Thus the integral I_1 is bounded by a constant $K_1 > 0$, independent of $z \in [0,\pi]$ and w,

$$I_1(z,w) \leq K_1. \tag{2.23}$$

Next, to majorate I_2 we split it into

$$I_2(z,w) = \underbrace{\int_0^{\gamma/w} \left|h(t^2)\right| |D(t,iw)|\, dt}_{I_2'(z,w)} + \underbrace{\int_{\gamma/w}^{z} \left|h(t^2)\right| |D(t,iw)|\, dt}_{I_2''(z,w)},$$

where $\gamma > 0$ is a parameter we will choose afterwards. A simple but quite tedious computation gives (using $J_0(z) = 1 - \frac{1}{4}z^2 + \circ(z^2)$)

$$D(z,s) = \sqrt{z}\ cs^2z^2(1+\mu(s^2z^2)),$$

where c is a constant and μ is a smooth function such that $\mu(0) = 0$. Using this last expression in I_2', we get

$$I_2'(z,w) \leq \sqrt{w}\frac{bc}{6}\gamma^{3/2}\left(1 + \sup_{|\xi|\leq\gamma^2} |\mu(\xi)|\right), \tag{2.24}$$

where $b > 0$ is such that $\left|h(t^2)\right| \leq b/(4t^2)$ for all $t \in]0,\pi]$. On the other hand, it is easy to check that

$$I_2''(z,w) \leq \frac{bw}{4\gamma} M(z,w). \tag{2.25}$$

Gathering (2.24) and (2.25), we get

$$I_2(z,w) \leq \sqrt{w}\frac{bc}{6}\gamma^{3/2}\left(1 + \sup_{|\xi|\leq\gamma^2} |\mu(\xi)|\right) + \frac{bw}{4\gamma} M(z,w). \tag{2.26}$$

Thanks to the majorations (2.23) and(2.26), we get

$$KwM(z,w) \leq K_1 + \sqrt{w}\frac{bc}{6}\gamma^{3/2}\left(1 + \sup_{|\xi|\leq\gamma^2} |\mu(\xi)|\right) + \frac{bw}{4\gamma} M(z,w).$$

This majoration is valid for $z \in]0,\pi]$, $w > 0$, and $\gamma > 0$ such that $\gamma/w \leq z$. Now we take

$$\gamma = \frac{b}{2K}.$$

Thus for each $z > 0$ and each $w > \gamma/z$, we have

$$(K - b/4\gamma)wM(z,w) \leq K_1 + \sqrt{w}\frac{bc}{6}\gamma^{3/2}\left(1 + \sup_{|\xi|\leq\gamma^2} |\mu(\xi)|\right).$$

Since $K - b/4\gamma = K/2$, we have

$$\frac{1}{2}\ KwM(z,w) \leq K_1 + \sqrt{w}\frac{bc}{6}\gamma^{3/2}\left(1 + \sup_{|\xi|\leq\gamma^2} |\mu(\xi)|\right). \tag{2.27}$$

Thus there exists $C_0 > 0$ such that for each $z \in]0, \pi]$ and for every $w > \gamma/z$,

$$|D(z, iw)| \leq \frac{C_0}{\sqrt{|w|}}. \tag{2.28}$$

Since $D(z, 0) \equiv 0$, we deduce for each $z > 0$ that $s \mapsto D(z, s)/s$ remains an entire function of s (of exponential type), and the above majoration says that $i\mathbb{R} \ni s \mapsto D(z, s)/s$ belongs to L^2.

Using the Paley–Wiener theorem. The Paley–Wiener theorem [17, p. 375] ensures that, for any $z \in [0, \pi]$, there exists $[-\frac{G^{-1}(z)}{\sqrt{a}}, \frac{G^{-1}(z)}{\sqrt{a}}] \ni t \mapsto \mathcal{K}(z, t)$ in L^2 such that

$$D(z, s)/s = \int_{-\frac{G^{-1}(z)}{\sqrt{a}}}^{\frac{G^{-1}(z)}{\sqrt{a}}} \mathcal{K}(z, \xi) \exp(s\xi) d\xi. \tag{2.29}$$

The integral bounds results from the following facts.

1. Via (2.16), $2\sqrt{x} = G^{-1}(z)$, and (2.13), we have
$$\forall s \in \mathbb{C}, \quad |(u(z, s)| \leq N(z) \exp\left(|s| \frac{G^{-1}(z)}{\sqrt{a}}\right)$$
for some $N(z) > 0$.
2. A well-known property on J_0 implies that
$$\forall s \in \mathbb{C}, \quad |(u_0(z, s)| \leq N_0(z) \exp(|s|\, zK)$$
for some $N_0(z) > 0$.
3. Since $\tau_1 x \geq ax$, (2.15) implies that $zK < \frac{G^{-1}(z)}{\sqrt{a}}$.
4. Thus
$$\forall s \in \mathbb{C}, \quad |D(z, s)| = |u(z, s) - u_0(z, s)| \leq (N(z) + N_0(z)) \exp\left(|s| \frac{G^{-1}(z)}{\sqrt{a}}\right).$$

Conclusion.

$$(u(z, s) - u_0(z, s))/s = \int_{-\frac{G^{-1}(z)}{\sqrt{a}}}^{\frac{G^{-1}(z)}{\sqrt{a}}} \mathcal{K}(z, \xi) \exp(s\xi) d\xi.$$

This gives

$$u(z, s) = \frac{(Lg)^{1/4}}{\sqrt{\pi}} \sqrt{z} J_0(iKsz) + \int_{-\frac{G^{-1}(z)}{\sqrt{a}}}^{\frac{G^{-1}(z)}{\sqrt{a}}} s\mathcal{K}(z, \xi) \exp(s\xi) d\xi.$$

Pulling back this relation in the (x, A) coordinates, we deduce using (2.16) that

$$\begin{aligned} A(x, s) =& \frac{(Lg)^{1/4}}{\sqrt{\pi}} \frac{1}{(\tau_1(x))^{1/4}} \sqrt{G(2\sqrt{x})} J_0(iKsG(2\sqrt{x})) \\ &+ \frac{1}{(\tau_1(x))^{1/4}} \int_{-2\sqrt{\frac{x}{a}}}^{2\sqrt{\frac{x}{a}}} s\mathcal{K}(G(2\sqrt{x}), \xi) \exp(s\xi) d\xi. \end{aligned}$$

Then we quickly get $Y(x,s) = Y(0,s)A(x,s)$. This gives in the time domain

$$\begin{aligned} Y(x,t) =& \frac{(Lg)^{1/4}}{\sqrt{\pi}} \frac{1}{(\tau_1(x))^{1/4}} \sqrt{G(2\sqrt{x})}\ \frac{1}{2\pi} \int_{-\pi}^{\pi} Y(0, t + KG(2\sqrt{x}) \sin\theta) d\theta \\ &+ \frac{1}{(\tau_1(x))^{1/4}} \int_{-2\sqrt{\frac{x}{a}}}^{2\sqrt{\frac{x}{a}}} \mathcal{K}(G(2\sqrt{x}), \xi) \left[\frac{\partial}{\partial t} Y(0, t+\xi) \right] d\xi. \end{aligned}$$

Then substituting

$$\begin{aligned} X(x,t) &= Y\left(\tau(x)/g, t\right), \\ Y(0,t) &= X(0,t), \\ \frac{\partial Y}{\partial t}(0,t) &= \frac{\partial X}{\partial t}(0,t), \end{aligned}$$

we get

(2.30)

$$\begin{aligned} X(x,t) =& \frac{L^{1/4}\sqrt{g}}{2\pi^{3/2}(\tau(x)\tau'(x))^{1/4}} \sqrt{G(2\sqrt{\tau(x)/g})}\ \int_{-\pi}^{\pi} y(t + KG(2\sqrt{\tau(x)/g}) \sin\theta) d\theta \\ &+ \frac{1}{(\tau(x)\tau'(x)/g)^{1/4}} \int_{-2\sqrt{\frac{\tau(x)}{ag}}}^{2\sqrt{\frac{\tau(x)}{ag}}} \mathcal{K}(G(2\sqrt{\tau(x)/g}), \xi)\ \dot{y}\left(t+\xi\right)\ d\xi \end{aligned}$$

with $y(t) = X(0,t)$. □

Remark. In the case of a homogeneous chain, we can substitute

$$\begin{aligned} &\tau(x) = gx, \quad \tau'(x) = g, \quad \tau_1(x) = gx = \tau(x), \\ &K = \frac{2}{\pi}\sqrt{\frac{L}{g}}, \quad z = G(2\sqrt{x}) = \pi\sqrt{\frac{x}{L}}, \mathcal{K} = 0, \end{aligned}$$

and (2.30) reads

$$X(x,t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} y\left(t + 2\sqrt{\frac{x}{g}} \sin\theta\right) d\theta,$$

which is indeed identical to (1.4).

3. The inhomogeneous chain with punctual load. The system of Figure 3.1 consists of a heavy chain with a variable section carrying a punctual load m. Small deviations $X(x,t) - u(t)$ from the vertical position are described by the partial differential system

$$\left\{ \begin{aligned} &\frac{\partial}{\partial x}\left(\tau(x)\frac{\partial X}{\partial x}\right) - \frac{\tau'(x)}{g}\frac{\partial^2 X}{\partial t^2} = 0, \\ &\frac{\partial^2 X}{\partial t^2}(0,t) = g\frac{\partial X}{\partial x}(0,t), \\ &X(L,t) = u(t), \end{aligned} \right. \tag{3.1}$$

where u is the control. The tension in the chain writes $\tau(x)$: $\tau(0) = mg$, and $\tau'(x)/g > 0$ is the mass distribution along the chain.

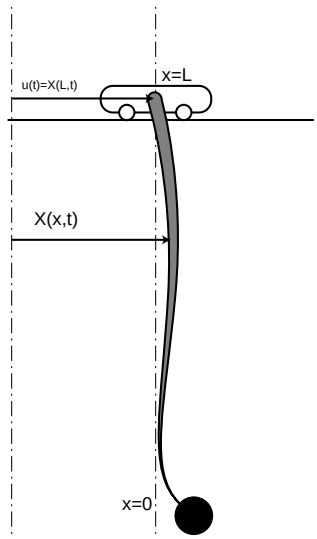


FIG. 3.1. *The inhomogeneous (variable section) chain with punctual load.*

THEOREM 2. *Consider* (3.1) *with* $[0,L] \ni x \mapsto \tau(x)$ *a smooth increasing function with* $\tau(0) = m$. *There is a one-to-one correspondence between the solutions* $[0,L]\times\mathbb{R} \ni (x,t) \mapsto (X(x,t), u(t))$ *that are* C^3 *in* t *and the* C^3 *functions* $\mathbb{R} \ni t \mapsto y(t)$ *via the following formulas:*

(3.2)
$$\begin{cases} X(x,t) = \phi(x)\ [y(t+\theta(x)) + y(t-\theta(x))] + \psi(x)\ [\dot{y}(t+\theta(x)) - \dot{y}(t-\theta(x))] \\ \qquad\qquad + \displaystyle\int_0^x \mathcal{B}(x,\xi)[y(t+\theta(\xi)) + y(t-\theta(\xi))]\ d\xi, \\ u(t) = X(L,t) \end{cases}$$

with

$$\begin{aligned} y(t) &= X(0,t), \\ \theta(x) &= \int_0^x \sqrt{\frac{\tau'}{g\tau}}, \\ \psi(x) &= \left(\frac{\tau(0)\tau'(0)}{\tau(x)\tau'(x)}\right)^{\frac{1}{4}} \frac{1}{2}\sqrt{\frac{\tau(0)}{g\tau'(0)}}, \\ \phi(x) &= \left(\frac{\tau(0)\tau'(0)}{\tau(x)\tau'(x)}\right)^{\frac{1}{4}} \cdots \\ &\quad \times \left[1 + \frac{1}{8}\sqrt{\frac{\tau(0)}{\tau'(0)}} \left(\left(\sqrt{\frac{\tau'}{\tau}} + \frac{\tau''}{\tau'}\sqrt{\frac{\tau}{\tau'}}\right)(x) - \left(\sqrt{\frac{\tau'}{\tau}} + \frac{\tau''}{\tau'}\sqrt{\frac{\tau}{\tau'}}\right)(0) \right.\right. \\ &\qquad\qquad \left.\left. + \cdots + \frac{1}{4}\int_0^x \left(\sqrt{\frac{\tau'}{\tau}} + \frac{\tau''}{\tau'}\sqrt{\frac{\tau}{\tau'}}\right)^2 \sqrt{\frac{\tau'}{\tau}}\right)\right], \end{aligned}$$

$B(x,\xi)$ a smooth function of x, and ξ defined by the function τ via formula (3.15).

Correspondence (3.2) defines a family of linear operators $\mathcal{A}_x$ with compact support such that, for any C^3 time function, $X(x,t) = \mathcal{A}_x y|_t$ is automatically the solution of (3.1) with $u(t) = X(L,t)$ and $X(0,t) = y(t)$.

The proof relies on the following points.

1. Symbolic computations where the time derivation is replaced by the Laplace variable s are performed. This yields a second order differential equation with nonconstant coefficients in the space variable x.
2. The solution $X(x,s)$ is factorized as $X(x,s) = X(0,s)A(x,s)$. A partial differential system is derived for $A(x,s)$.
3. The study of $s \mapsto A(x,s)$ is simplified by a Liouville transformation $(x,A) \mapsto (z,u)$.
4. The solution $A(x,s)$ of this differential equation is proven to be an entire function of s and of exponential type. (Volterra expansion and majoring series arguments are used.)
5. A careful study of the Volterra equation of the second kind satisfied by A shows that modulo some functions (exponentials of s, depending on x and explicitly calculated), for each x, the restriction of $A(x,s)$ to the imaginary axis is in L^2.
6. Thanks to the Paley–Wiener theorem and the last two properties of A, we prove that, for each x, A can be represented as a compact sum (discrete and continuous) of exponentials in s. This gives (3.2).

Proof. Symbolic computation. Replacing the time derivation by s gives

$$\begin{cases} \dfrac{\partial}{\partial x}\left(\tau(x)\dfrac{\partial X}{\partial x}\right) - \dfrac{\tau'(x)}{g}s^2 X = 0, \\ s^2 X(0,s) = gX'(0,s). \end{cases} \tag{3.3}$$

We do not consider the other boundary condition since u is the control and can be obtained explicitly from X via $u(t) = X(L,t)$.

Factorization. It is very easy to check that $X(x,s) = X(0,s)A(x,s)$ is the solution of (3.3), provided that $A(x,s)$ is the solution of the following partial differential system:

$$\begin{cases} \dfrac{\partial}{\partial x}\left(\tau(x)\dfrac{\partial A}{\partial x}\right) - \dfrac{\tau'(x)}{g}s^2 A = 0, \\ A(0,s) = 1, \\ gA'(0,s) = s^2. \end{cases} \tag{3.4}$$

Liouville transformation. This time we perform a Liouville transformation (already used in section 2)

$$(x,A) \mapsto (z,u)$$

with

$$p(x) = \tau(x), \quad \lambda = -\frac{s^2}{g}, \quad r(x) = \tau'(x), \quad q = 0, \quad x \in [0,L].$$

The new variables (z, u) are defined by the following formulas:

$$z = \frac{1}{K}\int_0^x \sqrt{\frac{\tau'}{\tau}}, \quad 0 \le z \le \pi, \quad K = \frac{1}{\pi}\int_0^L \sqrt{\frac{\tau'}{\tau}}, \tag{3.5}$$

$$u(z,s) = (\tau(x)\tau'(x))^{1/4} A(x,s). \tag{3.6}$$

System (3.4) is turned into

$$\frac{d^2u}{dz^2} + (\rho^2 - h(z))u = 0 \quad \text{with } \frac{du}{dz}(0) = (a + b\rho^2), \quad u(0) = 1, \tag{3.7}$$

where

$$\rho = \imath\, \frac{K}{\sqrt{g}} s, \quad \imath = \sqrt{-1},$$
$$h(z) = \frac{f''(z)}{f(z)} \quad \text{with } f(z) = (\tau(x)\tau'(x))^{1/4},$$
$$a = \frac{f'(0)}{f(0)}, \quad b = \frac{1}{K}\sqrt{\frac{\tau(0)}{\tau'(0)}}.$$

Proving that $\mathbb{C} \ni \rho \mapsto u(z,\rho)$ is an entire function of exponential type. We claim that, for each z, $\rho \mapsto u(z,\rho)$ is an entire function of exponential type.

Denote by $W(z,\rho)$ the 2×2 matrix solution of

$$\frac{dW}{dz} = \begin{pmatrix} 0 & 1 \\ h(z) - \rho^2 & 0 \end{pmatrix} W$$

with $W(0,\rho) = I$. Since

$$u(z,\rho) = \begin{pmatrix} 1 & 0 \end{pmatrix} W(z,\rho) \begin{pmatrix} 1 \\ a + b\rho^2 \end{pmatrix},$$

it suffices to prove that W is entire in ρ and of exponential type. Using the classical fixed point technique, W can be expressed as an absolutely convergent series of iterated integrals (Volterra expansion)

$$W(z,\rho) = \sum_{i \ge 0} W_i(z,\rho)$$

with

$$W_0(z,\rho) = I, \quad W_{i+1}(z) = \int_0^z \begin{pmatrix} 0 & 1 \\ h(\sigma) - \rho^2 & 0 \end{pmatrix} W_i(\sigma,\rho)\, d\sigma. \tag{3.8}$$

For each $i > 0$, $W_i(z,\rho)$ is a polynomial in ρ^2 of degree i with coefficients depending on z. Thus we have

$$\sum_{0 \le i \le k} W_i(z,\rho) = \sum_{0 \le j \le k} W^{j,k}(z)\, \rho^{2j}.$$

From step k to $k+1$, we add to $W^{j,k}(z)$ the coefficient of ρ^{2j} in W_{k+1}, say, $\mathcal{W}^{j,k+1}$, to obtain $W^{j,k+1}(z)$:

$$W^{j,k+1}(z) = W^{j,k}(z) + \mathcal{W}^{j,k+1}(z).$$

Let $\alpha = \sup_{[0,\pi]} |h|$. Then the absolute value of each entry of $W_i(z,\rho)$ is bounded by the corresponding entries in the following *majoring series* $M_i(z,\rho)$ defined by the induction (to be compared to (3.8)):

$$M_0(z,\rho) = I, \quad M_{i+1}(z) = \int_0^z \begin{pmatrix} 0 & 1 \\ \alpha+\rho^2 & 0 \end{pmatrix} M_i(\sigma,\rho)\, d\sigma. \tag{3.9}$$

As for W, we can define $M = \sum_{i\geq 0} M_i$ and, for each $k > 0$, the matrices $M^{j,k}$ and $\mathcal{M}^{j,k+1}$ satisfying

$$\sum_{0\leq i\leq k} M_i(z,\rho) = \sum_{0\leq j\leq k} M^{j,k}(z)\, \rho^{2j}, \quad M^{j,k+1}(z) = M^{j,k}(z) + \mathcal{M}^{j,k+1}(z).$$

Standard matrix computations show that

$$\begin{aligned} M(z,\rho) = I &+ \sum_{i>0} \frac{z^{2i}}{(2i)!} \begin{pmatrix} (\rho^2+\alpha)^i & 0 \\ 0 & (\rho^2+\alpha)^i \end{pmatrix} \\ &+ \sum_{i>0} \frac{z^{2i+1}}{(2i+1)!} \begin{pmatrix} 0 & (\rho^2+\alpha)^i \\ (\rho^2+\alpha)^{i+1} & 0 \end{pmatrix}. \end{aligned}$$

That is,

$$M(z,\rho) = \begin{pmatrix} \cosh(z\sqrt{\rho^2+\alpha}) & \sinh(z\sqrt{\rho^2+\alpha})/\sqrt{\rho^2+\alpha} \\ \sinh(z\sqrt{\rho^2+\alpha})\sqrt{\rho^2+\alpha} & \cosh(z\sqrt{\rho^2+\alpha}) \end{pmatrix}. \tag{3.10}$$

For each j, the matrices $M^{j,k} = \sum_{j\leq l\leq k-1} \mathcal{M}^{j,l}$ converge as k tends to ∞. Denote by M^j the limit. By construction, $M = \sum_{j\geq 0} M^j(z)\, \rho^{2j}$, and this series has an infinite radius of convergence in ρ, since, for each z, the functions $\rho \mapsto \cosh(z\sqrt{\rho^2+\alpha})$, $\rho \mapsto \sinh(z\sqrt{\rho^2+\alpha})/\sqrt{\rho^2+\alpha}$, and $\rho \mapsto \sinh(z\sqrt{\rho^2+\alpha})\sqrt{\rho^2+\alpha}$ are entire functions of ρ^2.

But, for each i, j, and k, the matrices $M^{j,k}$ and $\mathcal{M}^{j,k+1}$, whose entries are always nonnegative, dominate the absolute values of the entries of $W^{j,k}$ and $\mathcal{W}^{j,k+1}$, respectively. Thus for each j, the matrices $W^{j,k} = \sum_{j\leq l\leq k-1} \mathcal{W}^{j,l}$ converge as k tends to ∞. Denote by W^j the limit. By construction, $W = \sum_{j\geq 0} W^j(z)\rho^{2j}$, and this series has an infinite radius of convergence in ρ, since M has one. In other words, W is an entire function of ρ. Moreover, the entries of M are upper bounds of the entries of W. Thus W is of exponential type in ρ: for each $z \in [0,\pi]$, there exists $E > 0$ such that

$$\forall \rho \in \mathbb{C}, \quad |W(z,\rho)| \leq E \exp(z|\rho|).$$

We have proven that, for each $z \in [0,\pi]$, $u(z,\rho)$ is an entire function of ρ of exponential type with

$$\forall \rho \in \mathbb{C}, \quad |u(z,\rho)| \leq b(z) \exp(z|\rho|)$$

for some $b(z) > 0$ well-chosen.

Proving that "a part" of $\mathbb{R} \ni \rho \mapsto u(z,\rho)$ belongs to L^2. In general, $\mathbb{R} \ni \rho \mapsto u(z,\rho)$ does not belong to L^2. Thus the Paley–Wiener theorem does not apply directly. Removing some appropriate terms, the remaining is in L^2.

Let

$$v(z,\rho) = u(z,\rho) + b\rho\sin(\rho z) - \left(1 + \frac{b\int_0^z h}{2}\right)\ \cos(\rho z). \tag{3.11}$$

In the following we prove that this entire function of exponential type is such that $\mathbb{R} \ni \rho \mapsto v(z,\rho)$ belongs to L^2.

From the Volterra equation of the second kind satisfied by u (see [19, p. 111]),

$$u(z,\rho) = \left(\cos(\rho z) + (a - b\rho^2)\frac{\sin(\rho z)}{\rho}\right) + \frac{1}{\rho}\int_0^z \sin(\rho(z-\zeta))\ h(\zeta)\ u(\zeta,\rho)\ d\zeta,$$

we quickly derive a similar equation satisfied by v,

$$v(z,\rho) = \phi(z,\rho) + \frac{1}{\rho}\int_0^z \sin(\rho(z-\zeta))\ h(\zeta)\ v(\zeta,\rho)\ d\zeta,$$

where $\phi = \phi_1 - b\phi_2$ with

$$\begin{aligned}
\phi_1(z,\rho) =& a\frac{\sin(\rho z)}{\rho} + \frac{1}{\rho}\int_0^z \sin(\rho(z-\zeta))h(\zeta)\cos(\rho\zeta)\left(1 + (b/2)\int_0^\zeta h\right)\ d\zeta,\\
\phi_2(z,\rho) =& \cos(\rho z)\int_0^z h/2 + \int_0^z \sin(\rho(z-\zeta))h(\zeta)\sin(\zeta)\ d\zeta.
\end{aligned}$$

Clearly, there exists $D_1 > 0$ such that for all $z \in [0,\pi]$ and $\rho \in \mathbb{R}$,

$$|\phi_1(z,\rho)| \le \frac{D_1}{1+|\rho|}$$

(h is bounded). With $2\sin(\rho(z-\zeta))\sin(\zeta) = \cos(\rho(z-2\zeta)) - \cos(\rho z)$, we have

$$\phi_2(z,\rho) = \int_0^z \cos(\rho(z-2\zeta))h(\zeta)\ d\zeta.$$

The integration by part (by assumption τ is C^4 thus h is C^1)

$$\int_0^z \cos(\rho(z-2\zeta))h(\zeta)\ d\zeta = \frac{h(0)+h(z)}{2\rho}\sin(\rho z) + \frac{1}{2\rho}\ \int_0^z \sin(\rho(z-2\zeta))h'(\zeta)\ d\zeta$$

shows that for large $|\rho|$, ϕ_2 tends to zero at least as $1/|\rho|$. Thus there exists $D_2 > 0$ such that for all $z \in [0,\pi]$ and $\rho \in \mathbb{R}$,

$$|\phi_2(z,\rho)| \le \frac{D_2}{1+|\rho|}.$$

This proves that v satisfies

$$v(z,\rho) = \phi(z,\rho) + \frac{1}{\rho}\int_0^z \sin(\rho(z-\zeta))h(\zeta)v(\zeta,\rho)\ d\zeta \tag{3.12}$$

with $|\phi(z,\rho)| \le D/(1+|\rho|)$ for all $z \in [0,\pi]$ and $\rho \in \mathbb{R}$. ($D > 0$ is a well-chosen constant independent of z and ρ.)

This last inequality gives the desired conclusion by the following classical computation (see [19, p. 112], for instance).

Let $\beta(z,\rho) = \sup_{0\le\zeta\le z} |v(\zeta,\rho)|$. By (3.12) we have for each z_1 and z_2 in $[0,\pi]$, $z_1 \le z_2$

$$|v(z_1,\rho)| \le \frac{D}{1+|\rho|} + \frac{\alpha z_1 \beta(z_2,\rho)}{|\rho|} \le \frac{D}{1+|\rho|} + \frac{\alpha\pi}{|\rho|}\ \beta(z_2,\rho).$$

(Remember that $\alpha = \sup_{[0,\pi]} |h|$.) In particular, when $z_1 = z_2 = z$, we have

$$\beta(z,\rho)\left(1 - \frac{\alpha\pi}{|\rho|}\right) \le \frac{D}{1+|\rho|}. \tag{3.13}$$

Finally, for $|\rho| \ge 2\alpha\pi$, $\beta(z,\rho) \le 2D/(1+|\rho|)$. This proves that $\mathbb{R} \ni \rho \mapsto v(z,\rho)$ belongs to L^2.

Using the Paley–Wiener theorem. At last, the Paley–Wiener theorem ensures that the Fourier transform of $\rho \mapsto v(z,\rho)$ has a compact support included in $[-z,z]$ since for all $\rho \in \mathbb{C}, |v(z,\rho)| \le N\exp(z|\rho|)$ for some constant $N > 0$. This means that, for each $z \in [0,\pi]$, there exists $[-z,z] \ni \zeta \mapsto \mathcal{K}(z,\zeta)$ in $L^2([-z,z])$ such that

$$v(z,\rho) = \int_{-z}^{+z} \mathcal{K}(z,\zeta)\exp(\imath\zeta\rho)\ d\zeta.$$

Since v is an even function of ρ, $\mathcal{K}$ is also an even function of ζ. Thus we have, finally,

$$v(z,\rho) = \int_{0}^{+z} \mathcal{K}(z,\zeta)(\exp(\imath\zeta\rho) + \exp(-\imath\zeta\rho))\ d\zeta. \tag{3.14}$$

Conclusion. Pulling back this last relation in the (x,A) coordinates, noticing that $\rho = \imath Ks/\sqrt{g}$, that $\exp(-\theta s)$ is the Laplace transform of the θ-delay operator, and that $u(0,\rho)$ is, up to a constant, the Laplace transform of $X(0,t)$, we deduce after some standard but tedious computations formulae (3.2). The new function $\mathcal{B}(x,\xi)$ is related to $\mathcal{K}(z,\zeta)$ via

$$K\sqrt{\frac{\tau(\xi)}{\tau'(\xi)}}\ \mathcal{B}(x,\xi) = \left(\frac{\tau(0)\tau'(0)}{\tau(x)\tau'(x)}\right)^{\frac{1}{4}}\ \mathcal{K}\left(\frac{\sqrt{g}}{K}\theta(x), \frac{\sqrt{g}}{K}\theta(\xi)\right). \tag{3.15}$$

At last,

$$\begin{aligned} A(x,s) = \varphi(x)\,(\exp\theta(x)s + \exp\theta(x)s) + \psi(x)s\,(\exp\theta(x)s - \exp\theta(x)s) \\ + \int_0^x \mathcal{K}(x,\zeta)(\exp(\theta(\zeta)s) + \exp(-\theta(\zeta)s))\ d\zeta, \end{aligned}$$

so $X(x,s) = X(0,s)A(x,s)$ when turned back into the time-domain does give formulae (3.2). □

4. Conclusion. We have shown that, around the stable vertical position, heavy chain systems with or without load, with constant or variable section, are "flat": the trajectories of these systems are parameterizable by the trajectories of their free ends. Relations (1.4), (2.2), and (3.2) show that such parameterizations involve operators of compact supports.

It is surprising that such parameterizations can also be applied around the inverse and unstable vertical position. For the homogenous heavy chain, we have only to replace g by $-g$ to obtain a family of smooth solutions to the elliptic equation (singular at $x = 0$)

$$\frac{\partial}{\partial x}\left(gx\frac{\partial X}{\partial x}\right) + \frac{\partial^2 X}{\partial t^2} = 0$$

by the integral

$$X(x,t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} y(t + 2\imath\sqrt{x/g}\sin\theta)\ d\theta,$$

where y is now a holomorphic function in $\mathbb{R} \times [-2\sqrt{L/g}, +2\sqrt{L/g}]$ that is real on the real axis. This parameterization can still be used to solve the motion planning problem in spite of the fact that the Cauchy problem associated to this elliptic equation is not well-posed in the sense of Hadamard.

Acknowledgments. The authors are indebted to Michel Fliess and Philippe Martin for fruitful discussions relative to the Paley–Wiener theorem, series expansions, majoring series arguments, and Liouville transformations.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions*, Dover, New York, 1965.

[2] F. Boustany, *Commande Nonlinéaire Adaptative de Systèmes Mécaniques de Type Pont Roulant, Stabilisation Frontière d'EDP*, Ph.D. thesis, École des Mines de Paris, Paris, France, 1992.

[3] M. Fliess, J. Lévine, P. Martin, and P. Rouchon, *Flatness and defect of nonlinear systems: Introductory theory and examples*, Internat. J. Control, 61 (1995), pp. 1327–1361.

[4] M. Fliess, J. Lévine, P. Martin, and P. Rouchon, *A Lie-Bäcklund approach to equivalence and flatness of nonlinear systems*, IEEE Trans. Automat. Control, 44 (1999), pp. 922–937.

[5] M. Fliess, P. Martin, N. Petit, and P. Rouchon, *Active signal restoration for the telegraph equation*, in Proceedings of the 38th IEEE Conference on Decision and Control, IEEE Computer Society, Los Alamitos, CA, 1999, pp. 1007–1011.

[6] M. Fliess and H. Mounier, *Controllability and observability of linear delay systems: An algebraic approach*, ESAIM Control Optim. Calc. Var., 3 (1998), pp. 301–314.

[7] S. Hansen and E. Zuazua, *Exact controllability and stabilization of a vibrating string with an interior point mass*, SIAM J. Control Optim., 33 (1995), pp. 1357–1391.

[8] J.-L. Lions, *Contrôlabilité Exacte, Perturbations et Stabilisation de Systèmes Distribués*, Masson, Paris, 1988.

[9] J.-L. Lions, *Exact controllability, stabilization and perturbations for distributed systems*, SIAM Rev., 30 (1988), pp. 1–68.

[10] P. Martin and P. Rouchon, *Flatness and sampling control of induction motors*, in Proceedings of the IFAC World Congress, San Francisco, CA, 1996, pp. 389–394.

[11] H. Mounier, *Propriétés Structurelles des Systèmes Linéaires à Retards: Aspects Théoriques et Pratiques*, Ph.D. thesis, Université Paris Sud, Orsay, France, 1995.

[12] H. Mounier, J. Rudolph, M. Fliess, and P. Rouchon, *Tracking control of a vibrating string with an interior mass viewed as delay system*, ESAIM Control Optim. Calc. Var., 3 (1998), pp. 315–321.

[13] R. M. Murray, *Trajectory generation for a towed cable flight control system*, in Proceedings of the IFAC World Congress, San Francisco, CA, 1996, pp. 395–400.

[14] N. Petit, *Systèmes à Retards. Platitude en Génie des Procédés et Contrôle de Certaines Équations des Ondes*, Ph.D. thesis, École des Mines de Paris, Paris, France, 2000.

[15] N. Petit, Y. Creff, and P. Rouchon, *Motion planning for two classes of nonlinear systems with delays depending on the control*, in Proceedings of the 37th IEEE Conference on Decision and Control, IEEE Computer Society, Los Alamitos, CA, 1998, pp. 1107–1111.

[16] N. Petit and P. Rouchon, *Dynamics and Solutions to Some Control Problems for Water-Tank Systems*, CDS Technical Memo CIT-CDS 00-004, California Institute of Technology, Pasadena, CA, 2000.
[17] W. Rudin, *Real and Complex Analysis*, 2nd ed., McGraw-Hill, New York, St. Louis, Paris, 1974.
[18] G. N. Watson, *A Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge University Press, Cambridge, UK, 1958.
[19] K. Yosida, *Lectures on Differential and Integral Equations*, Interscience, New York, 1960.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 2, pp. 496–515

A GENERALIZATION OF ZUBOV'S METHOD TO PERTURBED SYSTEMS*

FABIO CAMILLI†, LARS GRÜNE‡, AND FABIAN WIRTH§

Abstract. A generalization of Zubov's theorem on representing the domain of attraction via the solution of a suitable partial differential equation is presented for the case of perturbed systems with a singular fixed point. For the construction it is necessary to consider solutions in the viscosity sense. As a consequence, maximal robust Lyapunov functions can be characterized as viscosity solutions.

Key words. Zubov's method, robust stability, domain of attraction, viscosity solutions

AMS subject classifications. 93D09, 49L25, 93D30

PII. S036301299936316X

1. Introduction.

The domain of attraction of an asymptotically stable fixed point has been one of the central objects in the study of continuous dynamical systems. In the late 1960's there was a particular surge of activity with a number of papers by Coleman [8], [9], Wilson [25], and Bhatia [6] analyzing properties of the domains. One of the celebrated results of that era was what came to be known as Zubov's method [26], which asserts that the domain of attraction of an asymptotically stable fixed point x^* of

$$\dot{x} = f(x)\,, \quad x \in \mathbb{R}^n,$$

may be characterized by solutions v of the partial differential equation

$$Dv(x) \cdot f(x) = -h(x)(1 - v(x))\sqrt{1 + \|f(x)\|^2}\,. \tag{1.1}$$

Namely, under suitable assumptions on h, the set $v^{-1}([0,1))$ is equal to the domain of attraction. These results are presented in several books; see [11] or [14]. For the case of real-analytic systems a constructive procedure is presented in [11] that allows for the approximation of the domain of attraction. This method was extended and simplified in [24], where again a constructive approach for the case of analytic systems is presented. The construction was extended to the case of asymptotically stable periodic orbits in [2].

In recent years much effort has been devoted to the development of numerical methods for the approximation of domains of attractions. Zubov's method also lends itself to the construction of such schemes; see [24], [13] and the paper [1], which considers a particular application.

*Received by the editors October 29, 1999; accepted for publication (in revised form) March 14, 2001; published electronically July 19, 2001. This research was supported by the TMR Networks "Nonlinear Control Network" and "Viscosity Solutions and Their Applications" and by the DFG Priority Research Program "Ergodentheorie, Analysis und effiziente Simulation dynamischer Systeme."
http://www.siam.org/journals/sicon/40-2/36316.html

†Dipartimento di Energetica, Fac. di Ingegneria, Università de l'Aquila, 67040 Roio Poggio (AQ), Italy (camilli@axcasp.caspur.it).

‡Fachbereich Mathematik, J. W. Goethe-Universität, Postfach 11 19 32, 60054 Frankfurt a.M., Germany (gruene@math.uni-frankfurt.de).

§Zentrum für Technomathematik, Universität Bremen, 28334 Bremen, Germany (fabian@math.uni-bremen.de).

In this paper our aim is to generalize Zubov's basic result by incorporating perturbations into the setup. That is, we consider systems of the form

$$\dot{x} = f(x, a)$$

with the property that the fixed point (which we take to be zero) is not perturbed under all perturbations. Under a local stability assumption, which guarantees that it is reasonable to consider domains of attraction we are interested in the set of points that is attracted to the fixed point regardless of the perturbation considered. This is what we call the robust domain of attraction. This subset of the domain of the unperturbed system $\dot{x} = f(x, a_0)$ is also studied in [18], [19], where, in particular, an approximation scheme for the robust domain of attraction is presented based on ideas of optimal control. In this paper we concentrate on proving an existence and uniqueness result for a Zubov-type equation and examining properties of the solutions that can be obtained. Numerical aspects and actual examples are presented in [7].

In section 2 we begin defining robust domains of attraction for the class of systems under consideration and state some fundamental properties. In section 3 we define the generalization of (1.1) suitable for our case and discuss the question of solvability of this equation. For this we turn to the methodology of viscosity solutions. We refer to [3] for an introduction to this theory in the context of optimal control. Using viscosity solutions, we obtain an existence and uniqueness result for the generalized equation. In sections 4 and 5 we note some properties of the constructed solutions. In particular, the solutions can be interpreted as robust Lyapunov functions for the perturbed system, and via suitable choices of the parameters this Lyapunov function can be guaranteed to be globally Lipschitz, or smooth, at least, on subsets of the robust domain of attraction. Finally, in section 6 we provide a simple example illustrating our results.

2. The robust domain of attraction. Let $\varphi(t, x_0, a)$ be the solution of

$$\begin{cases} \dot{x}(t) = f(x(t), a(t)), & t \in [0, \infty), \\ x(0) = x_0, \end{cases} \tag{2.1}$$

where $a(\cdot) \in \mathcal{A} = L^\infty([0, +\infty), A)$ and A is a compact subset of $\mathbb{R}^m$. Throughout the paper the map f is taken to be continuous and bounded in $\mathbb{R}^n \times A$ and locally Lipschitz in x uniformly in $a \in A$. Furthermore, we assume that the fixed point $x = 0$ is singular; that is, $f(0, a) = 0$ for any $a \in A$.

We assume that the singular point 0 is uniformly locally exponentially stable for the system (2.1), i.e.,

(H1) there exist constants $C, \sigma, r > 0$ such that $\|\varphi(t, x_0, a)\| \le Ce^{-\sigma t}\|x_0\|$ for any $x_0 \in B(0, r)$ and any $a \in \mathcal{A}$.

We now define the following sets which describe domains of attraction for the equilibrium $x = 0$ of the system (2.1).

DEFINITION 2.1. *For the system* (2.1) *satisfying* (H1) *we define the* robust domain of attraction *as*

$$\mathcal{D} = \{x_0 \in \mathbb{R}^n : \varphi(t, x_0, a) \to 0 \text{ as } t \to +\infty \text{ for any } a \in \mathcal{A}\}$$

and the uniform robust domain of attraction *by*

$$\mathcal{D}_0 = \left\{ x_0 \in \mathbb{R}^n : \begin{array}{l} \text{there exists a function } \beta(t) \to 0 \text{ as } t \to \infty \\ \text{such that } \|\varphi(t, x_0, a)\| \le \beta(t) \text{ for all } t > 0,\ a \in \mathcal{A} \end{array} \right\}.$$

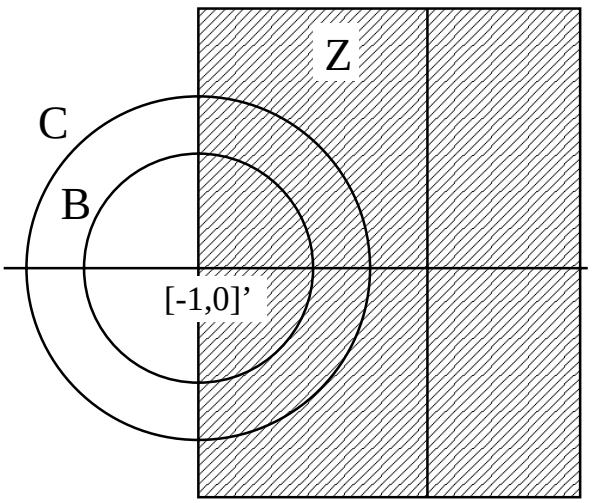


FIG. 2.1. *Sketch for Example* 2.1.

In order to obtain a different characterization of $\mathcal{D}_0$, we introduce the following "first hitting time" defined by $t(x,a) := \inf\{t > 0 : \ \varphi(t,x,a) \in B(0,r)\}$. Note that by the assumption on $B(0,r)$ there exists $T > 0$ independent of x and a such that $\varphi(t,x,a) \in B(0,r)$ for any $t \geq t(x,a) + T$.

LEMMA 2.2. *Assume* (H1); *then the robust domains of attraction* $\mathcal{D}$ *and* $\mathcal{D}_0$ *satisfy*

$$\mathcal{D} = \{x \in \mathbb{R}^n \, : \, t(x,a) < +\infty \text{ for any } a \in \mathcal{A}\} \, ,$$
$$\mathcal{D}_0 = \left\{ x \in \mathbb{R}^n \, : \, \sup_{a \in \mathcal{A}} \{t(x,a)\} < +\infty \right\} .$$

Proof. This is immediate from Definition 2.1. □

Before we begin analyzing some of the properties of $\mathcal{D}$ and $\mathcal{D}_0$, let us give an example that shows that for general nonlinear systems they are different.

Example 2.1. Let $n = 2$ and $y_0 = [\,-1\,,\,0\,]'$. We fix two discs around y_0 given by $B := B(y_0, 1/2)$ and $C := B(y_0, 3/4)$ and let $Z := \{x = [x_1, x_2] \in \mathbb{R}^2 \ : \ x_1 > -1\}$; see Figure 2.1.

Fix a C^∞ function $h : \mathbb{R}^2 \to \mathbb{R}$ such that

$$h \geq 0 \, , \quad h_{|B} \equiv 1 \, , \quad h_{|C^c} \equiv 0 \, .$$

Fix any $f : \mathbb{R}^2 \to \mathbb{R}^2$ such that $f_{|B} \equiv 0$, $f(0) = 0$, and such that the set $Z \setminus \operatorname{cl} B$ is contained in the domain of attraction of $x^* = 0$ for the system $\dot{x} = f(x)$. We may assume, furthermore, that for the first component function of f (denoted by f_1) we have $f_1(x) > 0$ on the annulus $C \setminus \operatorname{cl} B$. Now consider the system

$$\dot{x} = f(x) + h(x) \begin{bmatrix} (x_1+1)^2 + x_2^2 + 1 - a^2 \\ a \end{bmatrix} =: g(x,a) \, ,$$

where $a \in [-1,1]$. It is easy to see that for $x_0 \in B$ it holds that the first component of the solution $\varphi_1(t,x_0,a)$ is strictly increasing as long as $\varphi(t,x_0,a) \in B$. We even

have that for any $x_0 \in B \cap \operatorname{cl} Z$ and any $a \in \mathcal{A}$ there is a time $T = T(x_0, a)$ such that $\varphi(T, x_0, a) \notin B$. Also, by our assumption on f, h, and the construction of g, the first component of the solutions is strictly increasing on $C \setminus \operatorname{cl} B$. As a consequence, $y_0 \in \mathcal{D}$. On the other hand, for y_0 we have that for any time $t > 0$ and any $\varepsilon > 0$ there is some $a \in \mathcal{A}$ with $\|y_0 - \varphi(t, y_0, a)\| < \varepsilon$ by [12, Chap. 3, Theorem 6] as 0 is contained in the convex hull of $\{g(y_0, a) \ : \ a \in A\}$. Hence $t(y_0, a)$ is unbounded over $\mathcal{A}$, and so $y_0 \notin \mathcal{D}_0$.

In the following proposition we present some relevant properties of the (uniform) robust domain of attraction. Several of these bear a striking resemblance to those of the domain of attraction of an asymptotically stable fixed point of a time-invariant system; compare [11, Chap. IV]. It will frequently be convenient to consider the reachable set at time T from an initial condition $x_0 \in \mathbb{R}^n$ defined by

$$\mathcal{R}(x_0, T_0) := \{x \in \mathbb{R}^n \ : \ \exists\, 0 \le t \le T_0, a \in \mathcal{A} \text{ such that } x = \varphi(t, x_0, a)\}.$$

Note that by the boundedness of f it is immediate that the reachable set from a bounded set of initial conditions S given by

$$\mathcal{R}(S, T) := \bigcup_{x \in S} \mathcal{R}(x, T)$$

is bounded for any $T \ge 0$.

PROPOSITION 2.3. *Consider system* (2.1) *and assume* (H1)*; then the following hold.*

(i) $\operatorname{cl} B(0, r) \subset \mathcal{D}_0$.
(ii) $\mathcal{D}_0$ *is an open, connected, invariant set.* $\mathcal{D}$ *is a pathwise connected, invariant set.*
(iii) $\sup_{a \in \mathcal{A}}\{t(x, a)\} \to +\infty$ *for* $x \to x_0 \in \partial\mathcal{D}_0$ *or* $\|x\| \to \infty$.
(iv) $\mathcal{D} \subset \operatorname{cl} \mathcal{D}_0$.
(v) $\operatorname{cl} \mathcal{D}_0 = \operatorname{cl} \mathcal{D}$ *is an invariant set.*
(vi) $\mathcal{D}_0$, $\mathcal{D}$ *are contractible to* 0.
(vii) *If for some* $a_0 \in A$ $f(\cdot, a_0)$ *is of class* C^1*, then* $\mathcal{D}_0$ *is* C^1*-diffeomorphic to* $\mathbb{R}^n$.
(viii) *If for every* $x \in \partial\mathcal{D}_0$ *there exists* $a \in \mathcal{A}$ *such that* $\varphi(t, x, a) \in \partial\mathcal{D}_0$ *for all* $t \ge 0$*, then* $\mathcal{D} = \mathcal{D}_0$.
(ix) *If for all* $x \in \mathcal{D}$ *the set* $\{f(x, a) \ : \ a \in \mathcal{A}\}$ *is convex, then* $\mathcal{D}_0 = \mathcal{D}$.

Proof.

(i) This is a consequence of the exponential bound in (H1), which can easily be shown to extend to $\operatorname{cl} B(0, r)$.
(ii) Let $x_0 \in \mathcal{D}_0$ and $T_0 = \sup_{a \in \mathcal{A}}\{t(x_0, a)\}$. Then there exists T such that $\varphi(t, x_0, a) \in B(0, r/2C)$ for any $a \in \mathcal{A}$, $t \ge T$. Let δ be such that if $\|x_0 - x\| \le \delta$, then $\|\varphi(t, x_0, a) - \varphi(t, x, a)\| \le r/2C$ for any $t \le T$ and any $a \in \mathcal{A}$. Then $\varphi(t, x, a) \in B(0, r)$ for $t \ge T$ and $a \in \mathcal{A}$. Therefore, $x \in \mathcal{D}_0$, and it follows that $\mathcal{D}_0$ is open. By definition from each $x \in \mathcal{D}_0$ ($x \in \mathcal{D}$) there exists a trajectory $\varphi(\cdot, x, a)$ entering $B(0, r)$. This shows connectedness. To prove invariance assume that for some $x \in \mathcal{D}_0, a_1 \in \mathcal{A}$ there exists a $t > 0$ such that $y := \varphi(t, x, a_1) \notin \mathcal{D}_0$. This implies $\sup_{a \in \mathcal{A}}\{t(y, a)\} = \infty$. But clearly, $\sup_{a \in \mathcal{A}}\{t(x, a)\} \ge \sup_{a \in \mathcal{A}}\{t(y, a)\}$, contradicting the choice of x. A similar argument works for $\mathcal{D}$.
(iii) Let $x_n \to x_0 \in \partial\mathcal{D}_0$ and set $T_n = \sup_{a \in \mathcal{A}}\{t(x_n, a)\}$. If we assume that T_n is bounded and we take $r' < r$, we can find T such that, for any n, $\varphi(t, x_n, a) \in B(0, r')$ for any $t \ge T$ and for any $a \in \mathcal{A}$.

For any $\epsilon > 0$, there exists $\delta > 0$ such that if $\|x' - x''\| \leq \delta$, $\|\varphi(t, x', a) - \varphi(t, x'', a)\| \leq \epsilon$ for any $t \leq T$, for any $a \in \mathcal{A}$. Thus, setting $\epsilon = r - r'$ and choosing n sufficiently large such that $\|x_n - x_0\| \leq \delta$, we obtain that $\varphi(t, x_0, a) \in B(0, r)$ for any $t \geq T$ and for any $a \in \mathcal{A}$. Hence $x_0 \in \mathcal{D}_0$, which is a contradiction. The assertion is clear for $\|x_n\| \to \infty$, as our assumptions exclude solutions exploding in backward time.

(iv) The statement follows from an application of [20, Lemma III.2], which states that if $x \in \mathcal{D} \setminus \mathcal{D}_0$ or, equivalently, if $\sup_{a \in \mathcal{A}}\{t(x, a)\} = \infty$, while $t(x, a) < \infty$ for every $a \in \mathcal{A}$, then $x \in \partial\mathcal{D}$, as in every neighborhood of x there exists a point y and a control a_y such that $t(y, a_y) = \infty$.

(v) If for some $x \in \operatorname{cl}\mathcal{D}_0$ and $a \in \mathcal{A}$ we have $\varphi(t, x, a) \notin \operatorname{cl}\mathcal{D}_0$, then by continuous dependence on initial conditions we have that $\mathcal{D}_0$ is not invariant, contradicting (i). The equality of the two sets is an immediate consequence of (iv).

(vi) This follows by regarding the flow of $\dot{x} = f(x, a_0)$ for some $a_0 \in \mathcal{A}$.

(vii) In the proof we follow the outline given in [19]. Recall that a paracompact manifold M with the property that every compact subset of M has an open neighborhood that is diffeomorphic to $\mathbb{R}^n$ is itself diffeomorphic to $\mathbb{R}^n$; see [17, Lemma 3]. Let $K \subset \mathcal{D}_0$ be compact and consider a neighborhood U of K with $B(0, r) \subset U \subset \mathcal{D}_0$. Choose a relatively compact neighborhood U_2 of K with $B(0, r/2) \subset \operatorname{cl} U_2 \subset U$ and fix a C^∞ function $h : \mathbb{R}^n \to [0, 1]$ with $h_{|U_2} \equiv 1$ and $h_{|U^c} \equiv 0$. Now consider the system

$$\dot{x} = h(x) f(x, a_0)$$

with associated flow $\psi(t, x)$. It is clear that for some T large enough we have $K \subset \psi(-T, B(0, r/4)) \subset U$. This proves the assertion as $\psi(-T, B(0, r/4))$ is diffeomorphic to $B(0, r/4)$, which is in turn diffeomorphic to $\mathbb{R}^n$.

(viii) By the pathwise connectedness of $\mathcal{D}$ we have that $\mathcal{D} \cap \partial\mathcal{D}_0 \neq \emptyset$ if $\mathcal{D} \neq \mathcal{D}_0$. This contradicts our assumption.

(ix) Clearly, we need only show $\mathcal{D} \subset \mathcal{D}_0$. Assume that $x \in \mathcal{D}$ and there exist sequences $a_k \in \mathcal{A}$, $T_k \to \infty$ such that $\|\varphi(T_k, x, a_k)\| > r > 0$ for all $k \in \mathbb{N}$. By standard constructions there exists a subsequence (for which we use the index k again) which converges uniformly on compact time intervals to a solution $y(t)$ of the differential inclusion

$$\dot{y}(t) \in f(y(t), A)\,.$$

Now by convexity of $f(y(t), A), t \geq 0$, and Filippov's lemma [15, p. 267] there exists a control $a \in \mathcal{A}$ such that $y(t) = \varphi(t, x, a)$. By assumption there exists a t_0 such that $\|\varphi(t_0, x, a)\| < r/C$. As $\varphi(t_0, x, a_k) \to \varphi(t_0, x, a)$, this implies for all k large enough the inequality $\|\varphi(t, x, a_k)\| < r$ for $t \geq t_0$, which is a contradiction. □

3. Zubov's method for robust domains of attraction. It is our aim to show that the appropriate generalization of Zubov's equation (1.1) is given by

$$\inf_{a \in A}\{-Dv(x) f(x, a) - (1 - v(x)) g(x, a)\} = 0, \qquad x \in \mathbb{R}^n.$$

In this section we show the existence of a unique solution under a suitable "boundary condition" in the equilibrium $x = 0$. This solution will turn out to characterize the

uniform robust domain of attraction $\mathcal{D}_0$. Before turning to this equation, we introduce two optimal value functions and show certain properties of these functions.

Consider the following nonnegative, extended value function $V : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ by

$$V(x) = \sup_{a \in \mathcal{A}} \int_0^{+\infty} g(\varphi(t, x, a), a(t))dt \tag{3.1}$$

and its transformation via the Kruzkov transform

$$v(x) = 1 - e^{-V(x)}. \tag{3.2}$$

The function $g : \mathbb{R}^n \times A \to \mathbb{R}$ is supposed to be continuous and satisfies the following.

(H2)
- (i) For any $a \in A$, $g(0, a) = 0$ and $g(x, a) > 0$ for $x \neq 0$.
- (ii) There exists a constant $g_0 > 0$ such that $\inf_{x \notin B(0,r),\, a \in A} g(x, a) \geq g_0$.
- (iii) For every $R > 0$ there exists a constant L_R such that $\|g(x, a) - g(y, a)\| \leq L_R \|x - y\|$ for all $\|x\|$, $\|y\| \leq R$, and all $a \in A$.

Since g is nonnegative, it is immediate that $V(x) \geq 0$ and $v(x) \in [0, 1]$ for all $x \in \mathbb{R}^n$. Furthermore, standard techniques from optimal control (see, e.g., [3, Chapter III]) imply that V and v satisfy the dynamic programming principle; i.e., for each $t > 0$ we have

$$V(x) = \sup_{a \in \mathcal{A}} \left\{ \int_0^t g(\varphi(\tau, x, a), a(\tau))d\tau + V(\varphi(t, x, a)) \right\} \tag{3.3}$$

and

$$v(x) = \sup_{a \in \mathcal{A}} \{(1 - G(x, t, a)) + G(x, t, a)v(\varphi(t, x, a))\} \tag{3.4}$$

with

$$G(x, t, a) := \exp\left(- \int_0^t g(\varphi(\tau, x, a), a(\tau))d\tau \right). \tag{3.5}$$

The relation between V and v is immediate; we have

$$\begin{array}{lcl} V(x) = 0 & \Leftrightarrow & v(x) = 0\,, \\ V(x) \in (0, +\infty) & \Leftrightarrow & v(x) \in (0, 1)\,, \\ V(x) = +\infty & \Leftrightarrow & v(x) = 1\,. \end{array} \tag{3.6}$$

In the next proposition we investigate the relation between $\mathcal{D}_0$ and V (and thus also v) and the continuity of V and v.

PROPOSITION 3.1. *Assume* (H1) and (H2). *Then the following hold.*

(i) $V(x) < +\infty$ *if and only if* $x \in \mathcal{D}_0$.
(ii) $V(0) = 0$ *if and only if* $x = 0$.
(iii) V *is continuous on* $\mathcal{D}_0$.
(iv) $V(x) \to +\infty$ *for* $x \to x_0 \in \partial\mathcal{D}_0$ *and for* $\|x\| \to \infty$.
(v) $v(x) < 1$ *if and only if* $x \in \mathcal{D}_0$.
(vi) $v(0) = 0$ *if and only if* $x = 0$.
(vii) v *is continuous on* $\mathbb{R}^n$.

(viii) $v(x) \to 1$ *for* $x \to x_0 \in \partial\mathcal{D}_0$ *and for* $\|x\| \to \infty$.

Proof. (i) To show that $V(x_0) < +\infty$ for $x_0 \in \mathcal{D}_0$, observe that by Lemma 2.2 for each $x_0 \in \mathcal{D}_0$ there exists $T_0 > 0$ such that $\varphi(t, x_0, a) \in B(0, r)$ for all $t \geq T_0$ and all $a \in \mathcal{A}$. Also, the closure of the reachable set $\operatorname{cl}\mathcal{R}(x_0, T_0)$ is compact. Thus for any $a \in \mathcal{A}$

$$\begin{aligned}\int_0^{+\infty} g(\varphi(t), a(t))dt &\leq \int_0^{T_0} g(\varphi(t), a(t))dt + L_r \int_{T_0}^{+\infty} \|\varphi(t)\|dt \\ &\leq T_0 \sup_{x\in\mathcal{R}(x_0,T_0),a\in A} g(x, a) + L_r C \int_{T_0}^{+\infty} e^{-\sigma t} r dt \leq \tilde{C}\end{aligned}$$

with $\tilde{C}$ independent of $a \in \mathcal{A}$, and therefore, $V(x_0) < +\infty$.

Now let $x_0 \notin \mathcal{D}_0$. Then there exists a sequence $a_n \in \mathcal{A}$ such that $t(x_0, a_n)$ tends to ∞. Then for any $n \in \mathbb{N}$

$$\int_0^{+\infty} g(\varphi(t), a_n(t))dt \geq \int_0^{t(x_0,a_n)} g(\varphi(t), a_n(t))dt \geq g_0 t(x_0, a_n)\,,$$

where $g_0 > 0$ is defined as in (H2) (ii). It follows that $V(x) = +\infty$.

(ii) The proof follows immediately from (3.1), (H2) (i), and $f(0, a) = 0$.

(iii) Observe that

$$\begin{aligned}|V(x) - V(y)| &= \left|\sup_{a\in\mathcal{A}} \int_0^{+\infty} g(\varphi(t, x, a), a(t))dt - \sup_{a\in\mathcal{A}} \int_0^{+\infty} g(\varphi(t, y, a), a(t))dt\right| \\ &\leq \sup_{a\in\mathcal{A}} \int_0^{+\infty} |g(\varphi(t, x, a), a(t)) - g(\varphi(t, y, a), a(t))|\, dt\,. \qquad (3.7)\end{aligned}$$

We first prove that V is continuous on $B(0, r/C)$.

Fix some $x_0 \in B(0, r/C)$. Then (H1) and (H2) (iii) imply

$$\begin{aligned}\int_0^{+\infty} g(\varphi(t, x_0, a), a(t))dt &\leq L_r \int_0^{+\infty} \|\varphi(t, x_0, a)\|dt \\ &\leq L_r C \int_0^{+\infty} e^{-\sigma t}\|x_0\|dt \leq C_1\|x_0\|.\end{aligned}$$

Fix $\epsilon > 0$. From (H1) we can conclude that there exists $T > 0$ such that $C_1\|\varphi(t, x, a)\| \leq \epsilon/4$ for all $t \geq T$ and all $x \in B(0, r/C)$. Abbreviate $L = L_{r/C}$. Then by Lipschitz continuity of f there exists a $\delta > 0$ such that $\|\varphi(t, x_0, a) - \varphi(t, y_0, a)\| < \epsilon/(2LT)$ for all $t \in [0, T]$ and all $y_0 \in B(0, r/C)$ with $\|x_0 - y_0\| < \delta$.

Putting this together yields for every $a \in \mathcal{A}$

$$\begin{aligned}&\int_0^{+\infty} |g(\varphi(t, x_0, a), a(t)) - g(\varphi(t, y_0, a), a(t))|\, dt \\ &\quad\leq \int_0^T L\,\|\varphi(t, x_0, a) - \varphi(t, y_0, a)\|\, dt + C_1\|\varphi(T, x_0, a)\| + C_1\|\varphi(T, y_0, a)\| \\ &\quad\leq \epsilon/2 + \epsilon/4 + \epsilon/4 \;\leq\; \epsilon\,,\end{aligned}$$

which by (3.7) implies continuity.

For $x_0 \in \mathcal{D}_0$ we can use openness of $\mathcal{D}_0$ in order to conclude that there exists an open neighborhood N of x_0 and $T > 0$ such that $\varphi(t, y_0, a) \in B(0, r/C)$ for all

$y_0 \in N$, all $a \in \mathcal{A}$, and all $t \geq T$. Thus (3.3) and the continuity on $B(0, r/C)$ imply continuity in x_0.

(iv) The proof follows immediately from Proposition 2.3 (iii) since $\mathcal{D}_0$ is open and $g(x) \geq g_0 > 0$ for x outside $B(0, r)$ as assumed in (H2).

(v) and (vi) follow immediately from (3.6), (i), and (ii); (vii) follows from (3.6), (iii), and (iv); and (viii) follows from (3.6) and (iv). □

We now turn to the formulation of suitable partial differential equations for which V and v form solutions. Since in general these functions will not be differentiable, we have to work with a more general solution concept, namely, viscosity solutions.

Let us recall the definition of viscosity solutions. (For more details about this theory we refer to [3].)

DEFINITION 3.2. *Given an open subset Ω of $\mathbb{R}^n$ and a continuous function $H : \Omega \times \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$, we say that a lower semicontinuous (l.s.c.) function $u : \Omega \to \mathbb{R}$ (resp., an upper semicontinuous (u.s.c.) function $v : \Omega \to \mathbb{R}$) is a viscosity supersolution (resp., subsolution) of the equation*

$$H(x, u, Du) = 0, \qquad x \in \Omega, \tag{3.8}$$

if for all $\phi \in C^1(\Omega)$ and $x \in \operatorname{argmin}_\Omega(u - \phi)$ (resp., $x \in \operatorname{argmax}_\Omega(v - \phi)$) we have

$$H(x, u(x), D\phi(x)) \geq 0 \qquad \big(\textit{resp., } H(x, v(x), D\phi(x)) \leq 0\big).$$

A continuous function $u : \Omega \to \mathbb{R}$ is said to be a viscosity solution of (3.8) *if u is a viscosity supersolution and a viscosity subsolution of* (3.8).

Remark 3.1. It is not difficult to see (cf. [3, Lemma II.1.7]) that the set of derivatives $D\phi(x)$ for $x \in \operatorname{argmin}_\Omega(u - \phi)$ coincides with the set

$$D^- u(x) := \{p \in \mathbb{R}^n \mid u(x) - u(y) - p(x - y) \geq -o(\|x - y\|) \text{ for all } y \in \mathbb{R}^n\}$$

and that the set of derivatives $D\phi(x)$ for $x \in \operatorname{argmax}_\Omega(v - \phi)$ equals

$$D^+ v(x) := \{p \in \mathbb{R}^n \mid v(x) - v(y) - p(x - y) \leq o(\|x - y\|) \text{ for all } y \in \mathbb{R}^n\}.$$

Hence one can alternatively define viscosity solutions via the sets D^- and D^+, the so called *sub- and superdifferentials*. Note that if a function $w : \Omega \to \mathbb{R}$ is differentiable in some $x \in \Omega$, the equality $D^+ w(x) = D^- w(x) = \{Dw(x)\}$ follows; hence for smooth functions viscosity solutions coincide with classical solutions.

Recalling that V is locally bounded in $\mathcal{D}_0$ and v is locally bounded on $\mathbb{R}^n$, the following proposition follows from an easy application of the dynamic programming principles (3.3) and (3.4); cf. [3, Chapter III].

PROPOSITION 3.3. *V is a viscosity solution of*

$$\inf_{a \in A} \{-DV(x) f(x, a) - g(x, a)\} = 0, \qquad x \in \mathcal{D}_0, \tag{3.9}$$

and v is a viscosity solution of

$$\inf_{a \in A} \{-Dv(x) f(x, a) - (1 - v(x)) g(x, a)\} = 0, \qquad x \in \mathbb{R}^n. \tag{3.10}$$

Observe that (3.10) is the straightforward generalization of the classical Zubov equation (1.1) [26] multiplied by -1, which is necessary in order to obtain the proper sign for viscosity sub- and supersolutions. Equation (3.9), however, shows that also our "auxiliary function" V can be characterized as the solution of a suitable PDE.

In order to get a uniqueness result, we use the following super- and suboptimality principles. Our approach is closely related to that of Soravia [21, 22]; we quote the following result from [21].

THEOREM 3.4 (see [21, Theorem 3.2 (i)]). *Consider the equation*

$$\sup_{a\in A}\{-Du(x)f(x,a) - h(x,a) + k(x,a)u(x)\} = 0\,. \tag{3.11}$$

Then if u is a u.s.c. subsolution of (3.11)*, then it satisfies the lower optimality principle*

$$u(x) = \inf_{a\in\mathcal{A}}\inf_{t\geq 0}\left[\int_0^t \exp\left(-\int_0^s k(\varphi(r),a(r))dr\right) h(\varphi(s),a(s))ds + \exp\left(-\int_0^t k(\varphi(t),a(t))ds\right) u(\varphi(t))\right].$$

Recalling the definition of G from (3.5), we see that this result has immediate applications for (3.9), (3.10).

PROPOSITION 3.5.

(i) *Let w be an l.s.c. supersolution of* (3.10) *in $\mathbb{R}^n$; then for any $x\in\mathbb{R}^n$*

$$w(x) = \sup_{a\in\mathcal{A}}\sup_{t\geq 0}\{(1-G(x,t,a)) + G(x,t,a)w(\varphi(t))\}\,. \tag{3.12}$$

(ii) *Let W be an l.s.c. supersolution of* (3.9) *in $\mathcal{D}_0$; then for any $x\in\mathcal{D}_0$*

$$W(x) = \sup_{a\in\mathcal{A}}\sup_{t\geq 0}\left\{\int_0^t g(\varphi(s),a(s))ds + W(\varphi(t))\right\}. \tag{3.13}$$

(iii) *Let u be a u.s.c. subsolution of* (3.10) *in $\mathbb{R}^n$, and let $\tilde{u}:\mathbb{R}^n\to\mathbb{R}$ be a continuous function with $u\leq\tilde{u}$. Then for any $x\in\mathbb{R}^n$ and any $T\geq 0$*

$$u(x) \leq \sup_{a\in\mathcal{A}}\inf_{t\in[0,T]}\{(1-G(x,t,a)) + G(x,t,a)\tilde{u}(\varphi(t))\}\,. \tag{3.14}$$

(iv) *Let U be a u.s.c. subsolution of* (3.9) *in $\mathcal{D}_0$, and let $\tilde{U}:\mathcal{D}_0\to\mathbb{R}$ be a continuous function with $U\leq\tilde{U}$. Then for any $x\in\mathcal{D}_0$ and any $T\geq 0$*

$$U(x) \leq \sup_{a\in\mathcal{A}}\inf_{t\in[0,T]}\left\{\int_0^t g(\varphi(s),a(s))ds + \tilde{U}(\varphi(t))\right\}. \tag{3.15}$$

Proof. If w is an l.s.c. supersolution of (3.10), then it follows by multiplication by -1 and an application of the definition that $-w$ is a u.s.c. subsolution of

$$\sup_{a\in A}\{-Du(x)f(x,a) + (1+u(x))g(x,a)\} = 0, \qquad x\in\mathbb{R}^n\,. \tag{3.16}$$

This implies that we can directly apply Theorem 3.4 for the special case $h\equiv -g, k\equiv g$ to obtain that $-w$ satisfies

$$-w(x) = \inf_{a\in\mathcal{A}}\inf_{t\geq 0}\left[-\int_0^t \exp\left(-\int_0^s g(\varphi(r),a(r))dr\right) g(\varphi(s),a(s))ds - \exp\left(-\int_0^t g(\varphi(t),a(t))ds\right) w(\varphi(t))\right].$$

Now the assertion follows upon multiplication by -1 and using the fact that

$$\int_0^t G(\varphi(s), s, a(s))g(\varphi(s), a(s))ds = 1 - G(x, t, a)\,.$$

(ii) follows by insertion of $k \equiv 0, h \equiv -g$ in (3.11).

For the proof of (iii) we follow the ideas of [21] with minor modifications. Let $u : \mathbb{R}^n \to \mathbb{R}$ be a u.s.c. subsolution of (3.10), let $\tilde{u} : \mathbb{R}^n \to \mathbb{R}$ be a continuous function with $u \le \tilde{u}$, and define $u^* := -u$ and $\tilde{u}^* := -\tilde{u}$. Again, a straightforward verification of the definition shows that u^* is an l.s.c. viscosity supersolution of

$$\sup_{a\in A}\{-Dw(x)f(x,a) + (1 + w(x))g(x,a)\} = 0, \qquad x \in \mathbb{R}^n. \tag{3.17}$$

From this equation it is easy to see that the auxiliary function $\bar{u} : \mathbb{R}^{n+2} \to \mathbb{R}$ given by $\bar{u}(x, r, s) = e^{-s}u^*(x) + r$ is an l.s.c. supersolution of

$$\sup_{a\in A}\left\{-e^{-s}D_x v(x,r,s)f(x,a) + D_r v(x,r,s)e^{-s}g(x,a) - D_s v(x,r,s)g(x,a)\right\} = 0$$

for $x \in \mathbb{R}^n$, r, $s \in \mathbb{R}$.

We now introduce a change of variables by choosing $\rho : \mathbb{R} \to \mathbb{R}_+$ smooth, bounded, and such that $0 < \dot{\rho} \le M$ and $\rho(s) \to 0$ as $s \to -\infty$. Now consider the function

$$U(z) = U(x, r, s) := \rho(\bar{u}(x, r, s)) = \rho(e^{-s}u^*(x) + r)\,.$$

By the rules for changes of variables (cf. [3, Proposition II.2.5]) it can be shown that U is an l.s.c. supersolution of

$$\sup_{a\in A}\{-D_z u(z)F(z,a)\} = 0\,, \quad z \in \mathbb{R}^{n+2}\,, \tag{3.18}$$

where the underlying dynamics is given by

$$\dot{z} = \begin{pmatrix} \dot{x} \\ \dot{r} \\ \dot{s} \end{pmatrix} = F(z(t), a(t)) = \begin{pmatrix} f(\varphi(t), a(t)) \\ -e^{-s(t)}g(\varphi(t), a(t)) \\ g(\varphi(t), a(t)) \end{pmatrix}. \tag{3.19}$$

Note that the solution to this system corresponding to an initial value $z = (x, 0, 0)$ is given by

$$z(t) = \left[\varphi(t, x, a), (G(t, x, a) - 1), \int_0^t g(\varphi(s), a(s))ds\right]'. \tag{3.20}$$

In order to apply results from [21, Appendix] we need that F satisfies a global Lipschitz condition. Since this is not true in general, we localize the problem by considering for $k \in \mathbb{N}$ the family of smooth functions $\zeta_k : \mathbb{R}^{n+2} \to \mathbb{R}$ with $0 \le \zeta_k \le 1$, $\zeta_k \equiv 1$ in $B(0, k) \subset \mathbb{R}^{n+2}$, $\zeta_k \equiv 0$ in $B(0, k + 1)^c$, $|D\zeta_k| \le 2$, and setting $F_k = \zeta_k F$.

Then from (3.18) we can conclude that for each $k \in \mathbb{N}$ the function U is also a supersolution of

$$\sup_{a\in A}\{-D_z u(z)F_k(z,a)\} = 0\,,$$

as the multiplication with the nonnegative function ζ_k does not affect the inequality that a supersolution has to fulfill.

Now consider the continuous function $\phi : \mathbb{R}^{n+2} \to \mathbb{R}\,, \phi(z) = \phi(x,s,r), x \in \mathbb{R}^n, s, r \in \mathbb{R}$,

$$\phi(z) = \rho(e^{-s}\tilde{u}^*(x) + r)\,.$$

Since $U \geq 0$ (by the choice of ρ), we obtain for any fixed $\lambda > 0$ that U is also a supersolution of

$$\lambda u + \min\left\{\sup_{a\in A}\{-D_z u(z) F_k(z,a)\}\,,\; u - (1+\lambda)\phi\right\} = 0. \tag{3.21}$$

This equation has a unique continuous viscosity solution and it can be shown [21, Appendix] that this solution is given by the value function

$$V_k^\lambda(z) := \inf_{a\in\mathcal{A}} \sup_{t\geq 0} e^{-\lambda t}\phi(z_k(t)),$$

where $z_k(\cdot)$ solves $\dot{z}_k(t) = F_k(z_k(t), a(t))$, $z_k(0) = z$. By the usual comparison theorem for semicontinuous viscosity solutions (see, e.g., [3, Chapter V]), we obtain $U \geq V_k^\lambda$ for each $\lambda > 0$ and each $k \in \mathbb{N}$. Hence letting $\lambda \to 0$ yields for all $k \in \mathbb{N}$ and all $T > 0$ the inequality

$$\rho(e^{-s}u^*(x) + r) = U(z) \geq \inf_{a\in\mathcal{A}} \sup_{t\in[0,T]} \phi(z_k(t)).$$

By the boundedness of f the reachable set $\mathcal{R}(x,T)$ is bounded for each $x \in \mathbb{R}^n, T > 0$. Hence for each $z = (x,0,0) \in \mathbb{R}^{n+2}$ and each $T > 0$ there exists a $k \in \mathbb{N}$ such that $z(t) \in B(0,k)$ for all $a \in \mathcal{A}$ and all $t \in [0,T]$. Furthermore, on $B(0,k)$ the trajectories $z(\cdot)$ and $z_k(\cdot)$ coincide, and thus we can conclude by (3.20) and by the definition of ϕ that

$$\rho(u^*(x)) = U(x,0,0) \geq \inf_{a\in\mathcal{A}} \sup_{t\in[0,T]} \phi(z(t))$$

$$= \inf_{a\in\mathcal{A}} \sup_{t\in[0,T]} \rho\left((G(x,t,a) - 1) + G(x,t,a)\tilde{u}^*(\varphi(t))\right).$$

Using the monotonicity of ρ, we obtain

$$u^*(x) \geq \inf_{a\in\mathcal{A}} \sup_{t\in[0,T]} \left\{(G(x,t,a) - 1) + G(x,t,a)\tilde{u}^*(\varphi(t))\right\},$$

and hence

$$u(x) \leq \sup_{a\in\mathcal{A}} \inf_{t\in[0,T]} \left\{(1 - G(x,t,a)) + G(x,t,a)\tilde{u}(\varphi(t))\right\}$$

holds for each $T \geq 0$, which shows (iii).

Assertion (iv) is proved analogously. □

We can now apply these principles to the generalized version of Zubov's equation (3.10).

PROPOSITION 3.6. *Let w be a bounded l.s.c. supersolution of* (3.10) *on $\mathbb{R}^n$ with $w(0) \geq 0$. Then $w \geq v$ for v as defined in* (3.2).

Proof. First observe that the lower semicontinuity of w and the assumption $w(0) \geq 0$ imply that for each $\epsilon > 0$ there exists a $\delta > 0$ such that

$$w(x) \geq -\epsilon \text{ for all } x \in \mathbb{R}^n \text{ with } \|x\| \leq \delta. \tag{3.22}$$

Furthermore, the upper optimality principle (3.12) implies

$$w(x_0) \geq \sup_{a \in \mathcal{A}} \inf_{t \geq 0} \{1 + G(x_0, t, a)(w(\varphi(t, x_0, a)) - 1)\} . \tag{3.23}$$

Now we distinguish two cases.

(i) $x_0 \in \mathcal{D}_0$. In this case we know that for each $a \in \mathcal{A}$ we have $\varphi(t, x_0, a) \to 0$ as $t \to \infty$. Thus from (3.22) and (3.23), and using the definition of v, we can conclude

$$w(x_0) \geq \sup_{a \in \mathcal{A}} \left\{ \lim_{t \to \infty} (1 - G(x_0, t, a)) \right\} = v(x_0) ,$$

which shows the claim.

(ii) $x_0 \notin \mathcal{D}_0$. In this case by (3.6) and Proposition 3.1(v) it is sufficient to show that $w(x_0) \geq 1$. By the definition of $\mathcal{D}_0$ we know that for each $T > 0$ that there exists $a_T \in \mathcal{A}$ such that $t(x_0, a_T) > T$, which implies $G(x_0, T, a_T) \leq \exp(-Tg_0)$, which tends to 0 as $T \to \infty$. Thus, denoting the bound on $|w|$ by $M > 0$, the inequality (3.23) implies

$$w(x_0) \geq (1 - \exp(-Tg_0)) - \exp(-Tg_0)M$$

for every $T > 0$, and hence $w(x_0) \geq 1$. □

PROPOSITION 3.7. *Let u be a bounded u.s.c. subsolution of* (3.10) *on $\mathbb{R}^n$ with $u(0) \leq 0$. Then $u \leq v$ for v defined in* (3.2).

Proof. By the upper semicontinuity of u and $u(0) \leq 0$ we obtain that for every $\epsilon > 0$ there exists a $\delta > 0$ with $u(x) \leq \epsilon$ for all $x \in \mathbb{R}^n$ with $\|x\| \leq \delta$. Thus for each $\epsilon > 0$ we find a bounded and continuous function $\tilde{u}_\epsilon : \mathbb{R}^n \to \mathbb{R}$ with

$$\tilde{u}_\epsilon(0) < \epsilon \text{ and } u \leq \tilde{u}_\epsilon. \tag{3.24}$$

Now the lower optimality principle (3.14) implies for every $t \geq 0$ that

$$u(x_0) \leq \sup_{a \in \mathcal{A}} \{1 + G(x_0, t, a)(\tilde{u}_\epsilon(\varphi(t, x_0, a)) - 1)\} . \tag{3.25}$$

Again, we distinguish two cases.

(i) $x_0 \in \mathcal{D}_0$. In this case $\|\varphi(t, x_0, a)\| \to 0$ as $t \to \infty$ uniformly in $a \in \mathcal{A}$. Hence for each $\epsilon > 0$ there exists $t_\epsilon > 0$ such that

$$\tilde{u}_\epsilon(\varphi(t_\epsilon, x_0, a)) \leq \epsilon \quad \text{and} \quad |G(x_0, t_\epsilon, a) - G(x_0, \infty, a)| \leq \epsilon$$

for all $a \in \mathcal{A}$. Thus from (3.24) and (3.25), and using the definition of v, we can conclude

$$u(x_0) \leq \sup_{a \in \mathcal{A}} \{1 - (1 - \epsilon)G(x_0, t_\epsilon, a)\} \leq v(x_0) + \epsilon(1 - v(x_0)) + \epsilon ,$$

which shows the claim since v is bounded and $\epsilon > 0$ was arbitrary.

(ii) $x_0 \notin \mathcal{D}_0$. In this case by (3.6) and Proposition 3.1(v) it is sufficient to show that $u(x_0) \leq 1$. By (i) we know that $u(y) \leq v(y) < 1$ for each $y \in \mathcal{D}_0$; hence, analogous to (3.24) for each $\epsilon > 0$, we can conclude the existence of a continuous $\tilde{u}_\epsilon$ with $u \leq \tilde{u}_\epsilon$ and $\tilde{u}_\epsilon(y) \leq 1 + \epsilon$ for each $y \in \mathcal{D}_0$. Since u is bounded by assumption, we may choose $\tilde{u}_\epsilon$ such that $M_\epsilon := \sup_{x \in \mathbb{R}^n} \tilde{u}_\epsilon(x) < \infty$. If $M_\epsilon \leq 1$ for some $\epsilon > 0$, we are done. Otherwise, fix $\epsilon > 0$ and consider a sequence $t_n \to \infty$. Then (3.25) implies that there exists a sequence $a_n \in \mathcal{A}$ with

$$u(x_0) - \epsilon \leq 1 + G(x_0, t_n, a_n)(\tilde{u}_\epsilon(\varphi(t_n, x_0, a_n)) - 1).$$

If $\varphi(t_n, x_0, a_n) \in \mathcal{D}_0$, we know that $\tilde{u}_\epsilon(\varphi(t_n, x_0, a_n)) \leq 1 + \epsilon$, and since $G \leq 1$, we obtain $u(x_0) - \epsilon \leq 1 + \epsilon$. If $\varphi(t_n, x_0, a_n) \notin \mathcal{D}_0$, then $G(x_0, t_n, a_n) \leq \exp(-g_0 t_n)$; thus

$$1 + G(x_0, t_n, a_n)(\tilde{u}_\epsilon(\varphi(t_n, x_0, a_n)) - 1) \leq 1 + \exp(-g_0 t_n)(M_\epsilon - 1).$$

Thus for each $n \in \mathbb{N}$ we obtain

$$u(x_0) \leq 2\epsilon + 1 + \exp(-g_0 t_n)(M_\epsilon - 1)\,,$$

which for $n \to \infty$ implies $u(x_0) \leq 1 + 2\epsilon$. This proves the assertion since $\epsilon > 0$ was arbitrary. □

Using these propositions, we can now formulate an existence and uniqueness theorem for the generalized version of Zubov's equation (3.10).

THEOREM 3.8. *Consider the system* (2.1) *and a function* $g : \mathbb{R}^n \times A \to \mathbb{R}$ *such that* (H1) *and* (H2) *are satisfied. Then* (3.10) *has a unique bounded and continuous viscosity solution* v *on* $\mathbb{R}^n$ *satisfying* $v(0) = 0$.

This function coincides with v *from* (3.2). *In particular, the characterization* $\mathcal{D}_0 = \{x \in \mathbb{R}^n \mid v(x) < 1\}$ *holds.*

Proof. This is immediate from Propositions 3.6 and 3.7. □

For the sake of completeness we state the following analogous result for (3.9), which is proved with the same techniques, using (3.13) and (3.15) instead of (3.12) and (3.14). Observe that this result corresponds to the one in [4].

THEOREM 3.9. *Consider the system* (2.1) *and a function* $g : \mathbb{R}^n \times A \to \mathbb{R}$. *Assume* (H1) *and* (H2). *Let* $\mathcal{O} \subset \mathbb{R}^n$ *be an open set containing the origin, and let* $U : \mathcal{O} \to \mathbb{R}$ *be a positive and continuous function which is a viscosity solution of* (3.9) *on* $\mathcal{O}$ *and satisfies* $U(0) = 0$ *and* $U(x) \to \infty$ *for* $x \to \partial\mathcal{O}$ *and for* $|x| \to \infty$.

Then U *coincides with* V *from* (3.1), *and* $\mathcal{O} = \mathcal{D}_0$. *In particular, the function* V *from* (3.1) *is the unique positive continuous viscosity solution of* (3.9) *on* $\mathcal{D}_0$ *with* $V(0) = 0$.

For practical purposes, Theorem 3.8 might be inconvenient since we have to compute (or verify) a solution of (3.10) on the whole $\mathbb{R}^n$. The following fact can be exploited to show that this is not always necessary.

Remark 3.2. The optimality principles (i) and (iii) of Proposition 3.5 also hold if we have viscosity sub- or supersolutions of (3.10), which are defined only on some proper open subset $\mathcal{O} \subset \mathbb{R}^n$, except that in this case the "inf" and "sup" over the time t is taken only up to the first time when the trajectory under consideration leaves $\mathcal{O}$. More precisely, (3.12) becomes

$$w(x) = \sup_{a \in \mathcal{A}} \sup_{t \in [0, \tau_x(a)]} \{(1 - G(x,t,a)) + G(x,t,a) w(\varphi(t))\}\,, \tag{3.26}$$

and (3.14) becomes

$$u(x) \leq \sup_{a \in \mathcal{A}} \inf_{t \in [0, \tau_x(a)]} \{(1 - G(x,t,a)) + G(x,t,a) \tilde{u}(\varphi(t))\}\,, \tag{3.27}$$

where $\tau_x(a) := \inf\{t \geq 0 \,|\, \varphi(t,x,a) \notin \mathcal{O}\}$. We refer to [22] for a proof using the same arguments as in the $\mathbb{R}^n$ case combined with a localization technique.

Using these "nonglobal" optimality principles, we are now able to state nonglobal versions of the Propositions 3.6 and 3.7.

PROPOSITION 3.10. *Consider some open set $\mathcal{O} \subset \mathbb{R}^n$. Let $w : \mathrm{cl}\,\mathcal{O} \to \mathbb{R}$ be a bounded l.s.c. supersolution of* (3.10) *on $\mathcal{O}$ with $w(0) \geq 0$ and $w(x) \geq 1$ for all $x \in \partial\mathcal{O}$. Then $w \geq v|_{\mathcal{O}}$ for v as defined in* (3.2).

Proof. The proof follows with the same techniques as the proof of Proposition 3.6 using (3.26) instead of (3.12). ∎

In contrast to Proposition 3.10, we have to strengthen the assumption of Proposition 3.7 in order to get the corresponding nonglobal result.

PROPOSITION 3.11. *Consider some open set $\mathcal{O} \subset \mathbb{R}^n$. Let $u : \mathrm{cl}\,\mathcal{O} \to \mathbb{R}$ be a bounded continuous subsolution of* (3.10) *on $\mathcal{O}$ with $u(0) \leq 0$ and $u(x) = 1$ for all $x \in \partial\mathcal{O}$. Then $u \leq v|_{\mathcal{O}}$ for v as defined in* (3.2).

Proof. It is sufficient to show that $\mathcal{D}_0 \subseteq \mathcal{O}$ since in this case we get $v|_{\partial\mathcal{O}} \equiv 1$ and thus obtain the assertion with the same techniques as in the proof of Proposition 3.7 using (3.27) instead of (3.14).

In order to show $\mathcal{D}_0 \subseteq \mathcal{O}$, assume that $\mathcal{D}_0 \not\subseteq \mathcal{O}$. Then we obtain

$$r_0 := \sup\{r > 0 \,|\, \{x \in \mathbb{R}^n \,|\, v(x) \leq r\} \subset \mathcal{O}\} < 1.$$

We set $S := \{x \in \mathbb{R}^n \,|\, v(x) \leq r_0\}$. Note that from the optimality principle (3.4) we immediately obtain that v is strictly decreasing along each trajectory $\varphi(t,x_0,a)$; hence $\varphi(t,x_0,a) \in \mathrm{int}S \subseteq \mathcal{O}$ for all $t > 0, a \in \mathcal{A}$. By definition of r_0 there exists $x_0 \in \partial\mathcal{O}$ with $v(x_0) = r_0$ and $u(x_0) = 1$; hence by continuity of u there exists $\epsilon > 0$ and $\eta > 0$ such that $u(x) > r_0 + \epsilon$ for all $x \in \mathcal{O} \cap B(x_0,\eta)$. Fixing some arbitrary $a^* \in \mathcal{A}$ and some $\tau > 0$ sufficiently small, we set $x_1 := \varphi(\tau,x_0,a^*) \in \mathcal{O} \cap B(x_0,\eta)$. Then $\varphi(t,x_1,a) \in \mathrm{int}S \subseteq \mathcal{O}$ for all $t \geq 0, a \in \mathcal{A}$; i.e., the trajectory never reaches $\partial\mathcal{O}$, implying that (3.27) coincides with (3.14). (Note that by continuity of u we can choose $\tilde{u} = u$.) Thus we obtain

$$r_0 + \epsilon \leq u(x_1) \leq \sup_{a\in\mathcal{A}} \inf_{t\in[0,T]} \{(1 - G(x,t,a)) + G(x,t,a)u(\varphi(t,x_1,a))\}$$

for all $T > 0$. Since u is continuous with $u(0) \leq 0$ and $\varphi(t,x_1,a) \to 0$ as $t \to \infty$, we obtain by letting $T \to \infty$

$$r_0 + \epsilon \leq u(x_1) \leq \lim_{t\to\infty} \sup_{a\in\mathcal{A}} \{(1 - G(x,t,a))\} = v(x_1) \leq r_0,$$

which is a contradiction and hence shows $\mathcal{D}_0 \subseteq \mathcal{O}$. ∎

From these propositions we can now easily deduce the following theorem. It shows that we can restrict ourselves to a proper open subset $\mathcal{O}$ of the state space and still obtain our solution v, provided $\mathcal{D}_0 \subseteq \mathcal{O}$. Conversely, if $\mathcal{D}_0 \not\subseteq \mathcal{O}$, then no viscosity solution v with $v(x) = 1$ for all $x \in \partial\mathcal{O}$ can exist.

THEOREM 3.12. *Consider the system* (2.1) *and a function $g : \mathbb{R}^n \times A \to \mathbb{R}$. Assume* (H1) *and* (H2). *Let $\mathcal{O} \subset \mathbb{R}^n$ be an open set containing the origin, and let $v : \mathrm{cl}\mathcal{O} \to \mathbb{R}$ be a bounded and continuous function which is a viscosity solution of* (3.10) *on $\mathcal{O}$ and satisfies $v(0) = 0$ and $v(x) = 1$ for all $x \in \partial\mathcal{O}$.*

Then v coincides with the restriction $v|_{\mathcal{O}}$ of the function v from (3.2). *In particular, the characterization $\mathcal{D}_0 = \{x \in \mathcal{O} \,|\, v(x) < 1\}$ holds.*

Proof. The proof follows immediately from Propositions 3.10 and 3.11. ∎

4. Further properties of the solution. In this section we collect several properties of the solution v of Zubov's equation from Theorem 3.8. In particular, we show that this solution is a robust Lyapunov function on $\mathcal{D}_0$ and that additional assumptions on g ensure Lipschitz continuity of v.

THEOREM 4.1. *The function v is a robust Lyapunov function for the system* (2.1). *More precisely, we have*

$$v(\varphi(t,x_0,a(\cdot)))-v(x_0) \le \left[1-\exp\left(-\int_0^t g(\varphi(\tau),a(\tau))d\tau\right)\right](v(\varphi(t,x_0,a(\cdot)))-1) < 0$$

for all $x_0 \in \mathcal{D}_0 \setminus \{0\}$ and all $a(\cdot) \in \mathcal{A}$. In particular, each sublevel set of v is positively invariant.

Proof. The dynamic programming principle (3.4) implies

$$v(x) \ge 1-\exp\left(\int_0^t g(\varphi(\tau,x,a),a(\tau))d\tau\right)+\exp\left(\int_0^t g(\varphi(\tau,x,a),a(\tau))d\tau\right)v(\varphi(t,x,a))$$

for each $a \in \mathcal{A}$. This immediately yields the assertion. ∎

Remark 4.1.

(i) If v is differentiable in some point $0 \neq x_0 \in \mathcal{D}_0$, this yields the more familiar inequality

$$\sup_{a\in A} Dv(x_0)f(x,a) \le (v(x_0)-1)g(x,a) < 0,$$

which, in fact, can also be directly derived from (3.10).

(ii) It follows immediately from Proposition 3.5 (ii) that any viscosity supersolution w of (3.10) with $w(0)=0$ is a robust Lyapunov function on its sublevel set $\{x \in \mathbb{R}^d \,|\, w(x) < 1\}$.

Now we investigate regularity properties for the function v. In general, we cannot expect this function to be differentiable. A suitable choice of g, however, guarantees Lipschitz continuity. We start by investigating this for the function V defined in (3.1).

PROPOSITION 4.2. *Assume* (H1) *and* (H2) *and that $f(\cdot,a)$ is locally Lipschitz continuous uniformly in a; i.e., for any $R>0$ there exists a constant M_R such that*

$$\|f(x,a)-f(y,a)\| \le M_R\|x-y\| \quad \text{for all } x,y \in B(0,R),\ a \in A\,.$$

Assume, furthermore, that there exists a neighborhood N of the origin such that for all x, $y \in N$ the inequality

$$|g(x,a)-g(y,a)| \le K\max\{\|x\|,\|y\|\}^s\|x-y\|$$

holds for some $K>0$ and $s > M_r/\sigma$ with $r>0, \sigma>0$ as in (H1). *Then V is locally Lipschitz in $\mathcal{D}_0$.*

Proof. Let $S \subset \mathcal{D}_0$ be compact. According to (H2), there exists a time $T>0$ such that $\varphi(t,x,a) \in N \cap B(0,r)$ for all $t \ge T, x \in S, a \in \mathcal{A}$. Furthermore, the set $\mathcal{R}(S,T)$ is bounded, and we may choose $R>0$ large enough so that $\mathcal{R}(S,T) \subset B(0,R)$. Now fix x, $y \in S$. Analogously to the proof of Proposition 3.1(iii), we obtain

$$\begin{aligned}|V(x)-V(y)| &\le \sup_{a\in\mathcal{A}} \int_0^{+\infty} |g(\varphi(t,x,a),a(t))-g(\varphi(t,y,a),a(t))|\,dt \\ &\le \sup_{a\in\mathcal{A}} \int_0^{T} |g(\varphi(t,x,a),a(t))-g(\varphi(t,y,a),a(t))|\,dt\end{aligned}$$

$$
\begin{aligned}
&+\sup_{a\in\mathcal{A}}\int_T^{+\infty}|g(\varphi(t,x,a),a(t))-g(\varphi(t,y,a),a(t))|\,dt\\
&\le\int_0^T L_R e^{M_R t}\|x-y\|dt\\
&\quad+\int_T^{+\infty}K\max\{\|\varphi(T)\|,\,\|y(T)\|\}^s C^s e^{-s\sigma(t-T)}e^{M_r t}\|x-y\|dt\\
&\le\underbrace{\left(L_R\frac{e^{M_R T}-1}{M_R}+Kr^sC^se^{\sigma T}\frac{e^{(M_r-s\sigma)T}}{s\sigma-M_r}\right)}_{=L_S}\|x-y\|\,.
\end{aligned}
$$

This shows the assertion. □

Obviously, this result immediately carries over to v on $\mathcal{D}_0$. In order to obtain Lipschitz continuity of v on the rest of $\mathbb{R}^n$, it is convenient to consider a generalization of the transformation (3.2) by defining

$$v_\delta(x):=1-\exp(-\delta V(x))$$

for $\delta>0$. Observe that this results in the equation

$$\inf_{a\in A}\{-f(x,a)Dv(x)-\delta(1-v(x))g(x,a)\}=0,\qquad x\in\mathbb{R}^n. \tag{4.1}$$

Thus this transformation is equivalent to an appropriate choice of g in (3.10). Observe that for $\delta\to 0$ the function v_δ converges to 0 on $\mathcal{D}_0$ and is equal to 1 outside $\mathcal{D}_0$. Note that this convergence to a piecewise constant function is a typical behavior of discounted optimal value functions, see, e.g., [10].

In the opposite case, i.e., for sufficiently large $\delta>0$, the following result holds for v_δ.

PROPOSITION 4.3. *Assume that $f(\cdot,a)$ and $g(\cdot,a)$ are globally Lipschitz continuous in $\mathbb{R}^n$, with constants L_f, $L_g>0$ independent of $a\in A$, and assume that there exists a neighborhood N of the origin such that for all x, $y\in N$ the inequality*

$$|g(x,a)-g(y,a)|\le K\max\{\|x\|,\,\|y\|\}^s\|x-y\|$$

holds for some $K>0$ and $s>L_f/\sigma$ with $\sigma>0$ given by (H1). *Then the function v_δ is Lipschitz continuous in $\mathbb{R}^n$ for all $\delta>0$ sufficiently large.*

Proof. Let L_0 denote the Lipschitz constant of V on $B(0,r)$ guaranteed by Proposition 4.2. For $x\in\mathcal{D}_0$, define $T_x=\sup\{t(x,a):a\in\mathcal{A}\}$ and observe that $V(x)\ge g_0T_x$, where $g_0>0$ is given by (H2). If x, $y\in\mathcal{D}_0$, then for any $\epsilon>0$, there exists a control $a\in\mathcal{A}$ such that

$$
\begin{aligned}
|V(x)-V(y)|&\le\int_0^{T_x\vee T_y}|g(\varphi(t,x,a),a(t))-g(\varphi(t,y,a),a(t))|dt\\
&\quad+|V(\varphi(T_x\vee T_y,x,a))-V(\varphi(T_x\vee T_y,y,a))|+\epsilon\\
&\le\int_0^{T_x\vee T_y}L_g\exp(L_ft)\|x-y\|dt+L_0\|x-y\|\exp(L_f(T_x\vee T_y))+\epsilon\\
&\le(L_0+L_g/L_f)\exp(L_fV(x)/g_0)\|x-y\|+\epsilon.
\end{aligned}
$$

So we see that V is locally Lipschitz continuous in $\mathcal{D}_0$ with a constant of the form $L\exp(L_fV(x)/g_0)$.

Let $\phi \in C^1(\mathbb{R}^n)$ be such that $v_\delta(x) - \phi$ has a local maximum at $x_0 \in \mathcal{D}_0$, where we may assume that $v_\delta(x_0) - \phi(x_0) = 0$ and $\phi(x) \le 1, \forall x \in \mathbb{R}^n$. Then $V - \psi$ has a local maximum at x_0 for $\psi(x) = -\ln(1 - \phi(x))/\delta$.

It follows that

$$|D\phi(x_0)| \le \delta |D\psi(x_0)| \exp(-\delta V(x)) \le L\delta \exp((L_f/g_0 - \delta)V(x)).$$

Hence, letting $\delta \ge L_f/g_0$ and recalling that $v_\delta \equiv 1$ in $\mathbb{R}^n \setminus \mathcal{D}_0$, we have that $|D\phi(x_0)| \le \delta L$ for any $x \in \mathbb{R}^n$ and for any $\phi \in C^1(\mathbb{R}^n)$ such that $v_\delta(x) - \phi$ has a local maximum at x. This implies that v_δ is Lipschitz continuous in $\mathbb{R}^n$ with Lipschitz constant δL, cf. [5, Lemma 2.10]. □

5. Smooth solutions. It is always of interest to know whether for a given stability property there are Lyapunov functions with certain regularity properties. In [16] it is shown that under the condition of global uniform asymptotic stability, that is, under the condition $\mathcal{D}_0 = \mathbb{R}^n$ in our terminology, there exists a C^∞ Lyapunov function $V : \mathbb{R}^n \to \mathbb{R}$ such that

$$DV(x)f(x,a) \le -\alpha_1(\|x\|) \tag{5.1}$$

for some class $\mathcal{K}_\infty$ function α_1. Furthermore, there exist class $\mathcal{K}_\infty$ functions α_2, α_3 such that

$$\alpha_2(\|x\|) \le V(x) \le \alpha_3(\|x\|). \tag{5.2}$$

(As usual in stability theory, we call a function $\alpha : [0,\infty) \to [0,\infty)$ of class $\mathcal{K}_\infty$ if it is continuous, strictly increasing, unbounded, and satisfies $\alpha(0) = 0$). By [23, Theorems 1 and 2, Proposition 3] it follows that if we add the assumption that $f(x, A)$ be convex for all $x \in \mathbb{R}^n$, then there exists a C^∞ Lyapunov function V on $\mathcal{D}_0$ (which is in this case equal to $\mathcal{D}$ by Proposition 2.3 (iii)). Assuming that $\omega : \mathcal{D} \to \mathbb{R}_{\ge 0}$ is continuous and satisfies $\omega(x) = 0$ if and only if $x = 0$, and that $\omega(x_n) \to \infty$ for any sequence $\{x_n\}$ with $\lim x_n \in \partial\mathcal{D}$ or $\lim \|x_n\| = \infty$, then V can be chosen in such a manner that it has the properties (5.1), (5.2), where $\|x\|$ has to be replaced by $\omega(x)$. It is of interest, and therefore the topic of our last section, to know whether we are able to reproduce these functions via our approach.

We first treat the case of global stability.

LEMMA 5.1. *Assume that system* (2.1) *is globally uniformly asymptotically stable at* 0*; then* $g(x,a)$ *can be chosen such that the corresponding solutions* V *of* (3.9) *and* v *of* (3.10) *are* C^∞*. Furthermore, for any smooth Lyapunov function* V *satisfying* (5.1) *and* (5.2) *there exists a function* $g : \mathbb{R}^n \to \mathbb{R}$ *such that* V *is the corresponding solution of* (3.9)*.*

Proof. By [16, Theorem 2.9, Remark 4.1] there exists a C^∞ Lyapunov function $V : \mathbb{R}^n \to \mathbb{R}$ for (2.1). Now define $v(x) = 1 - e^{-V(x)}$ as before, and

$$\begin{aligned} g(x,a) := g(x) &:= -\sup_{a\in\mathcal{A}} \frac{Dv(x)f(x,a)}{1 - v(x)} \\ &= -\sup_{a\in\mathcal{A}} \frac{e^{-V(x)}DV(x)f(x,a)}{e^{-V(x)}} = -\sup_{a\in A} DV(x)f(x,a)\,. \end{aligned} \tag{5.3}$$

It is clear that the function g thus defined satisfies condition (i) of (H2). By (5.1) we have $g(x) \ge \alpha_1(\|x\|)$, which implies (ii). The third condition is implied by the

Lipschitz continuity of f and smoothness of V. A straightforward computation yields that V and v are the respective (unique) solutions of (3.9) and (3.10).

The second statement is clear by the previous construction. □

It is now tempting to try to copy this argument for the nonglobal case by utilizing the smooth maximal Lyapunov functions defined on the domain of attraction which are obtained in [23]. In this way one might hope to construct smooth Lyapunov functions that are representable as suitable solutions of (3.9), respectively, (3.10). This approach, however, has one problem. It is by no means clear that g as defined in the proof of Lemma 5.1 can be continuously extended to $\mathbb{R}^n$ so that (H2) is satisfied. We can, however, reconstruct smooth solutions on any subset of $\mathcal{D}_0$ that is bounded away from $\partial\mathcal{D}_0$.

PROPOSITION 5.2. *Assume* (H1), (H2)*, and that $f(x, A)$ is convex for all $x \in \mathbb{R}^n$. Let $B \subset \mathcal{D}_0$ satisfy* $\operatorname{dist}(B, \partial\mathcal{D}_0) > 0$*; then there exists a function $g : \mathbb{R}^n \to \mathbb{R}$ such that the corresponding solution v of* (3.10) *is C^∞ on a neighborhood of B.*

Proof. Let V denote a smooth Lyapunov function for system (2.1) defined on $\mathcal{D}$. Let U be an open neighborhood of B contained in $\mathcal{D}_0$, and define $g_{|U}$ by (5.3). Then g can be extended to a continuous function on $\mathbb{R}^n$ satisfying (H2). The corresponding unique solution v of (3.10) is C^∞ on U. □

6. Example. In this section we illustrate our results by a simple example, where we explicitly verify a (nonsmooth) solution of the generalized version of Zubov's equation (3.10). Consider the system

$$\begin{aligned} \dot{x}_1 &= -x_1 + ax_1^2, \\ \dot{x}_2 &= -x_2 + ax_2^2 \end{aligned}$$

with $x = (x_1, x_2)^T \in \mathbb{R}^2$ and $A = [-1, 1]$. We claim that for $g(x, a) = \|x\|^2 = x_1^2 + x_2^2$ the function v defined by

$$v(x) = \begin{cases} 1 - e^{-V(x)}, & x \in (-1, 1)^2, \\ 1, & x \notin (-1, 1)^2, \end{cases}$$

where $V : (-1, 1)^2 \to \mathbb{R}$ is given by

$$V(x) = \begin{cases} -\ln(1 - x_1) - \ln(1 - x_2) - x_1 - x_2, & x_1 \geq -x_2, \\ -\ln(1 + x_1) - \ln(1 + x_2) + x_1 + x_2, & x_1 \leq -x_2, \end{cases}$$

solves (3.10).

Note that by Theorem 3.12 it suffices to verify (3.10) on $(-1, 1)^2$, since $v|_{(-1,1)^2}$ satisfies the assumptions of this theorem with $\mathcal{O} = (-1, 1)^2$.

Using Remark 3.1, we identify the set of possible derivatives of functions ϕ such that $v - \phi$ has a local extremum for $x \in (-1, 1)$. First note that v is smooth on $(-1, 1)^2 \setminus D_1$, where D_1 is the diagonal $\{x \in (-1, 1)^2 \mid x_1 = -x_2\}$. In this region $D\phi$ must coincide with Dv, which is computed to be

$$Dv(x) = \begin{cases} (x_1(1 - x_2)e^{+x_1+x_2},\ x_2(1 - x_1)e^{+x_1+x_2}), & x \in (-1, 1)^2,\ x_1 > -x_2, \\ (x_1(1 + x_2)e^{-x_1-x_2},\ x_2(1 + x_1)e^{-x_1-x_2}), & x \in (-1, 1)^2,\ x_1 < -x_2. \end{cases}$$

On D_1 (setting $x = (y, -y)^T$) one verifies that the superdifferential D^+v is empty, while the subdifferential D^-v satisfies

$$D^-v(y, -y) = \{\theta p_1 + (1 - \theta)p_2 \mid \theta \in [0, 1]\},$$

where

$$p_1 = (+y(y+1),\, +y(y-1))\,,$$
$$p_2 = (-y(y-1),\, -y(y+1))\,.$$

Using these computations, we obtain that on $(-1,1)^2$ (3.10) becomes

$$\min_{a\in[-1,1]} \{-e^{x_1+x_2}(1-a)(x_1^3+x_2^3-x_1x_2^3-x_1^3x_2)\} = 0 \quad \text{for } x_1 > -x_2,$$

$$\min_{a\in[-1,1]} \{e^{-x_1-x_2}(1+a)(x_1^3+x_2^3+x_1^3x_2+x_1x_2^3)\} = 0 \quad \text{for } x_1 < -x_2,$$

and

$$\min_{a\in[-1,1]} \{2(1-a+2\theta a)y^4)\} \geq 0 \quad \text{for } x_1 = -x_2 =: y.$$

It turns out that in the first case the minimizer is $a = 1$, and in the second case it is $a = -1$, while in the third case it is $a = 1$ for $\theta \in [0, 1/2)$, $a = -1$ for $\theta \in (1/2, 1]$, and any $a \in [-1, 1]$ for $\theta = 1/2$. In all cases we see that the desired (in)equalities are satisfied, which, in particular, shows that $\mathcal{D}_0 = (-1, 1)^2$.

Acknowledgments. The authors would like to thank Pierpaolo Soravia and Maurizio Falcone for useful discussions.

REFERENCES

[1] M. Abu Hassan and C. Storey, *Numerical determination of domains of attraction for electrical power systems using the method of Zubov*, Internat. J. Control, 34 (1981), pp. 371–381.
[2] B. Aulbach, *Asymptotic stability regions via extensions of Zubov's method.* I *and* II, Nonlinear Anal., 7 (1983), pp. 1431–1440, 1441–1454.
[3] M. Bardi and I. Capuzzo Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman equations*, Birkhäuser Boston, Boston, 1997.
[4] M. Bardi and P. Soravia, *Hamilton-Jacobi equations with singular boundary conditions on a free boundary and applications to differential games*, Trans. Amer. Math. Soc., 325 (1991), pp. 205–229.
[5] G. Barles, *Solutions de viscosité des equations de Hamilton-Jacobi*, Springer-Verlag, Paris, 1994.
[6] N. Bhatia, *On asymptotic stability in dynamical systems*, Math. Systems Theory, 1 (1967), pp. 113–127.
[7] F. Camilli, L. Grüne, and F. Wirth, *A regularization of Zubov's equation for robust domains of attraction*, in Nonlinear Control in the Year 2000, Volume 1, Lecture Notes in Control and Inform. Sci. 258, A. Isidori et al., eds., Springer-Verlag, London, 2000, pp. 277–290.
[8] C. Coleman, *Local trajectory equivalence of differential systems*, Proc. Amer. Math. Soc., 16 (1965), pp. 890–892.
[9] C. Coleman, *Addendum to: Local trajectory equivalence of differential systems*, Proc. Amer. Math. Soc., 17 (1966), p. 770.
[10] L. Grüne, *On the relation between discounted and average optimal control problems*, J. Differential Equations, 148 (1998), pp. 65–99.
[11] W. Hahn, *Stability of Motion*, Springer-Verlag, Berlin, 1967.
[12] V. Jurdjevic, *Geometric Control Theory*, Cambridge University Press, Cambridge, UK, 1997.
[13] N. E. Kirin, R. A. Nelepin, and V. N. Bajdaev, *Construction of the attraction region by Zubov's method*, Differential Equations, 17 (1982), pp. 871–880.
[14] H. K. Khalil, *Nonlinear Systems*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1996.
[15] E. B. Lee and L. Markus, *Foundations of Optimal Control Theory*, John Wiley and Sons, New York, 1967.
[16] Y. Lin, E. D. Sontag, and Y. Wang, *A smooth converse Lyapunov theorem for robust stability*, SIAM J. Control Optim., 34 (1996), pp. 124–160.

[17] J. W. Milnor, *Differential topology*, in Lectures in Modern Mathematics II, John Wiley and Sons, New York, 1964, pp. 165–183.

[18] A. D. B. Paice and F. Wirth, *Robustness analysis of domains of attraction of nonlinear systems*, in Proceedings of the Mathematical Theory of Networks and Systems, Il Poligrafo, Padova, Italy, 1998, pp. 353–356.

[19] A. D. B. Paice and F. Wirth, *Robustness of nonlinear systems subject to time-varying perturbations*, in Advances in Mathematical Systems Theory, F. Colonius et al., eds., Birkhäuser Boston, Boston, 2000, pp. 31–54.

[20] E. D. Sontag and Y. Wang, *New characterizations of input to state stability*, IEEE Trans. Automat. Control, 41 (1996), pp. 1283–1294.

[21] P. Soravia, *Optimality principles and representation formulas for viscosity solutions of Hamilton-Jacobi equations* I: *Equations of unbounded and degenerate control problems without uniqueness*, Adv. Differential Equations, 4 (1999), pp. 275–296.

[22] P. Soravia, *Optimality principles and representation formulas for viscosity solutions of Hamilton-Jacobi equations* II: *Equations of control problems with state constraints*, Differential Integral Equations, 12 (1999), pp. 275–293.

[23] A. Teel and L. Praly, *A smooth Lyapunov function from a class-$\mathcal{KL}$ estimate involving two positive semidefinite functions*, ESAIM Control Optim. Cal. Var., 5 (2000), pp. 313–367.

[24] A. Vannelli and M. Vidyasagar, *Maximal Lyapunov functions and domains of attraction for autonomous nonlinear systems*, Automatica J. IFAC, 21 (1985), pp. 69–80.

[25] F. W. Wilson, *The structure of the level surfaces of a Lyapunov function*, J. Differential Equations, 3 (1967), pp. 323–329.

[26] V. I. Zubov, *Methods of A. M. Lyapunov and Their Application*, P. Noordhoff, Groningen, The Netherlands, 1964.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 2, pp. 516–524

VARIATIONAL INEQUALITIES, SYSTEM OF FUNCTIONAL EQUATIONS, AND INCOMPLETE INFORMATION REPEATED GAMES*

RIDA LARAKI†

Abstract. We consider a pair of functional equations obtained by Mertens and Zamir (*Internat. J. Game Theory,* 1 (1971–72), pp. 39–64; *J. Math. Anal. Appl.* 60 (1977), pp. 550–558) to characterize the asymptotic value of a two person zero sum repeated with lack of information on both sides (Aumann and Maschler (*Repeated Games with Incomplete Information,* MIT Press, Cambridge, MA 1995)). We give a new proof for the convergence of the discounted values of the repeated game and a new characterization of the limit using variational inequalities. The same idea allows us to prove existence and uniqueness of a Lipschitz solution for the pair of functional equations in a general framework using an auxiliary game: "the splitting game," introduced by Sorin (*A First Course on Zero-Sum Repeated Games,* preprint).

Key words. incomplete information repeated games, variational inequalities, functional equations, convexification operator, convex-concave functions

AMS subject classifications. 39B72, 49K40, 52B99, 91A05, 91A15, 91A20

PII. S0363012900366601

Introduction. A two person zero sum repeated game with lack of information on both sides (Aumann and Maschler (1995)) is a multistage game where the payoff function depends on two parameters and where each player knows only one of the parameters (see section 1).

Mertens and Zamir (1971–1972) have shown that the sequence of the values converges when the length of the game grows to infinity, to the unique solution of the following system of functional equations with unknown v:

$$\begin{cases} v(p,q) = Cav_{p\in\Delta(K)}\left[\min(u,v)\right](p,q) & (S1)\,, \\ v(p,q) = Vex_{q\in\Delta(L)}\left[\max(u,v)\right](p,q) & (S2)\,. \end{cases} \qquad (S)$$

In this system, K and L are two finite sets, $\Delta(K)$ is the unit simplex of R^K, $\Delta(L)$ is the unit simplex of R^L, and u is the value of the average game (that is, the game where the players do not use their information). For a function φ and a convex set C, $Cav_C[\varphi]$ (resp., $Vex_C[\varphi]$) is the smallest concave function on C greater than φ (resp., the greatest convex function on C smaller than φ).

In fact, Mertens and Zamir (1977) also studied the "functional equations" (S) in a general framework without reference to game theoretical tools: u is not necessarily the value of a game, and $\Delta(K)$ and $\Delta(L)$ are replaced by any convex-compact sets C and D in finite dimension. Remark that when u does not depend on the first variable, the unique solution of (S) is $Vex_D[u]$. The example of Kruskal (1969) shows that the convexification operator does not conserve the continuity for any convex compact set D. (That is, $Vex_D[u]$ is not always continuous even if u is continuous.) This implies that the Mertens–Zamir system does not always admit a continuous solution

*Received by the editors January 24, 2000; accepted for publication (in revised form) March 7, 2001; published electronically July 25, 2001.

http://www.siam.org/journals/sicon/40-2/36660.html

†Ecole Polytechnique, Laboratoire d'Econométrie, 1 rue Descartes, 75005 Paris and Modal'X, UFR-SEGMI, Université Paris, 10 Nanterre, 200, Avenue de la République, 92001 Nanterre Cédex, France (laraki@poly.polytechnique.fr).

for arbitrary convex compact sets C and D. In a recent work (Laraki (2001a)), we studied necessary and sufficient conditions on the geometry of a convex-compact set X in order that the convexification operator on X conserves the continuity (resp., uniformly the Lipschitz property). In Laraki (2001b), we studied the existence of a continuous solution for (S) when C and D are in the class of convex-compact sets characterized in Laraki (2001b). This is the class of convex-compact sets X such that the convexification operator on X conserves the continuity.

In this paper we will consider the existence of a Lipschitz solution for (S) in a general framework. In Laraki (2001a) we proved that when X is a polytope in a normed real vector space, then the convexification operator conserves uniformly the Lipschitz property. We showed that in finite dimension, being a polytope is also a necessary condition to preserve uniformly the Lipschitz property. Hence it is natural to ask if (S) admits a Lipschitz solution when C and D are two polytopes in a normed real vector space and u is Lipschitz.

The main contributions of this paper are

- a new characterization ($P1$ and $P2$, below) equivalent to (S) in a very general framework;
- a new simple proof for the convergence of the discounted values of the repeated game with incomplete information on both sides by providing a game interpretation of $P1$ and $P2$ (section 2);
- the same idea allows us to show that (S) admits a Lipschitz solution (where C and D are two polytopes in a normed vector space and u is Lipschitz) by using an auxiliary (stochastic) game called the "splitting game" (introduced by Sorin (2000)) (see sections 3–4).

A function $f(p,q)$ is concave-convex if it is concave in p and convex in q.

More precisely, we prove that the limit of the discounted values is the unique continuous concave-convex function, $v(p,q)$, satisfying the following variational inequalities.

- $P1$: For all q_0, if $[p_0, v(p_0,q_0)]$ is an extreme point of (the hypograph) of $v(\cdot,q_0)$, then $v(p_0,q_0) \leq u(p_0,q_0)$.
- $P2$: For all q_0, if $[q_0, v(p_0,q_0)]$ is an extreme point of (the epigraph) of $v(q_0,\cdot)$, then $v(p_0,q_0) \geq u(p_0,q_0)$.

In section 2.1 we will show that any accumulation point, v, of the family of the discounted values $\{v_\lambda\}$ satisfies $P1$ and $P2$. This very simple proof translates the following intuitive idea: if p_0 is an extreme point of $v(\cdot,q_0)$, then " asymptotically" Player 1 must not use his information, and then Player 2 can guarantee asymptotically (just by not using his information) $u(p_0,q_0)$. This implies that $v(p_0,q_0) \leq u(p_0,q_0)$.

In section 2.2 we deduce the uniqueness of a continuous solution by proving a comparison theorem.

In section 3 we study the existence of a Lipschitz solution by using the splitting game.

Finally, we prove the equivalence with Mertens–Zamir's system in a very general framework (section 4).

1. Preliminary results. We recall here the framework of zero sum repeated games with incomplete information (Aumann and Maschler (1995)).

I and J are two finite sets, $X = \Delta(I)$ (the set of probabilities on I), and $Y = \Delta(J)$. $\left(A^{k,l}\right)_{k\in K, l\in L}$ is a family of $I \times J$-matrices (I rows and J columns).

DEFINITION 1. *For each $p \in \Delta(K)$ and $q \in \Delta(L)$, the game form $G_F(p,q)$ is as follows.*

- *At stage 0, k is chosen according to the probability p and announced to Player 1 only; l is chosen according to the probability q and announced to Player 2 only.*
- *At stage 1, Player 1 chooses a move $i_1 \in I$, Player 2 chooses a move $j_1 \in J$, and the couple (i_1, j_1) is told to both players. The payoff is $A^{k,l}_{i_1,j_1}$ but is not announced.*
- *Inductively, at stage m, knowing the past history $h_m = (i_1, j_1, \ldots, i_{m-1}, j_{m-1})$, Player 1 chooses a move $i_m \in I$, Player 2 chooses a move $j_m \in J$, and the new history $h_{m+1} = (h_m, i_m, j_m)$ is told to both players. The payoff is $A^{k,l}_{i_m,j_m}$ and is not announced.*
- *Both players know the above description (public knowledge).*

We denote by Σ (resp., Υ) the set of behavioral strategies of Player 1 (resp., Player 2).

Several games are associated to this game form and differ only in the way the stream of payoffs is evaluated. We will be interested in the λ-discounted game $G_\lambda(p,q)$ $(0 < \lambda < 1)$, where, if the play is $(k, l, (i_1, j_1), \ldots, (i_n, j_n), \ldots)$, Player 2 gives Player 1 the amount $\sum_{m=1}^{\infty} \lambda(1-\lambda)^{m-1} A^{k,l}_{i_m j_m}$. The stage payoff being uniformly bounded, the payoff function is jointly continuous and bilinear on $\Sigma \times \Upsilon$. Hence (Sion (1958)) this game has a value $v_\lambda(p,q)$.

Notations.

- Let $u(p,q)$ be the value of the one shot average game $G(p,q)$ with matrix payoff $A_{i,j}(p,q) = \sum_{k \in K} p^k q^l A^{k,l}_{ij}$. Then u is a Lipschitz function on $\Delta(K) \times \Delta(L)$ with constant $\|A\|_\infty = \max_{k,l,i,j} \left\| A^{k,l}_{i,j} \right\|$.
- Let us call a function $f(p,q)$ concave in p and convex in q a saddle function.
- Let F be the set of saddle Lipschitz functions on $\Delta(K) \times \Delta(L)$ with constant $\|A\|_\infty$.
- Fix $(p,q) \in \Delta(K) \times \Delta(L)$ initial probabilities and $(x,y) \in X^K \times Y^L$ one stage strategies of the players. Then, for all $(i,j) \in I \times J$, we define:
  $\overline{x}(i) = \sum_{k \in K} p^k x^k(i)$ (the total probability of playing i),
  $\overline{y}(j) = \sum_{l \in L} q^l y^l(j)$ (the total probability of playing j),
  $p(i)$: the conditional probability over K knowing i given by $p^k(i) = \frac{p^k x^k(i)}{\overline{x}(i)}$,
  $q(j)$: the conditional probability over L knowing j given by $q^l(j) = \frac{p^l y^l(j)}{\overline{y}(j)}$,
  $A_{x,y}(p,q) = \sum_{k,l,i,j} p^k q^l x^k(i) y^l(j) A^{k,l}_{i,j}$.

Then we have the following property (see Mertens, Sorin and Zamir (1994)).

PROPOSITION 1. *v_λ is in F and satisfies the following recursive formula:*

$$v_\lambda(p,q) = \max_{x \in X^K} \min_{y \in Y^l} \left[\lambda A_{x,y}(p,q) + (1-\lambda) \sum_{i \in I, j \in J} \overline{x}(i)\overline{y}(j) v_\lambda\left(p(i), q(j)\right) \right].$$

2. The convergence of v_λ. We prove in this section that the asymptotic value, $v = \lim_{\lambda \to 0} v_\lambda$, exists and is the unique saddle continuous function on $\Delta(K) \times \Delta(L)$ satisfying $P1$ and $P2$. In the first subsection we give a game theoretical interpretation of these properties by proving that any accumulation point of v_λ satisfies $P1$ and $P2$. In the second subsection we prove uniqueness via a comparison theorem.

2.1. Existence. We want to express mathematically the fact that if p_0 is an extreme point of $v(\cdot, q_0)$, then "asymptotically" Player 1 must not use his information. We will follow the operator approach of Rosenberg and Sorin (2001) to study repeated games with incomplete information. (We use here some of their notations.)

Define for $0 \leq \lambda \leq 1$ an operator $T(\lambda, \cdot)$ which associates to a function $f \in F$ the function $T(\lambda, f)$ defined by

$$T(\lambda, f)(p,q) = \max_{x \in X^K} \min_{y \in Y^L} \left[\lambda A_{x,y}(p,q) + (1-\lambda) \sum_{i,j} \overline{x}(i)\overline{y}(j) f\left(p(i), q(j)\right) \right]$$

$$= \min_{y \in Y^L} \max_{x \in X^K} \left[\lambda A_{x,y}(p,q) + (1-\lambda) \sum_{i,j} \overline{x}(i)\overline{y}(j) f\left(p(i), q(j)\right) \right].$$

Denote by $X_\lambda(f)(p,q)$ (resp., $Y_\lambda(f)(p,q)$) the set of optimal strategies of Player 1 (resp., Player 2) in the above one shot game.

We introduce NR_p^1, the set of nonrevealing strategies of Player 1, i.e., the set of $x \in X^K$ such that the conditional probability distribution on K induced by x is constant (thus equal to p), which is the case if and only if, for all $k \neq k'$ such that $p(k)p(k') > 0$, $x^k = x^{k'}$. NR_q^2 is defined in the same way.

For a function g defined on $\Delta(K)$, p is an extreme point of g if $g(p) = \alpha g(p_1) + (1-\alpha)\, g(p_2)$ with $p = \alpha p_1 + (1-\alpha)\, p_2$ and $0 < \alpha < 1$ implies $p_1 = p_2 = p$.

Because the family $\{v_\lambda\}$ is uniformly Lipschitz, there exist $v \in F$ and $(\lambda_n) \to 0$ such that v_{λ_n} converges uniformly to v. Such a v is called an accumulation point of the family $\{v_\lambda\}$.

We have the following properties.

PROPOSITION 2 (see Rosenberg and Sorin (2001)).

(i) $v_\lambda = T(\lambda, v_\lambda)$.

For all v, an accumulation point of $\{v_\lambda\}$, we have the following.

(ii) $v = T(0, v)$.

(iii) *If p_0 is an extreme point of $v(\cdot, q_0)$, then $X_0(v)(p_0, q_0) \subset NR_{p_0}^1$.*

(iv) *If q_0 is an extreme point of $v(p_0, \cdot)$, then $Y_0(v)(p_0, q_0) \subset NR_{q_0}^2$.*

Proof. (i) and (ii) are consequences of Proposition 1, the definition and the continuity of T.

For (iii), let $x^* \in X_0(v)(p_0, q_0)$. Then we have

$$v(p_0, q_0) = \min_{y \in Y^L} \left[\sum_{i,j} \overline{x}^*(i)\overline{y}(j) v\left(p_0(i), q_0(j)\right) \right]$$

$$\leq \min_{y \in Y} \left[\sum_{i} \overline{x}^*(i) v\left(p_0(i), q_0\right) \right]$$

$$\leq \sum_{i} \overline{x}^*(i) v\left(p_0(i), q_0\right).$$

But, as $v\left(\cdot, q_0\right)$ is concave, we have

$$v(p_0, q_0) = \sum_{i} \overline{x}^*(i) v\left(p_0(i), q_0\right).$$

Since p_0 is an extreme point of $v\left(\cdot, q_0\right)$, we deduce that $p_0(i) = p_0$ for all i. Thus $x^* \in NR_{p_0}^1$. ∎

Let us recall the basic variational inequalities $P1$ and $P2$ for a function v.

- $P1$: For all $q_0 \in \Delta(L)$, if $p_0 \in \Delta(K)$ is an extreme point of $v(\cdot, q_0)$, then $v(p_0, q_0) \leq u(p_0, q_0)$.
- $P2$: For all $p_0 \in \Delta(K)$, if $q_0 \in \Delta(L)$ is an extreme point of $v(q_0, \cdot)$, then $v(p_0, q_0) \geq u(p_0, q_0)$.

PROPOSITION 3. *Any accumulation point v of $\{v_\lambda\}$ satisfies $P1$ and $P2$.*

Proof. Let p_0 be an extreme point of $v(\cdot, q_0)$.

Denote $E_{x,y}[f](p_0, q_0) = \sum_{i\in I, j\in J} \overline{x}(i)\overline{y}(j) f(p_0(i), q_0(j))$. Then, we have

$$\begin{aligned}
&v_{\lambda_n}(p_0, q_0)\\
&= \max_{x\in X^K} \min_{y\in Y^l} \left[\lambda_n \left[A_{x,y}(p_0, q_0) - E_{x,y}\left[v_{\lambda_n}\right](p_0, q_0)\right] + E_{x,y}\left[v_{\lambda_n}\right](p_0, q_0)\right]\\
&= \max_{x\in X_{\lambda_n}(v_{\lambda_n})(p_0,q_0)} \min_{y\in Y^l} \left[\lambda_n \left[A_{x,y}(p_0, q_0) - E_{x,y}\left[v_{\lambda_n}\right](p_0, q_0)\right] + E_{x,y}\left[v_{\lambda_n}\right](p_0, q_0)\right]\\
&\leq \max_{x\in X_{\lambda_n}(v_{\lambda_n})(p_0,q_0)} \min_{y\in Y} \left[\begin{array}{c} \lambda_n \left[A_{x,y}(p_0, q_0) - \sum_{i\in I} \overline{x}(i) v_{\lambda_n}(p_0(i), q_0)\right] \\ + \sum_{i\in I} \overline{x}(i) v_{\lambda_n}(p_0(i), q_0) \end{array}\right].
\end{aligned}$$

Thus

$$\max_{x\in X_{\lambda_n}(v_{\lambda_n})(p_0,q_0)} \min_{y\in Y} \left(\lambda_n \left[A_{x,y}(p_0, q_0) - \sum_{i\in I} \overline{x}(i) v_{\lambda_n}(p_0(i), q_0)\right] + \sum_{i\in I} \overline{x}(i) v_{\lambda_n}(p_0(i), q_0) - v_{\lambda_n}(p_0, q_0)\right) \geq 0.$$

But v_{λ_n} concave yields

$$\sum_{i\in I} \overline{x}(i) v_{\lambda_n}(p_0(i), q_0) - v_{\lambda_n}(p_0, q_0) \leq 0,$$

so that

$$\max_{x\in X_{\lambda_n}(v_{\lambda_n})(p_0,q_0)} \min_{y\in Y} \lambda_n \left[A_{x,y}(p_0, q_0) - \sum_{i\in I} \overline{x}(i) v_{\lambda_n}(p_0(i), q_0)\right] \geq 0.$$

Since $\lambda_n > 0$, this gives

$$\max_{x\in X_{\lambda_n}(v_{\lambda_n})(p_0,q_0)} \min_{y\in Y} \left[A_{x,y}(p_0, q_0) - \sum_{i\in I} \overline{x}(i) v_{\lambda_n}(p_0(i), q_0)\right] \geq 0.$$

Now let $X^*(v)(p_0, q_0)$ be the set of accumulation points of $X_{\lambda_n}(v_{\lambda_n})(p_0, q_0)$. Since the correspondence $(\lambda, f) \to X_\lambda(f)$ is upper-semicontinuous, we deduce that

$$X^*(v)(p_0, q_0) \subset X_0(v)(p_0, q_0).$$

But p_0 is an extreme point of $v(\cdot, q_0)$; thus by (iii) in Proposition 2 we deduce that

$$X_0(v)(p_0, q_0) \subset NR^1_{p_0}.$$

Hence, letting $n \to \infty$ and using the uniform convergence of v_{λ_n} to v, we deduce that

$$\max_{x\in NR^1_{p_0}} \min_{y\in Y} \left[\sum_{k,l} p_0^k q_0^l x^k A^{k,l} y - v(p_0, q_0)\right] \geq 0.$$

Thus

$$\max_{x\in X}\min_{y\in Y}\left[\sum_{k,l}p_0^k q_0^l x A^{k,l} y - v\left(p_0,q_0\right)\right]\geq 0,$$

which implies that

$$u\left(p_0,q_0\right)\geq v\left(p_0,q_0\right).$$

Hence v satisfies $P1$ and, similarly, $P2$. □

2.2. Uniqueness. To show uniqueness, we use a comparison result and basically follow the idea of Mertens and Zamir (1971–72, 1977).

PROPOSITION 4. (maximum principle). *Let v_1 and v_2 be two saddle continuous functions satisfying $P1$ and $P2$, respectively. Then $v_1\leq v_2$.*

Proof. Let $\delta=\max_{p,q}\left[v_1(p,q)-v_2(p,q)\right]$. We will show that $\delta\leq 0$. Let $C=\operatorname{Argmax}_{p,q}\left[v_1(p,q)-v_2(p,q)\right]$. C is a nonempty compact set. □

We first prove the following.

LEMMA 2. *If (p_0,q_0) is an extreme point of the convex-hull of C, then p_0 is an extreme point of $v_1\left(\cdot,q_0\right)$ and q_0 is an extreme point of $v_2\left(p_0,\cdot\right)$.*

Proof. Assume that $v_1(p_0,q_0)=\alpha v_1(p_1,q_0)+\left(1-\alpha\right)v_1(p_2,q_0)$ with $p_0=\alpha p_1+\left(1-\alpha\right)p_2$ and $0<\alpha<1$.

Since v_2 is concave in p, we have

$$\alpha v_2(p_1,q_0)+\left(1-\alpha\right)v_2(p_2,q_0)\leq v_2(p_0,q_0),$$

so that

$$\begin{aligned}\alpha\left[v_1(p_1,q_0)-v_2(p_1,q_0)\right]+\left(1-\alpha\right)\left[v_1(p_2,q_0)-v_2(p_2,q_0)\right]\\ \geq v_1(p_0,q_0)-v_2(p_0,q_0)=\delta.\end{aligned}$$

Since $v_1(p,q)-v_2(p,q)\leq\delta$ for all (p,q), we necessarily have equality. Hence $(p_i,q_0)\in C$ for $i=1,2$, which is a contradiction.

Now we continue with the proof of the proposition.

Consider an extreme point (p_0,q_0) of the convex hull of C. By the previous lemma, $P1$, and $P2$, we deduce that $v_1(p_0,q_0)\leq u(p_0,q_0)$ and $v_2(p_0,q_0)\geq u(p_0,q_0)$.

Thus $\delta=v_1(p_0,q_0)-v_2(p_0,q_0)\leq u(p_0,q_0)-u(p_0,q_0)=0$. □

THEOREM 3. *v_λ converges uniformly to the unique continuous saddle function on $\Delta(K)\times\Delta(L)$ satisfying $P1$ and $P2$.*

Proof. The proof follows from the existence result (Proposition 3) and the comparison result (Proposition 4). □

3. The general case: Existence via the splitting game. Here H is a Lipschitz function on $C\times D$, where C and D are two polytopes in a normed real vector space.

We want to study the existence of a Lipschitz solution to the functional equations with unknown Ψ:

$$\begin{cases}\Psi(c,d)=Cav_{c\in C}\left[\min(H,\Psi)\right](c,d),\\ \Psi(c,d)=Vex_{d\in D}\left[\max(H,\Psi)\right](c,d).\end{cases}$$

In section 4 we will prove in a more general framework that this system is equivalent to the properties $P1[H,C,D]$ and $P2[H,C,D]$, below. Thus the comparaison theorem implies the uniqueness of a continuous solution. Here we use the same proof as

for a repeated game with incomplete information to prove the existence of a Lipschitz solution to the functional equations by considering an auxiliary stochastic game introduced by Sorin (2000) called the splitting game.

DEFINITION 4. *For each $(c_0, d_0) \in C \times D$, the splitting game $SG(c_0, d_0)$ is a zerosum stochastic game, described as follows.*

- *At stage 1 Player 1 chooses a probability P_{c_0} on C centered at c_0 and Player 2 chooses a probability Q_{d_0} on D centered at d_0. Then c_1 is selected according to P_{c_0}, and d_1 is selected according to Q_{d_0}. Finally, $h_1 = (c_0, d_0, c_1, d_1)$ is announced to both players. The stage payoff (from Player 2 to Player 1) is $H(c_1, d_1)$.*
- *Inductively, at stage $m+1$, knowing the past history h_m, Player 1 chooses a probability P_{c_m} on C centered at c_m, and Player 2 chooses a probability Q_{d_m} on D centered at d_m. Then c_{m+1} follows the low P_{c_m}, and d_{m+1} follows Q_{d_m}. Finally, $h_{m+1} = (h_m, c_{m+1}, d_{m+1})$ is announced to both players. The stage payoff is $H(c_{m+1}, d_{m+1})$.*

We consider the discounted evaluation $\sum_{m=1}^{\infty} \lambda(1-\lambda)^{m-1} H(c_m, d_m)$, where $0 < \lambda < 1$, and we call SG_λ the associated (discounted) splitting game.

PROPOSITION 5. *$SG_\lambda(c, d)$ has a value $V_\lambda(c, d)$.*

V_λ is a saddle function on $C \times D$ and satisfies the following recursive equation:

$$V_\lambda(c,d) = \max_{P \in \Delta_C(c)} \min_{Q \in \Delta_D(d)} \left[\int_{C \times D} \left[\lambda H(\widetilde{c}, \widetilde{d}) + (1-\lambda) V_\lambda(\widetilde{c}, \widetilde{d}) \right] dP(\widetilde{c}) dQ(\widetilde{d}) \right]$$
$$= \min_{Q \in \Delta_D(d)} \max_{P \in \Delta_C(c)} \left[\int_{C \times D} \left[\lambda H(\widetilde{c}, \widetilde{d}) + (1-\lambda) V_\lambda(\widetilde{c}, \widetilde{d}) \right] dP(\widetilde{c}) dQ(\widetilde{d}) \right],$$

where $\Delta_C(c)$ is the set of probabilities on C, centered at c, and $\Delta_D(d)$ is the set of probabilities on D, centered at d.

Moreover, there exists a norm on $C \times D$ with respect to which the family (V_λ) has the same Lipschitz constant as H.

Proof. Let $\mathcal{UL}$ be the space of real valued functions on $C \times D$ which are upper-semicontinuous–lower-semicontinuous, bounded by $\|H\|_\infty$. This space is complete for uniform convergence.

Let Φ be the splitting operator from $\mathcal{UL}$ to itself (Laraki (2001b)) defined by

$$\Phi[f](c,d) = \max_{P \in \Delta_C(c)} \min_{Q \in \Delta_D(d)} \left[\int_{C \times D} f(\widetilde{c}, \widetilde{d}) dP(\widetilde{c}) dQ(\widetilde{d}) \right]$$
$$= \min_{Q \in \Delta_D(d)} \max_{P \in \Delta_C(c)} \left[\int_{C \times D} f(\widetilde{c}, \widetilde{d}) dP(\widetilde{c}) dQ(\widetilde{d}) \right].$$

$f \to \Phi[\lambda H + (1-\lambda)\cdot]$ is contracting; hence it admits a fixed point $V_\lambda \in \mathcal{UL}$. It is standard and easy to show that both players can guarantee V_λ in the splitting game (see Mertens–Sorin–Zamir (1994)).

By Laraki (2001b) we deduce that V_λ is a saddle function.

Now, since C and D are polytopes, by Laraki (2001b) we deduce that there exists an equivalent norm on $C \times D$ ($\|\|\|_{C \times D}$) with respect to which the splitting operator conserves the Lipschitz constant. Hence, if M is the Lipschitz constant of H with respect to $\|\|\|_{C \times D}$, then the operator $f \to \Phi[\lambda H + (1-\lambda)\cdot]$ associates to an M-Lipschitz function an M-Lipschitz one. By the completeness of the space of uniformly Lipschitz functions, we deduce that the last operator admits a unique M-Lipschitz fixed point (which is V_λ, of course). ∎

DEFINITION 5. *A function φ on $C \times D$ satisfies*

- *$P1[H,C,D]$ if for all $d_0 \in D$, if $c_0 \in C$ is an extreme point of $\varphi(\cdot, d_0)$ in C, then $\varphi(p_0,q_0) \leq H(p_0,q_0)$;*
- *$P2[H,C,D]$ if for all $c_0 \in C$, if $d_0 \in D$ is an extreme point of $\varphi(c_0,\cdot)$ in D, then $\varphi(p_0,q_0) \geq H(p_0,q_0)$.*

PROPOSITION 6. *V_λ converges uniformly to the unique saddle continuous function V satisfying $P1[H,C,D]$ and $P2[H,C,D]$. In addition, for some equivalent norm depending (only) on C and D, V is Lipschitz with the same constant of Lipschitz as H.*

Proof. The proof here is exactly the same proof as in section 2. □

4. Equivalence with Mertens–Zamir's system. Here C and D are two convex-compact sets in a metric real vector space endowed with a locally convex topology.

LEMMA 6. *For all lower-semicontinuous bounded functions φ on D, there exists a unique lower-semicontinuous convex function (say, ψ) on D satisfying the following.*

(α) $\psi \leq \varphi$;

(β) If d_0 is an extreme point of ψ, then $\psi(d_0) \geq \varphi(d_0)$.

This function is $Vex_D[\varphi]$.

Proof. For a function f on D, $Epi(f)$ is the epigraph of f:

$$Epi(f) = \{(d,r) \in D \times BbbR : r \geq f(d)\}.$$

The property (β) and the fact that φ is bounded and lower-semicontinuous implies that $Epi(\psi) \subset Epi(\varphi)$. Since $Epi(\psi)$ is convex (since ψ is convex), we deduce that $Epi(\psi) \subset co[Epi(\varphi)] = Epi(Vex_D(\varphi))$. Hence $\psi \geq Vex_D[\varphi]$.

The property (α) and the fact that ψ is convex implies that $\psi \leq Vex_D[\varphi]$.

The fact that $(Vex_D[\varphi])$ satisfies (α) and (β) and is lower-semicontinuous is clear. □

PROPOSITION 7. *Let H and ψ be two upper-semicontinuous–lower-semicontinuous bounded functions on $C \times D$. Then the following hold.*

(i) *ψ is concave on C and satisfies $P1[H,C,D] \Leftrightarrow \psi = Cav_C[\min(H,\psi)]$.*

(ii) *ψ is convex on D and satisfies $P2[H,C,D] \Leftrightarrow \psi = Vex_D[\max(H,\psi)]$.*

Proof. Let us prove (ii) (the proof of (i) is similar).

It is clear that if $\psi = Vex_D[\max(H,\psi)]$, then the following hold.

- ψ is convex on D.
- If d_0 is an extreme point of $\psi(c_0,\cdot)$, then $\psi(c_0,d_0) = \max[H,\psi](c_0,d_0) \geq H(c_0,d_0)$.

Now suppose that ψ is convex on D and satisfies $P2[H,C,D]$. Let $c_0 \in C$ and put $\varphi(\cdot) = \max(H,\psi)(c_0,\cdot)$. Then it is clear that ψ satisfies (α) and (β) and is convex. By the last lemma we deduce that $\psi = Vex_D[\varphi]$. Hence $\psi = Vex_D[\max(H,\psi)]$. □

5. Concluding remarks.

- In fact, the last proposition can be deduced implicitly from the proof of Proposition 18 in Rosenberg and Sorin (2001). Our contribution is (a) how to extract properties $P1$ and $P2$ from the discounted games, and (b) their use to prove the existence of a Lipschitz solution of (S).
- The proof for the finitely repeated game (the study of $\lim v_n$) is much more complicated and needs all the machinery of the operator approach in Rosenberg and Sorin (2001). Since the goal in this paper is to give some new ideas with simple proofs, we omit this question.

- In fact, Mertens and Zamir (1971–1972) studied the asymptotic value in a more general framework where the private information received by the players is dependent. (The probability over the types is not necessarily the product of the marginals.) It is easy to see that our proof holds also in this case, but for clarity (because the formulation of the problem is very technical) we cover only the independent case.
- We remark that if the splitting operator does not conserve uniformly the Lipschitz property, then the familly (V_λ) will not be uniformly Lipschitz. Hence our proof does not apply directly in this case.
- In Laraki (2001b) we study the regularity properties of the splitting operator, and we address the problem of the existence of a continuous solution for the Mertens–Zamir system when C and D are in the class of convex-compact sets satisfying some necessary geometric conditions. (This class strictly contains the polytopes.)

Acknowledgment. My gratitude goes to Sylvain Sorin for supervising and motivating this work by his useful comments and advice.

REFERENCES

R. J. Aumann and M. Maschler with the collaboration of R. B. Stearns (1995), *Repeated Games with Incomplete Information*, MIT Press, Cambridge, MA.

J. B. Kruskal (1969), *Two convex counterexamples: A discontinuous envelope function and a nondifferentiable nearest-point matching*, Proc. Amer. Math. Soc., 23, pp. 697–703.

R. Laraki (2001a), *On the regularity of the convexification operator on a compact set*, Cahiers du Laboratoire d'Econométrie de l'Ecole Polytechnique, 2001–005, Paris, France.

R. Laraki (2001b), *The splitting game and applications*, Cahiers du Laboratoire d'Econométrie de l'Ecole Polytechnique, 2001–006, Paris, France. Internat. J. Game Theory, to appear.

J. F. Mertens, S. Sorin, and S. Zamir (1994), *Repeated Games*, Core Discussion Paper, 9420-9421-9422, Université Catholique de Louvain, Louvain la Neuve, Belgium.

J. F. Mertens and S. Zamir (1971–1972). *The value of two person zero sum repeated games with lack of information on both sides*, Internat. J. Game Theory, 1, pp. 39–64.

J. F. Mertens and S. Zamir (1977), *A duality theorem on a pair of simultaneous functional equations*, J. Math. Anal. Appl., 60, pp. 550–558.

D. Rosenberg and S. Sorin (2001), *An operator approach to zero-sum repeated games*, Israel J. Math., 121, pp. 221–246.

M. Sion (1958), *On general minmax theorems*, Pacific J. Math., 8, pp. 171–176.

S. Sorin (2000), *A First Course on Zero-Sum Repeated Games*, preprint.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 2, pp. 525–539

A DYNAMIC PROGRAMMING ALGORITHM FOR THE OPTIMAL CONTROL OF PIECEWISE DETERMINISTIC MARKOV PROCESSES*

ANTHONY ALMUDEVAR†

Abstract. A piecewise deterministic Markov process (PDP) is a continuous time Markov process consisting of continuous, deterministic trajectories interrupted by random jumps. The trajectories may be controlled with the object of minimizing the expected costs associated with the process. A method of representing this controlled PDP as a discrete time decision process is presented, allowing the value function for the problem to be expressed as the fixed point of a dynamic programming operator. Decisions take the form of trajectory segments. The expected costs may then be minimized through a dynamic programming algorithm, rather than through the solution of the Bellman–Hamilton–Jacobi equation, assuming the trajectory segments are numerically tractable. The technique is applied to the optimal capacity expansion problem, that is, the problem of planning the construction of new production facilities to meet rising demand.

Key words. piecewise deterministic Markov process, dynamic programming, capacity expansion

AMS subject classifications. 93E20, 60J75

PII. S0363012999364474

1. Introduction.

In this paper a technique for minimizing expected costs associated with piecewise deterministic Markov processes (PDPs) is developed. Such processes may be described as continuous time Markov processes consisting of continuous, deterministic trajectories interrupted by random jumps. A comprehensive definition and theoretical development of these processes can be found in Davis [4]. Many problems in operations research can be naturally expressed in this framework; hence there is a great deal of interest in optimization problems associated with these processes.

A PDP is usually defined on a state space $E \subset \Re^p$ partitioned into a boundary E_δ and interior E_o, although the state space definition in [4] is somewhat more general. We let $\mathcal{E}$ denote the Borel subsets of E, and we will let $\mathcal{P}(E)$ be the space of probability measures on the measurable space $(E, \mathcal{E})$, endowed with the topology of weak convergence. Under suitable regularity conditions a PDP can be uniquely determined by a vector field $f : E \to \Re^p$, an intensity function $\lambda : E \to \Re^+$, and stochastic kernels $q_o : E_o \to \mathcal{P}(E)$ and $q_\delta : E_\delta \to \mathcal{P}(E)$. Between jumps the PDP $\hat{x}(t)$ obeys $d\hat{x}(t)/dt = f(\hat{x}(t))$, and jumps occur at rate $\lambda(x)$ when the process is at state x, independently of the process history. If a jump occurs at $x \in E_o$, the process is transferred immediately to a new state given randomly by probability measure $q_o(dx \mid x)$. If the process reaches the boundary at $x \in E_\delta$, the process is transferred immediately to a new state given randomly by probability measure $q_\delta(dx \mid x)$. We will always assume that $q_o(E_o \mid x) = 1$ and $q_\delta(E_o \mid x) = 1$.

A controlled PDP is defined when the quadruple $(f, \lambda, q_o, q_\delta)$ is allowed to depend on a control parameter u. In addition, cost is assumed at a rate $l_o(x, u)$ when the process is at $x \in E_o$ and control u is applied, and a discrete cost $l_\delta(x, u)$ is assumed

*Received by the editors November 10, 1999; accepted for publication (in revised form) February 19, 2001; published electronically July 25, 2001. This research was supported by the Natural Sciences and Engineering Research Council of Canada.
http://www.siam.org/journals/sicon/40-2/36447.html
†Department of Mathematics and Computing Science, Saint Mary's University, Halifax, NS, Canada, B3H 3C3 (anthony.almudevar@stmarys.ca).

when the process reaches the boundary at $x \in E_\delta$ and control u is applied. A *control policy* Φ is equivalent to a specification for each $x \in E$ of an open loop continuous time control function to be applied from x until the next jump (Vermes [11]).

If we define $J_\Phi(x)$ to be the expected cost under control policy Φ from initial state x, possibly under geometric discounting, the *value function* is then defined as

$$J^*(x) = \inf_\Phi J_\Phi(x),$$

where the infimum is taken over all admissible control policies. The object is to find, if it exists, a control policy whose expected cost achieves this infimum.

In the existing literature the value function for this problem is typically given as a solution to a Bellman–Hamilton–Jacobi (BHJ) equation. In [11] a limiting form of the BHJ equation is given as a necessary and sufficient optimality condition. In Dempster and Ye [9] a generalized BHJ equation, expressed in terms of the generalized Clarke gradient (Clarke [3]), is given as a necessary and sufficient optimality condition. In Soner [10] a viscosity solution approach to the BHJ equation is proposed, and, more recently, the viscosity solution to the BHJ equation has been developed in Davis and Farid [8], which has advantages with respect to the availability of numerical methods for solution.

In this paper we use an approach similar to that introduced by Davis [5] and developed in [9] and Davis [6], in which the problem is reformulated in terms of an imbedded discrete time process, in which a stage consists of the intrajump deterministic portion of the process. The principal difference is that in the approach proposed in this article the problem remains in the discrete time domain up to and including the solution algorithm. The concept of a continuously applied control parameter will play no role. Instead, a discrete time decision process is defined in which a decision consists of the selection of a trajectory segment, in this way constructing the deterministic trajectory in a piecewise fashion. This means that the BHJ equation plays no role. Ultimately, the value function is calculable as a fixed point of a dynamic programming operator in discrete time. Here we do not admit direct control over the cost function and the jump rate, unlike the other models cited in the above literature, although in principle the methodology could be extended to incorporate the cost function and the jump rate into the action space.

Apart from the more limited control, this allows a more uniform approach to the calculation of optimal policies and a weakening of regularity conditions. For the generalized BHJ equation in [9], conditions are imposed which guarantee that the value function is Lipschitz, which excludes many problems of practical importance (see [6]). The viscosity solution approach allows milder assumptions. In [8] the state space is required to be bounded, and the cost rate and jump rate are assumed to be bounded and uniformly continuous. In comparison, in this article the state space need not be bounded, the jump rate is bounded but not necessarily uniformly continuous, and the cost rate may be lower semicontinuous and need not be bounded. In fact, conditions are placed only on suitably defined integrals of the cost rate (see section 2). As for the trajectory, in the context of the BHJ equation the vector field f is generally assumed to be Lipschitz. In the approach presented here there is no explicit vector field and no other restrictions other than that the path can be constructed in a piecewise manner from trajectories selected from a compact set. This admits a wider variety of control structures, including certain types of impulse controls.

In section 2, we define a discrete time decision process imbedded in the PDP and obtain conditions under which the resulting transition measure will be continuous

on the state-action product space. In section 3, we discuss some results for discrete time decision processes from Bertsekas and Shreve [1] which may be applied to the problem under consideration here. In addition, with some additional assumptions we show that the dynamic programming operator is a contraction mapping. In section 4 we show how this may be applied to the capacity expansion problem considered in Davis et al. [7]. Section 5 contains some concluding remarks and possible extensions of this work.

2. Reduction of a PDP to a discrete time process. Let $E \subset \Re^p$ be a state space containing a boundary E_δ. Let $E_o = E - E_\delta$ be the interior of E. Possibly, $E_\delta = \emptyset$, the empty set. We also have the intensity function and stochastic kernels (λ, q_o, q_δ) as defined in section 1, all assumed to be Borel measurable mappings.

Then let I_T be a time scale interval $[0, T]$ if $T < \infty$ and $[0, \infty)$ if $T = \infty$. Let A be an action space consisting of a family of continuous trajectories $\alpha : I_T \to \Re^p$ with $\alpha(0) = 0$. It will be assumed that A is a compact metric space in which convergence implies pointwise convergence. We define

$$B(x, \alpha) = \inf\{t \in I_T : x + \alpha(t) \in E_\delta\},$$

which is the time taken for the trajectory $x + \alpha(t)$ to reach the boundary, and let

$$t_f(x, \alpha) = \min\{B(x, \alpha), T\},$$

adopting the convention that $\inf \emptyset = \infty$. For each $x \in E$ let $A_x \subset A$ be a subset of trajectories available at state x, which gives the *state-action space*

$$\Gamma = \{(x, \alpha) \in E \times A : \alpha \in A_x\}.$$

We assume that $x + \alpha(t) \in E$ when $t \le t_f(x, \alpha)$ for all $(x, \alpha) \in \Gamma$. Generally, the following condition will be satisfied:

(A.1) $x+\alpha(t) = x+\alpha(B(x,\alpha))\,\forall\, t \ge B(x,\alpha), t \in I_T,$ when $(x, \alpha) \in \Gamma$ and $B(x, \alpha) < \infty$,

which implies that a trajectory comes to rest upon reaching the boundary. In addition, (A.1) implies $x+\alpha(B(x, \alpha)) \in E_\delta$ for all $(x, \alpha) \in \Gamma$ and that the only admissable action when $x \in E_\delta$ is $\alpha \equiv 0$.

We can define iteratively the continuous time process $\{\hat{x}(t) \in E : t \ge 0\}$ and the imbedded discrete time decision process $\{(\hat{x}_n, \hat{\alpha}_n) \in \Gamma : n \ge 0\}$ with the associated event time process $\{\hat{t}_n \ge 0 : n \ge 0\}$. Suppose we have state $\hat{x}_n \in E_o$, decision $\hat{\alpha}_n \in A_{x_n}$, and time $\hat{t}_n$. The process then follows the trajectory

$$\hat{x}(t) = \hat{x}_n + \hat{\alpha}_n(t - \hat{t}_n), \quad t \ge \hat{t}_n, \tag{2.1}$$

until time $\hat{t}_n + t_f(\hat{x}_n, \hat{\alpha}_n)$, unless a random jump occurs along the trajectory before then, say, at time $t' \in (\hat{t}_n, \hat{t}_n + t_f(\hat{x}_n, \hat{\alpha}_n))$, in which case (2.1) holds until t'. These jumps occur at rate $\lambda(x)$ when the process is in state $x \in E_o$, independently of the process history. If such a jump occurs at state x', then the new state $\hat{x}_{n+1} \in E_o$ is given randomly by the distribution $q_o(dx \mid x')$, and we set $\hat{t}_{n+1} = t'$. If no jump occurs before $\hat{t}_n + t_f(\hat{x}_n, \hat{\alpha}_n)$ then set $\hat{t}_{n+1} = \hat{t}_n + t_f(\hat{x}_n, \hat{\alpha}_n)$. In this case, if the process has reached the boundary at state $x' \in E_\delta$ (i.e., $B(\hat{x}_n, \hat{\alpha}_n) < \infty$), then the new state $\hat{x}_{n+1} \in E_o$ is given randomly by the distribution $q_\delta(dx \mid x')$. Otherwise, if the end of the trajectory segment $\hat{\alpha}_n$ has been reached before the boundary (i.e.,

$B(\hat{x}_n, \hat{\alpha}_n) = \infty$, $T < \infty$), then set $\hat{x}_{n+1} = \hat{x}_n + \hat{\alpha}_n(\hat{t}_{n+1} - \hat{t}_n)$. A new decision $\hat{\alpha}_{n+1} \in A_{\hat{x}_{n+1}}$ is then made. An initial state and decision $(\hat{x}_0, \hat{\alpha}_0) \in \Gamma$ is specified, with $\hat{t}_0 = 0$. If $\hat{x}_0 \in E_\delta$, we will set $\hat{t}_1 = \hat{t}_0 = 0$, $\hat{\alpha}_0 \equiv 0$, and $\hat{x}_1$ will be determined by $q_\delta(dx \mid \hat{x}_0)$. Then $\hat{x}_k \in E_o$ for $k \geq 1$.

This defines the transition measure $Q : \Gamma \to \mathcal{P}(E)$ for the process $(\hat{x}_n, \hat{\alpha}_n)$, where $Q(K \mid x, \alpha)$ is the probability that $\hat{x}_{n+1} \in K$ given that trajectory $\hat{\alpha}_n = \alpha$ is selected at state $\hat{x}_n = x$. Assuming (A.1) holds, this is given explicitly by

$$\begin{aligned} Q(K \mid x, \alpha) = &\int_0^{t_f(x,\alpha)} q_o(K \mid x + \alpha(t))\lambda(x + \alpha(t)) \exp(-\Lambda(t, x, \alpha))\, dt \\ &+ q_\delta(K \mid x + \alpha(B(x, \alpha))) \exp(-\Lambda(B(x, \alpha), x, \alpha)) I\{B(x, \alpha) < \infty\} \\ &+ I\{x + \alpha(T) \in K\} \exp(-\Lambda(T, x, \alpha)) I\{B(x, \alpha) = \infty, T < \infty\}, \end{aligned} \tag{2.2}$$

where

$$\Lambda(t, x, \alpha) = \int_0^t \lambda(x + \alpha(w))\, dw.$$

Here, $I\{S\}$ is the indicator function of set S. Since we assume $q_o(E_o \mid x) = 1$ and $q_\delta(E_o \mid x) = 1$, we necessarily have $Q(E_o \mid x, \alpha) = 1$ for $(x, \alpha) \in \Gamma$. Note also that if $x \in E_\delta$, we have $\alpha \equiv 0$, $t_f(x, \alpha) = B(x, \alpha) = 0$, and $Q(K \mid x, \alpha) = q_\delta(K \mid x)$. It will be useful to know when the transition measure is continuous with respect to weak convergence on Γ (with $E \times A$ assuming the product topology). We prove below that Q will be continuous under the following assumptions:

(B.1) $q_o(dx \mid x)$ is continuous on E_o with respect to weak convergence.

(B.2) $q_\delta(dx \mid x)$ is continuous on E_δ with respect to weak convergence.

(B.3) λ is continuous on E, $\lambda \leq M_\lambda$ for some $M_\lambda < \infty$.

(B.4) The sets $B_1 = \{(x, \alpha) \in \Gamma : B(x, \alpha) < \infty\}$ and $B_2 = \{(x, \alpha) \in \Gamma : B(x, \alpha) = \infty\}$ are both closed.

(B.5) $B(x, \alpha)$ is continuous on B_1.

Remark. If a nontrivial boundary is present, a special condition is typically necessary for the continuity of Q to hold. Generally, some assumption which governs the behavior of the trajectory near the boundary is required. Informally, these assumptions typically require that if a trajectory approaches the boundary, it does so in some direct manner. In [11] the minimum velocity in the direction normal to the boundary is bounded away from 0. This condition is weakened in [8] to require only that where the boundary is approachable it is approachable nontangentially. Assumptions (B.4) and (B.5) are used here to govern trajectory behavior near the boundary. They will not be natural to many problems and are not satisfied by the capacity expansion problem considered in [7]. However, we show in section 4 how a reasonable redefinition of the problem can force (B.4) and (B.5) to hold.

THEOREM 2.1. *If assumptions* (A.1) *and* (B.1)–(B.5) *hold, then* $Q(dx \mid x, \alpha)$ *is continuous on* Γ *with respect to weak convergence.*

Proof. Let $\gamma = \{(x_n, \alpha_n) : n \geq 1\}$ be a convergent sequence in Γ with limit (x_0, α_0). We then have

$$\lim_{n \to \infty} x_n + \alpha_n(t) = x_0 + \alpha_0(t) \quad \forall t \in I_T,$$

and hence by (B.3)

$$\lim_{n\to\infty} \lambda(x_n + \alpha_n(t)) = \lambda(x_0 + \alpha_0(t)) \quad \forall t \in I_T. \tag{2.3}$$

By (B.3) λ is bounded, so applying the dominated convergence theorem gives

$$\lim_{n\to\infty} \Lambda(t, x_n, \alpha_n) = \Lambda(t, x_0, \alpha_0) \quad \forall t \in I_T. \tag{2.4}$$

Next, recall that if $\{\mu_n : n \geq 1\}$ is any sequence of probability measures in $\mathcal{P}(E)$, an equivalent definition of weak convergence of the sequence to a probability measure μ_0 is

$$\liminf_{n\to\infty} \mu_n(K) \geq \mu_0(K) \quad \forall \text{ open sets } K$$

(see, for example, Theorem 29.1 in Billingsley [2]), so it suffices to show that

$$\liminf_{n\to\infty} Q(K \mid x_n, \alpha_n) \geq Q(K \mid x_0, \alpha_0) \quad \forall \text{ open sets } K \in \mathcal{E}, \tag{2.5}$$

for each convergent sequence γ. Since B_1 and B_2 are closed, we may assume that $\gamma \subset B_1$ or $\gamma \subset B_2$. We now examine separately the three following cases.

Case 1: $T = \infty$, $\gamma \subset B_2$. In this case we have

$$Q(K \mid x_n, \alpha_n) = \int_0^\infty q_o(K \mid x_n + \alpha_n(t))\lambda(x_n + \alpha_n(t)) \exp(-\Lambda(t, x_n, \alpha_n))\, dt.$$

By (B.1), (2.3), and (2.4) we may assert for open K

$$\begin{aligned} &\liminf_{n\to\infty} q_o(K \mid x_n + \alpha_n(t))\lambda(x_n + \alpha_n(t)) \exp(-\Lambda(t, x_n, \alpha_n)) \\ &\qquad \geq q_o(K \mid x_0 + \alpha_0(t))\lambda(x_0 + \alpha_0(t)) \exp(-\Lambda(t, x_0, \alpha_0)); \end{aligned} \tag{2.6}$$

hence (2.5) holds by Fatou's lemma.

Case 2: $T < \infty$, $\gamma \subset B_2$. In this case we have

$$\begin{aligned} Q(K \mid x_n, \alpha_n) = &\int_0^T q_o(K \mid x_n + \alpha_n(t))\lambda(x_n + \alpha_n(t)) \exp(-\Lambda(t, x_n, \alpha_n))\, dt \\ &+ I\{x_n + \alpha_n(T) \in K\} \exp(-\Lambda(T, x_n, \alpha_n)) \end{aligned} \tag{2.7}$$

for all $n \geq 0$. Using an argument similar to that used for Case 1, we have

$$\begin{aligned} &\liminf_{n\to\infty} \int_0^T q_o(K \mid x_n + \alpha_n(t))\lambda(x_n + \alpha_n(t)) \exp(-\Lambda(t, x_n, \alpha_n))\, dt \\ &\qquad \geq \int_0^T q_o(K \mid x_0 + \alpha_0(t))\lambda(x_0 + \alpha_0(t)) \exp(-\Lambda(t, x_0, \alpha_0))\, dt \end{aligned} \tag{2.8}$$

for all open sets $K \in \mathcal{E}$. Then

$$\liminf_{n\to\infty} I\{x_n + \alpha_n(T) \in K\} \geq I\{x_0 + \alpha_0(T) \in K\}$$

for open $K \in \mathcal{E}$, which, when combined with (2.4), (2.7), and (2.8), gives (2.5) for Case 2.

Case 3: $\gamma \subset B_1$. In this case we necessarily have $t_f(x_n, \alpha_n) = B(x_n, \alpha_n)$, $n \geq 0$, so that

$$Q(K \mid x_n, \alpha_n) = \int_0^{B(x_n,\alpha_n)} q_o(K \mid x_n + \alpha_n(t))\lambda(x_n + \alpha_n(t)) \exp(-\Lambda(t, x_n, \alpha_n))\, dt \\ + q_\delta(K \mid x_n + \alpha_n(B(x_n, \alpha_n))) \exp(-\Lambda(B(x_n, \alpha_n), x_n, \alpha_n))$$

for all $n \geq 0$. By assumption (B.5) $B(x_n, \alpha_n) \to_n B(x_0, \alpha_0)$. Then, using (2.6), we have

$$\liminf_{n\to\infty} q_o(K \mid x_n + \alpha_n(t))\lambda(x_n + \alpha_n(t)) \exp(-\Lambda(t, x_n, \alpha_n)) I\{t \leq B(x_n, \alpha_n)\} \\ \geq q_o(K \mid x_0 + \alpha_0(t))\lambda(x_0 + \alpha_0(t)) \exp(-\Lambda(t, x_0, \alpha_0)) I\{t < B(x_0, \alpha_0)\}$$

for open $K \in \mathcal{E}$, so by Fatou's lemma

$$\liminf_{n\to\infty} \int_0^{B(x_n,\alpha_n)} q_o(K \mid x_n + \alpha_n(t))\lambda(x_n + \alpha_n(t)) \exp(-\Lambda(t, x_n, \alpha_n))\, dt \\ \geq \int_0^{B(x_0,\alpha_0)} q_o(K \mid x_0 + \alpha_0(t))\lambda(x_0 + \alpha_0(t)) \exp(-\Lambda(t, x_0, \alpha_0))\, dt \tag{2.9}$$

for all open sets $K \in \mathcal{E}$. By assumption (A.1) we must have

$$\lim_{n\to\infty} x_n + \alpha_n(B(x_n, \alpha_n)) = x_0 + \alpha_0(B(x_0, \alpha_0)),$$

and by assumption (B.2)

$$\liminf_{n\to\infty} q_\delta(K \mid x_n + \alpha_n(B(x_n, \alpha_n))) \geq q_\delta(K \mid x_0 + \alpha_0(B(x_0, \alpha_0)))$$

for all open sets $K \in \mathcal{E}$. We then have

$$\lim_{n\to\infty} \lambda(x_n + \alpha_n(t)) I\{t \leq B(x_n, \alpha_n)\} = \lambda(x_0 + \alpha_0(t)) I\{t \leq B(x_0, \alpha_0)\},$$

except possibly at $t = B(x_0, \alpha_0)$. The sequence $\{B(x_n, \alpha_n) : n \geq 1\}$ is bounded since B_1 is closed. Then with assumption (B.3) the dominated convergence theorem applies, giving

$$\lim_{n\to\infty} \Lambda(B(x_n, \alpha_n), x_n, \alpha_n) = \Lambda(B(x_0, \alpha_0), x_0, \alpha_0),$$

so that

$$\liminf_{n\to\infty} q_\delta(K \mid x_n + \alpha_n(B(x_n, \alpha_n))) \exp(-\Lambda(B(x_n, \alpha_n), x_n, \alpha_n)) \\ \geq q_\delta(K \mid x_0 + \alpha_0(B(x_0, \alpha_0))) \exp(-\Lambda(B(x_0, \alpha_0), x_0, \alpha_0))$$

for all open sets $K \in \mathcal{E}$, which with (2.9) gives (2.5) for Case 3, which completes the proof. □

With respect to assumption (B.5), assumption (A.1) is sufficient to guarantee the lower semicontinuity of $B(x, \alpha)$ on B_1, as shown in Lemma 2.2 below, but upper semicontinuity must be verified separately.

LEMMA 2.2. *If assumption* (A.1) *holds,* $B(x, \alpha)$ *is lower semicontinuous on* B_1.

Proof. Suppose that $\gamma = \{(x_n, \alpha_n) : n \geq 1\}$ is a convergent sequence in B_1 with limit $(x_0, \alpha_0) \in B_1$. We show that

$$\liminf_{n\to\infty} B(x_n, \alpha_n) \geq B(x_0, \alpha_0) \tag{2.10}$$

for any such sequence. Suppose there exists an infinite subsequence $\{(x_{n_k}, \alpha_{n_k}) : k \geq 1\}$ and a $\beta < B(x_0, \alpha_0)$ such that $B(x_{n_k}, \alpha_{n_k}) \leq \beta$ for all $k \geq 1$. Then

$$\lim_{k\to\infty} x_{n_k} + \alpha_{n_k}(B(x_0, \alpha_0)) = x_0 + \alpha_0(B(x_0, \alpha_0)), \tag{2.11}$$

and by assumption (A.1) and the fact that $B(x_{n_k}, \alpha_{n_k}) \leq \beta < B(x_0, \alpha_0)$, $k \geq 1$, we must have

$$\begin{aligned} \lim_{k\to\infty} x_{n_k} + \alpha_{n_k}(B(x_0, \alpha_0)) &= \lim_{k\to\infty} x_{n_k} + \alpha_{n_k}(\beta) \\ &= x_0 + \alpha_0(\beta). \end{aligned} \tag{2.12}$$

However, (2.11) and (2.12) are contradictory since $x_0 + \alpha_0(B(x_0, \alpha_0)) \in E_\delta$, but $x_0 + \alpha_0(\beta) \in E_o$; hence any convergent sequence must satisfy (2.10). □

Finally, we assume there is a nonnegative expected cost $g : \Gamma \to \Re^+$ associated with each stage. This cost may be specified by letting C_T be the family of measurable functions $c : I_T \to \Re^+$. The cost of a stage is then determined by a mapping $h_o : \Gamma \to C_T$ which represents the rate at which cost is assumed at a time t after decision α is made from state x. We may also have a boundary cost $h_\delta(x)$, $x \in E_\delta$, assumed when the process reaches the boundary at x. Then if $W_{(x,\alpha)}$ is the random time spent in the stage, the cost assumed in the stage given $W_{(x,\alpha)} = w$ is

$$H_{(x,\alpha)}(w) = \int_0^w h_o(t \mid x, \alpha)\, dt + h_\delta(x + \alpha(B(x, \alpha)))I\{w = B(x, \alpha), B(x, \alpha) < \infty\}.$$

Then g is given by

$$\begin{aligned} g(x, \alpha) &= E[H_{(x,\alpha)}(W_{(x,\alpha)})] \\ &= \int_0^{t_f(x,\alpha)} h_o(t \mid x, \alpha) \exp(-\Lambda(t, x, \alpha))\, dt \end{aligned} \tag{2.13}$$

$$+ h_\delta(x + \alpha(B(x, \alpha))) \exp(-\Lambda(B(x, \alpha), x, \alpha))I\{B(x, \alpha) < \infty\}. \tag{2.14}$$

In the following discussion any regularity condition will be placed on g directly.

3. Optimization for lower semicontinuous costs. We give a general definition (following [1]) of a stochastic discrete time decision process $\{(\hat{x}_n, \hat{\alpha}_n) : n \geq 0\}$, where $\hat{x}_n$ and $\hat{\alpha}_n$ are elements of a state space and action space E and A, both assumed to be Borel spaces. Let $\mathcal{P}(E)$ and $\mathcal{P}(A)$ be the space of all probability measures on the Borel sets of E and A, respectively, endowed with the topology of weak convergence. For each $x \in E$ we assume that there is a set of available actions $A_x \subset A$. We then have state-action space

$$\Gamma = \{(x, \alpha) \in E \times A : \alpha \in A_x\},$$

where $E \times A$ is endowed with the product topology (and is also a Borel space). We assume there is a stochastic kernel $Q(dx \mid x, \alpha)$ which is a Borel measurable mapping

from Γ to $\mathcal{P}(E)$. Finally, we have a lower semianalytic cost function $g : \Gamma \to \Re^+$. Define a policy

$$\Phi = \{\tilde{\phi}_n : n \geq 0\}$$

as a sequence of stochastic kernels $\tilde{\phi}_n(dy \mid x_0, \alpha_0, \ldots, x_{n-1}, \alpha_{n-1}, x_n)$ which are universally measurable mappings from $(\times^n \Gamma) \times E$ to $\mathcal{P}(A)$ satisfying

$$\tilde{\phi}_n(A_{x_n} \mid x_0, \alpha_0, \ldots, x_{n-1}, \alpha_{n-1}, x_n) = 1,$$

and let Π be the class of all such policies. For a given policy $\Phi \in \Pi$ the process $(\hat{x}_n, \hat{\alpha}_n)$ can then be defined iteratively by considering a current state $\hat{x}_n$ and the process history $\{(\hat{x}_k, \hat{\alpha}_k) : k = 0, \ldots, n-1\}$. Decision $\hat{\alpha}_n$ is then given randomly by the distribution $\tilde{\phi}_n(dy \mid \hat{x}_0, \hat{\alpha}_0, \ldots, \hat{x}_{n-1}, \hat{\alpha}_{n-1}, \hat{x}_n)$, and then state $\hat{x}_{n+1}$ is given randomly by the distribution $Q(dx \mid \hat{x}_n, \hat{\alpha}_n)$. We are given an initial state $\hat{x}_0$. Then a cost of $\sum_n g(\hat{x}_n, \hat{\alpha}_n)$ is assumed. (We do not consider at this point geometric discounting.) Define

$$J_\Phi(x) = E\left[\sum_{n=0}^{\infty} g(\hat{x}_n, \hat{\alpha}_n) \mid \hat{x}_0 = x\right],$$

which denotes the expected cost assumed by the process under policy Φ with initial state $\hat{x}_0 = x$. If $\tilde{\phi}_n$ is parametrized by x_n only, then Φ is a Markov policy. Let Π_1 be the class of all mappings $\phi : E \to A$ with $\phi(x) \in A_x$ for all $x \in E$. We will be interested primarily in *nonrandomized stationary Markov policies*, that is, policies for which there is some $\phi \in \Pi_1$ such that for all $n \geq 0$, $\tilde{\phi}_n(dy \mid x_n)$ is a point mass at $\phi(x_n)$. (In this case we will simply write $\tilde{\phi}_n(dy \mid x_n) = \phi(x_n)$.)

We then define the problem:

(P) minimize $J_\Phi(x)$ over all policies $\Phi \in \Pi$ for each $x \in E$.

Define the value function

$$J^*(x) = \inf_{\Phi \in \Pi} J_\Phi(x), \quad x \in E.$$

For universally measurable $J : E \to \Re^+$ define the operator T mapping J to $TJ : E \to \Re^+$ by

$$(TJ)(x) = \inf_{\alpha \in A_x} \left(g(x, \alpha) + \int_E J(x') Q(dx' | x, \alpha) \right) \tag{3.1}$$

for all $x \in E$. For $\phi \in \Pi_1$, define the operator T_ϕ mapping universally measurable $J : E \to \Re^+$ to $T_\phi J : E \to \Re^+$ by

$$(T_\phi J)(x) = g(x, \phi(x)) + \int_E J(x') Q(dx' \mid x, \phi(x))$$

for all $x \in E$. Letting $J_0 \equiv 0$, define the sequence

$$J_{k+1} = TJ_k, \quad k \geq 0. \tag{3.2}$$

(The sequence is well defined, since if J is lower semianalytic, so is TJ. See [1, Section 8.2].)

It is easy to verify that $g \geq 0$ implies that T is monotone in the sense that $TJ_2 \geq TJ_1$ if $J_2 \geq J_1$. Then $J_1 \geq J_0$, and hence $J_2 = TJ_1 \geq TJ_0 = J_1$. By extending this argument we conclude that $\{J_k\}$ is increasing, so that the limit

$$J_\infty = \lim_{k\to\infty} J_k \tag{3.3}$$

exists.

The model defined in this section is a *lower semicontinuous model with positive cost* according to the definition given in [1, Definition 8.7, p. 208] if the following conditions hold:

(C.1) A is compact.

(C.2) Γ is a closed subset of $E \times A$.

(C.3) $g(x, \alpha)$ is lower semicontinuous on Γ.

(C.4) The transition measure $Q(dx \mid x, \alpha)$ is weakly continuous on Γ.

We summarize some results from [1, Proposition 8.6, Corollary 9.4.1, Proposition 9.8, Corollary 9.17.2, Proposition 9.18, pp. 209, 221, 225, 235, 236] in the following theorem.

THEOREM 3.1. *Under assumptions* (C.1)–(C.4), *the following hold.*

(i) *If $J \in \mathcal{J}$ is lower semicontinuous, then so is TJ (from proof of Proposition* 8.6*).*

(ii) *J^* is lower semianalytic, and $J^* = TJ^*$ (Corollary* 9.4.1*, Proposition* 9.8*).*

(iii) *There exists a Borel measurable nonrandomized stationary Markov policy Φ^* such that $J_{\Phi^*} = J^*$ (Corollary* 9.17.2*).*

(iv) *$J^* = J_\infty$, where J_∞ is lower semicontinuous (Corollary* 9.17.2*).*

(v) *There exists a sequence $\{\phi_k \in \Pi_1 : k \geq 0\}$, where ϕ_k is universally measurable, such that $T_{\phi_k} J_k = TJ_k$, $k \geq 0$. Each sequence $\{\phi_k(x)\}$, $x \in E$, has an accumulation point. If $\phi^* \in \Pi_1$ is universally measurable and $\phi^*(x)$ is an accumulation point of $\{\phi_k(x)\}$ when $J^*(x) < \infty$, then $\Phi^* = (\phi^*, \phi^*, \ldots)$ is an optimal policy (Proposition* 9.18*).*

Suppose the state space E contains a measurable set E_K such that once the process enters E_K it does not leave and it assumes no further cost. Let $\mathcal{J}_K$ be the set of all $J : E \to \Re^+$ with $J(x) = 0$ for all $x \in E_K$. We must then have $J_\Phi \in \mathcal{J}_K$ for any policy Φ. Furthermore, suppose there is some $r > 0$ such that from any state the probability of subsequently entering E_K is at least r for all $(x, \alpha) \in \Gamma$. Under these assumptions, it is shown below that T is a contraction mapping on $\mathcal{J}_K$; hence there is at most one fixed point of T in $\mathcal{J}_K$. This can be summarized by the following assumptions:

(D.1) $\exists r > 0$ such that $Q(E_K \mid x, \alpha) \geq r$ for all $(x, \alpha) \in \Gamma$.

(D.2) $Q(E_K \mid x, \alpha) = 1$ for all $x \in E_K, \alpha \in A$.

(D.3) $g(x, \alpha) = 0$ for all $x \in E_K, \alpha \in A$.

THEOREM 3.2. *Suppose* (D.1)–(D.3) *hold. Then T is a contraction mapping of universally measurable $J \in \mathcal{J}_K$ to $\mathcal{J}_K$ with contraction constant $1 - r$.*

Proof. By (D.2), (D.3), and the definition of T, if $J(x) = 0$ on E_K for universally measurable J, then $TJ \in \mathcal{J}_K$.

If $J_1, J_2 \in \mathcal{J}_K$ are universally measurable, then for any $\phi \in \Pi_1$ we have

$$\begin{aligned} \|T_\phi J_2 - T_\phi J_1\| &= \sup_{x\in E} \left| \int_E J_2(x') Q(dx' \mid x, \phi(x)) - \int_E J_1(x') Q(dx' \mid x, \phi(x)) \right| \\ &\leq \sup_{x\in E} \int_E |J_2(x') - J_1(x')| Q(dx' \mid x, \phi(x)) \end{aligned}$$

$$\begin{aligned}
&\leq \sup_{x\in E} \int_{E-E_K} |J_2(x') - J_1(x')| Q(dx' \mid x, \phi(x)) \\
&\leq \sup_{x\in E} \|J_2 - J_1\|(1 - Q(E_K \mid x, \phi(x))) \\
&\leq \|J_2 - J_1\|(1 - r)
\end{aligned}$$

since $|J_2(x) - J_1(x)| = 0$ when $x \in E_K$. For $\epsilon > 0$ we may select ϕ so that

$$T_\phi J_1 \leq TJ_1 + \epsilon,$$

and then

$$\begin{aligned}
TJ_2 - TJ_1 &\leq TJ_2 - T_\phi J_1 + \epsilon \\
&\leq T_\phi J_2 - T_\phi J_1 + \epsilon \\
&\leq \|J_2 - J_1\|(1 - r) + \epsilon.
\end{aligned}$$

This holds for all $\epsilon > 0$, so $TJ_2 - TJ_1 \leq \|J_2 - J_1\|(1 - r)$. A similar argument gives $TJ_1 - TJ_2 \leq \|J_2 - J_1\|(1 - r)$, completing the proof. □

The model discussed in this section is directly applicable to the imbedded discrete time decision process introduced in section 2. Using the notation of that section, if E and E_o are measurable subsets of $\Re^p$ and if A can be defined as a compact metric space, then E and A are both Borel spaces; then it remains to verify that Γ is closed. It must then be verified that the transition measure (2.2) is continuous on Γ, possibly through Theorem 2.1. Then the cost g must be lower semicontinuous on Γ. Under these conditions, assumptions (C.1)–(C.4) hold and Theorem 3.1 applies, and the optimum expected cost may be calculated through the dynamic programming algorithm (3.2)–(3.3). An optimal policy may be obtained as the limit defined in Theorem 3.1(v).

With respect to the process of section 2, assumption (D.1) will hold under various circumstances. Geometric discounting may be introduced by adding to E a kill state Δ and assuming that the process jumps to Δ at some fixed rate. If λ is bounded and $B(x, \alpha)$ is bounded away from 0, then assumption (D.1) will be satisfied. Alternatively, there may be some target set which the state-action space is constrained to reach in one stage within some bounded time, barring a jump. If the process remains in this set with no further costs, then assumption (D.1) will be satisfied.

4. The capacity expansion problem. We now consider the optimal capacity expansion problem considered by Davis et al. [7]. We suppose that for a certain commodity there is a demand rate d which increases in time according to a compound Poisson process with constant rate $\lambda > 0$. Suppose there are enough plants to supply the commodity at rate s. At any time a decision to build a new plant may be made, which requires a total cost of C. Let y be the amount already invested in the plant being currently built. If no plant is currently being built, then $y = 0$. The rate of investment will then be $\dot{y} \in [0, c]$, where c represents the maximum possible investment rate. Once a plant is completed, capacity s is increased by L units. We let $z = s - d$. If $z > 0$, then there is overcapacity, and if $z < 0$, there is undercapacity. Let $h : \Re \to \Re^+$ represent the rate at which cost is assumed due to overcapacity or undercapacity z with $h(0) = 0$.

The problem is to derive a policy, giving the investment rate at any state, which minimizes the total expected cost under geometric discounting. In [7] a technique for solving the BHJ equation for this problem is given. It should be noted that an

optimal solution does not necessarily exist. Examples are given in [7] of a problem in which for certain values of z it is ϵ-optimal to build the current plant to within a small amount β of completion, with the expected cost function improving as β approaches 0, but not optimal to complete it. This suggests introducing as a control constraint the requirement that a plant be completed if it is within some fixed amount of completion. It will be shown below that this constraint forces assumptions (B.4) and (B.5) to hold. It is also shown in [7] that any optimal policy will specify either maximum investment rate c or minimum investment rate 0.

We need to specify $E_\delta, E_o, A, \Gamma, \lambda, q_o, q_\delta, g$ as defined in section 2. The state space will be

$$\begin{aligned} E &= [0, C] \times \Re, \\ E_o &= [0, C) \times \Re, \\ E_\delta &= E - E_o. \end{aligned}$$

Then we interpret $(y, z) \in E$ as the state at which the current plant has y currently invested and $z = s - d$. As in [7], we will suppose that the investment rate is either 0 or c. Hence from a starting point (y, z) the decision will consist of determining how much to invest in the current plant at rate c. The action space A is then the family of parametric curves $\alpha : [0, \infty) \to \Re^2$ of the form

$$\alpha(t) = \begin{cases} (ct, 0), & 0 \le t < a/c, \\ (a, 0), & a/c \le t, \end{cases} \tag{4.1}$$

for $a \in [0, C]$. Thus we have $T = \infty$. The trajectories in (4.1) are homeomorphic to the interval $[0, C]$, so that A will subsequently be represented by $[0, C]$. We let β be a positive constant less than C. The action space will be constrained so that if $y \in (C - \beta, C]$, the project must be completed at rate c, and if $y \in [0, C - \beta]$, then an amount $a \le C - \beta - y$ or $a = C - y$ may be invested. Hence we have state-action space

$$\begin{aligned} \Gamma &= \Gamma_1 \cup \Gamma_2, \\ \Gamma_1 &= \{(y, z, a) \in E \times [0, C] : y + a = C\}, \\ \Gamma_2 &= \{(y, z, a) \in E \times [0, C] : y + a \le C - \beta\}. \end{aligned}$$

We introduce geometric discounting by adding to E a kill state Δ to which the process jumps at a rate $\eta > 0$. At this state no further costs are assumed. We can then define the overall jump intensity as $\lambda(y, z) \equiv \lambda + \eta$. If the magnitude of any demand jump equals in distribution some nonnegative random variable Z, let $P_Z(\cdot \mid y, z)$ be the probability measure of the random vector equal in distribution to $(y, z - Z) \in E$. Then

$$q_o(K \mid y, z) = \frac{\lambda}{\lambda + \eta} P_Z(K \mid y, z) + \frac{\eta}{\lambda + \eta} I\{\Delta \in K\}, \quad (y, z) \in E_o,$$

and

$$q_\delta(K \mid y, z) = I\{(0, z + L) \in K\}, \quad (y, z) \in E_\delta,$$

for any $K \in \mathcal{E}$, the Borel subsets of E. Then q_0 and q_δ are continuous on E_o and E_δ, respectively. We also have $B(y, z, a) < \infty$ in Γ_1 and $B(y, z, a) = \infty$ in Γ_2, which

are both closed sets. Assumptions (B.1)–(B.4) of Theorem 2.1 are then satisfied. Furthermore, $B(y,z,a) = a/c$ on Γ_1 so that assumption (B.5) of Theorem 2.1 is satisfied. Since assumption (A.1) also holds, we may conclude that the transition measure Q defined in (2.2) is continuous on Γ.

For state (y,z) and decision a we can then calculate the immediate stage cost g,

$$g(y,z,a) = \frac{h(z)+c}{\lambda+\eta}(1-\exp(-(\lambda+\eta)(a/c)))$$

for $(y,z,a) \in \Gamma_1$ and

$$g(y,z,a) = \frac{h(z)}{\lambda+\eta} + \frac{c}{\lambda+\eta}(1-\exp(-(\lambda+\eta)(a/c)))$$

for $(y,z,a) \in \Gamma_2$. If $h(z)$ is lower semicontinuous, then so is g on Γ. Then assumptions (C.1)–(C.4) are satisfied so that Theorem 3.1 applies and algorithm (3.2)–(3.3) becomes

$$J_0(y,z) \equiv 0, \tag{4.2}$$

$$J_{k+1}(y,z) = \inf_a \left(g(y,z,a) + \int_E J_k(y',z')Q(dy',dz' \mid y,z,a) \right) \tag{4.3}$$

for all $(y,z) \in E$, $k \geq 0$, where the infimum is taken over $a \in [0, C-\beta-y] \cup \{C-y\}$ if $y \leq C-\beta$, and over the singleton $\{C-y\}$ if $y > C-\beta$. Since at any state (C,z) the process transfers immediately to $(0, z+L)$, we may set $J_k(C,z) = J_k(0,z+L)$ for all $k \geq 1$, $z \in \Re$. Note that $J_k(\Delta) = 0$.

As an example, we will apply this algorithm to a case considered in [7], in which jumps in demand consist of one unit with probability one. The integral in (4.3) becomes

$$\begin{aligned}\int_E J_k(y',z')Q(dy',dz' \mid y,z,a) = &\int_0^{a/c} J_k(y+ct, z-1)\lambda \exp(-(\lambda+\eta)t)\,dt \\ &+ J_k(0, z+L)\exp(-(\lambda+\eta)(a/c))\end{aligned}$$

for $(y,z,a) \in \Gamma_1$, and

$$\begin{aligned}\int_E J_k(y',z')Q(dy',dz' \mid y,z,a) = &\int_0^{a/c} J_k(y+ct, z-1)\lambda \exp(-(\lambda+\eta)t)\,dt \\ &+ J_k(y+a, z-1)\frac{\lambda}{\lambda+\eta}\exp(-(\lambda+\eta)(a/c))\end{aligned}$$

for $(y,z,a) \in \Gamma_2$. Then let

$$J_{k+1}^a(y,z) = g(y,z,a) + \int_E J_k(y',z')Q(dy',dz' \mid y,z,a)$$

for all $(y,z,a) \in \Gamma$, $k \geq 0$. If L is an integer, then we may confine attention to a semigrid on E by constraining z to be an integer. We will discretize the problem by considering only states $\{(Ci/n, z) : i = 0,1,\ldots,n\}$ for some large n. Choose $\beta = C(i^*/n)$ for some positive integer $i^* < n$. Then (4.3) can be calculated for a

given J_k numerically. To reduce the number of calculations necessary we can evaluate the discretized version of (4.3) using backwards recursion by setting

$$J_{k+1}(C-\beta,z) = \min\{J^0_{k+1}(C-\beta,z), J^\beta_{k+1}(C-\beta,z)\},$$
$$J_{k+1}(C(1-i/n),z) = \min\{V_{wait}, V_{go}\}, \quad i = i^*+1,\dots n,$$

where

$$V_{wait} = \frac{h(z)+\lambda J_k(C(1-i/n),z-1)}{\lambda+\eta}, \tag{4.4}$$

$$V_{go} = \left(C\frac{h(z)+c}{nc} + J_{k+1}(C(1-(i-1)/n),z)\right)\exp(-(\lambda+\eta)C/(nc))$$
$$+ \lambda/(\lambda+\eta)J_k(C(1-i/n),z-1)(1-\exp(-(\lambda+\eta)C/(nc))). \tag{4.5}$$

Intuitively, when the process is in state $(C-\beta,z)$ there are only two options available: completing the project or waiting. So we calculate the expected cost for each option and set $J_{k+1}(C-\beta,z)$ to be the smaller value. Then consider state $(C-\beta-C/n,z)$. Again, there are two choices: either proceeding to point $(C-\beta,z)$ or waiting. Then V_{wait} in (4.4) with $i=i^*+1$ represents the expected cost of waiting. If the choice is to proceed, the assumption is that the process reaches state $(C-\beta,z)$ with probability $\exp(-(\lambda+\eta)C/(nc))$ and then assumes the optimal choice there. Otherwise, the process jumps to point $(C-\beta-C/n,z-1)$ or Δ, with probabilities $\lambda/(\lambda+\eta)$ and $\eta/(\lambda+\eta)$, respectively. For this choice V_{go} in (4.5) with $i=i^*+1$ represents the expected cost. Then set $J_{k+1}(C-\beta-C/n,z)$ to be the smaller of these two values. Continue in this manner, decreasing y by C/n, until J_{k+1} is calculated for state $(0,z)$, and then repeat this algorithm for all values of z. Then J_{k+1} is used to calculate J_{k+2} in the same manner, continuing in this way until convergence is achieved.

This algorithm was applied to a set of parameters $L=1$, $C=1$, $c=1$, $\lambda=0.8$, $\eta=0.05$, and $h(z)=1.5|z|$ on the range $10\le z\le 10$ with $n=50$ and $\beta=3/50$. Note that to calculate $J_{k+1}(y,z)$ the values of $J_k(y,z-1)$ and $J_k(0,z+1)$ are required; hence the range over which J_k can be calculated will decrease by one unit in each direction of z with each iteration. In [7] this is dealt with by setting appropriate boundary conditions. We do the same here with the constraint $J_k(y,-10)=250$, $k\ge 0$. This quantity is roughly the expected cost when construction continues indefinitely from state $(0,-10)$. We also assumed that it will be optimal to wait at all states $(y,10)$. These constraints allow the calculation of J_k on the entire range of interest.

It was found that the optimal policy could be expressed by the quantities $w(z)$, $z=-10,-9\dots,10$, where it will be optimal to construct as long as $y<1-w(z)$. The quantities found were $w(z)=0$ for $z=-10,\dots,-1$; $w(z)=1$ for $z=4,\dots,10$; and $w(0)=0.06$, $w(1)=0.06$, $w(2)=0.22$, and $w(3)=0.62$. Note that $0.06=\beta$. The same example calculated in [7] gives $w(1)=0.0158$, $w(2)=0.2225$, and $w(3)=0.6612$. Also in [7], for $z=0$ it was found that the expected cost improved as $w(0)\to 0$, but it was not optimal to set $w(0)=0$. Accordingly, the algorithm proposed here calculated $w(0)=\beta$. Similarly, where $w(1)=0.0158\le\beta$ in [7], $w(1)$ by the above algorithm was found to be β. The other values were the same using both methods. Convergence was achieved by 50 iterations.

It should be noted that the solution techniques used in [7] require some prior assumption about the form of the optimal policy. Two classes of policy are considered: the "invest until complete" (IUC) policy and the "follow realized demand" (FRD) policy. For an IUC policy there is nonincreasing $w(z)\in[0,C]$ such that construction

takes place when $y \geq 1-w(z)$. Essentially, a plant is completed once started under this policy. For an FRD policy there is nondecreasing $w(z) \in [0, C]$ such that construction takes place as long as $y < 1-w(z)$. (The optimal policy calculated in this section is an FRD policy.) Necessary and sufficient optimality conditions are developed separately for each class and are then investigated separately. No such distinction has to be made in the technique presented in this paper.

5. Concluding remarks. The problem of minimum cost piecewise deterministic processes under a broad class of controls was considered with the objective of verifying the existence of an optimal control and with proposing a unified approach to a numerical solution. The approach is fundamentally different from other discussions of this problem in the literature in that the control problem is presented as a discrete time decision process in which a decision consists of the selection of a trajectory segment from a compact space. The BHJ equation plays no role. If the action space is numerically tractable, a straightfoward fixed point algorithm based on a dynamic programming operator can be used to calculate the optimal control.

In the BHJ equation method the velocity field is commonly assumed to be Lipschitz-continuous. This means that a solution to the BHJ equation could also be constructed from a sequence of trajectory segments taken from a suitably defined compact space, making the theory presented here applicable also to control models treated in the literature cited above (although one would need to establish some smoothness conditions on an optimal trajectory as a necessary condition). The solution methodology, however, is more natural for problems in which the trajectory segments are parametrizable in finite dimensions, although the infinite dimension control could be approximated with splines. It is important to note that the discrete time decision process also admits more coarse varieties of control. For example, we may define piecewise linear control policies, which would have the effect of allowing control to be exerted only at regular time intervals.

It is anticipated that further work in this area will result in an expansion of the definition of the action space to include some control over jump rate and cost function. This would make the range of applicable models similar to that of methods based on the BHJ equation. It would also be of some value to allow trajectory time lengths to vary and hence be subject to control. This would significantly expand the classes of admissible control structures. A more complete treatment of, for example, impulse-type controls would then be possible.

REFERENCES

[1] D. Bertsekas and S. E. Shreve, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.
[2] P. Billingsley, *Probability and Measure*, 3rd ed., John Wiley and Sons, New York, 1995.
[3] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley and Sons, New York, 1983.
[4] M. H. A. Davis, *Piecewise-deterministic Markov processes: A general class of non-diffusion stochastic models*, J. R. Stat. Soc., Ser. B Stat. Methodol., 46 (1984), pp. 353–388.
[5] M. H. A. Davis, *Control of piecewise-deterministic processes via discrete time dynamic programming*, in Proceedings of the 3rd Bad Honnef Conference on Stochastic Differential Systems, Lecture Notes in Control and Inform. Sci. 16, Springer-Verlag, Berlin, 1986, pp. 140–150.
[6] M. H. A. Davis, *Markov Models and Optimization*, Monogr. Statist. Appl. Probab. 49, Chapman and Hall, London, 1993.
[7] M. H. A. Davis, M. A. H. Dempster, S. P. Sethi, and D. Vermes, *Optimal capacity expansion under uncertainty*, Adv. Appl. Probab., 19 (1987), pp. 156–176.

[8] M. H. A. Davis and M. Farid, *Piecewise-deterministic processes and viscosity solutions*, in Stochastic Analysis, Control, Optimization and Applications. A Volume in Honor of W. H. Fleming, Systems Control Found. Appl., W. M. McEneaney, G. Yin, and Q. Zhang, eds., Birkhäuser Boston, Boston, 1999, pp. 249–268.

[9] M. A. H. Dempster and J. J. Ye, *Necessary and sufficient optimality conditions for control of piecewise deterministic Markov processes*, Stochastics Stochastics Rep., 40 (1992), pp. 125–145.

[10] H. M. Soner, *Optimal control with a state space constraint* II, SIAM J. Control Optim., 24 (1986), pp. 1110–1122.

[11] D. Vermes, *Optimal control of piecewise deterministic Markov processes*, Stochastics, 14 (1985), pp. 165–208.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 2, pp. 540–556

NONNEGATIVE REALIZATION OF AUTONOMOUS SYSTEMS IN THE BEHAVIORAL APPROACH*

MARIA ELENA VALCHER†

Abstract. Nonnegative linear systems, which have traditionally been investigated within the state-space framework, have been recently introduced and analyzed by means of the behavioral approach. In a couple of recent papers [J. W. Nieuwenhuis, *Linear Algebra Appl.*, 281 (1998), pp. 43–58, M. E. Valcher, *Linear Algebra Appl.*, 319 (2000), pp. 147–162], several general definitions and results about nonnegative behaviors, as well as a complete analysis of nonnegativity property for autonomous behaviors, have been presented. In this contribution, by focusing our interest again on autonomous behaviors, we explore the nonnegative realization problem by deriving an extended set of necessary and sufficient (geometric) conditions for an autonomous behavior to be nonnegative realizable. In the scalar case, in particular, necessary and sufficient conditions for nonnegative realizability, which refer to the set of zeros of any polynomial involved in the kernel description of the behavior, are provided. Finally, a comparison between the nonnegative realizability property, here investigated, and K-realizability, addressed in [H. Maeda and S. Kodama, *IEEE Trans. Circuits, Systems* I *Fund. Theory Appl.*, CAS-281 (1981), pp. 39–47] is carried on.

Key words. autonomous behavior, most powerful unfalsified model (MPUM), nonnegative behavior, state-space realization, proper (polyhedral) cones left invariant by a linear transformation, nonnegative realization

AMS subject classifications. 12D10, 15A18, 15A48, 37N35, 39A10, 52B99, 93A30, 93B07, 93B20, 93B60

PII. S0363012900378206

1. Introduction.

Since the early seventies positive linear systems have attracted the interest of several researchers, both from a mathematical and from an engineering background. In fact, on the one hand, the theory of positive linear systems is deep and elegant and relies on a family of nice results that essentially draw on the celebrated Perron–Frobenius theorem [2]. On the other hand, concrete applications of this seemingly abstract theory arise in various fields, like econometrics, bioengineering, chemistry, and, generally speaking, in every context where the variables involved are intrinsically nonnegative. A fundamental reference for the elementary definitions and results of positive system theory, as well as a good source of examples where positive system modeling mostly finds interesting applications, is [13]. A very good and recent reference is [6].

A few years ago, there was a first attempt to develop a general theory of a positive linear system within the behavioral framework [17]. In a very nice paper [15], Nieuwenhuis has introduced the notion of nonnegative discrete behavior (whose trajectories are defined on the time axis $\mathbb{Z}_+$), based on the notion of most powerful unfalsified behavior [9, 24], and later given some preliminary results, mostly concerned with behaviors which are one-dimensional (namely, with trajectories in $(\mathbb{R})^{\mathbb{Z}_+}$) and autonomous, or two-dimensional (with trajectories in $(\mathbb{R}^2)^{\mathbb{Z}_+}$) and controllable. More recently, these definitions and results have stimulated a special interest in autonomous behaviors [20], thus leading to a rather complete characterization of nonnegative au-

*Received by the editors September 18, 2000; accepted for publication (in revised form) March 13, 2001; published electronically July 25, 2001.
http://www.siam.org/journals/sicon/40-2/37820.html
†Dipartimento di Ingegneria dell'Innovazione, Università di Lecce, strada per Monteroni, 73100 Lecce, Italy (elena.valcher@unile.it).

tonomous behaviors, mostly based on geometric tools, like invariant cones, and on some new entities, as the nonnegative part of a behavior.

In this contribution, we aim to further extend our analysis of nonnegative autonomous behaviors by addressing what undoubtedly has been, and still is, the most challenging problem in positive system theory: the nonnegative realization problem [1, 4, 14]. In fact, the question we aim to answer is: given a nonnegative autonomous behavior $\mathfrak{B} \subseteq (\mathbb{R}^q)^{\mathbb{Z}_+}$, under what conditions does there exist a nonnegative (autonomous) state-space model

$$\begin{aligned} \mathbf{x}(t+1) &= A\mathbf{x}(t), \\ \mathbf{w}(t) &= C\mathbf{x}(t), \qquad t \in \mathbb{Z}_+, \end{aligned}$$

with A and C nonnegative matrices, such that $\mathfrak{B}$ can be characterized as the set of all trajectories $\mathbf{w}$ which are obtained from the above model, corresponding to the set of all possible initial conditions $\mathbf{x}(0)$?

As we will see, the above question finds quite an exhaustive answer. In fact, after having obtained a rather wide set of equivalent geometric conditions for the problem solvability, by resorting to some key results [1, 4, 11] about the nonnegative realizability of strictly proper transfer functions, we will derive, in the scalar case, a complete spectral characterization of those autonomous behaviors which are nonnegative realizable.

The paper is organized as follows: section 2 summarizes up the basic definitions and results about (linear, left shift-invariant, complete) behaviors whose trajectories are defined on $\mathbb{Z}_+$ and take values in $\mathbb{R}^q$. Also, the fundamental definitions required to introduce positive behaviors are recalled. In section 3, the nonnegativity property for autonomous systems, in the behavioral approach, is recalled, and necessary and sufficient conditions for an autonomous behavior to be nonnegative are presented [20]. Some of these conditions are stated in slightly different terms with respect to [20] in order to obtain more suitable statements for the following analysis. Moreover, some new conditions have been introduced. Finally, a necessary spectral condition, derived in [20] for the nonnegativity of a scalar autonomous behavior, is here strengthened, and proved to be also sufficient.

Section 4 analyzes the nonnegative realization problem for autonomous behaviors and, in section 5, a complete spectral characterization of scalar autonomous behaviors which are nonnegative realizable is given. To conclude the paper, a quick comparison with the notion of nonnegative realizability (K-realizability) previously introduced for scalar autonomous behaviors in [14] is carried on.

Throughout the paper we let $\mathbb{R}^n_+$ denote the nonnegative orthant, namely, the set of nonnegative vectors in the n-dimensional Euclidean space $\mathbb{R}^n$. A set $\mathcal{K} \subset \mathbb{R}^n$ is said to be a *cone* if all finite nonnegative linear combinations of elements of $\mathcal{K}$ belong to $\mathcal{K}$. A cone $\mathcal{K}$ is *convex* if it contains, with any two points, the line segment between them, namely, $\alpha\mathbf{v}_1 + (1-\alpha)\mathbf{v}_2 \in \mathcal{K}$, for every $\alpha \in [0,1]$ and every pair of vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ in $\mathcal{K}$. A convex cone $\mathcal{K}$ is *solid* if it contains an open set (a ball) of $\mathbb{R}^n$, and it is *pointed* if $\mathcal{K} \cap \{-\mathcal{K}\} = \{0\}$. A closed, pointed, solid convex cone is called a *proper cone*. A cone $\mathcal{K}$ is said to be *polyhedral* if it can be expressed as the set of nonnegative linear combinations of a finite set of *generating vectors*. This amounts to saying that a positive integer r and r vectors in $\mathbb{R}^n$, $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$, can be found, such that $\mathcal{K}$ coincides with the set of nonnegative combinations of $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$. In this case, we adopt the notation $\mathcal{K} := \mathrm{Cone}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r)$. This notation can be extended to the case when the vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$ are replaced by matrices $M_1, M_2, \ldots, M_r$ (with

the same number of rows). In this case, by $\text{Cone}(M_1, M_2, \ldots, M_r)$ we mean the set of nonnegative combinations of the columns of $M_1, M_2, \ldots, M_r$. Also, the extensions of the previous definitions to an arbitrary real vector space, on which it has been introduced some topology, are straightforward. The *dual* of a cone $\mathcal{K}$ in $\mathbb{R}^n$ is denoted by $\mathcal{K}^*$ and is defined as

$$\mathcal{K}^* := \{\mathbf{v} : \mathbf{w}^T\mathbf{v} \geq 0 \ \forall \ \mathbf{w} \in \mathcal{K}\}.$$

For further details we refer to [2].

If A is an $n \times n$ real matrix, we denote by $\sigma(A)$ its *spectrum* and by $\rho(A)$ its *spectral radius*, i.e., $\rho(A) := \max\{|\lambda| : \lambda \in \sigma(A)\}$. For every $\lambda \in \sigma(A)$, the *degree* of λ in A, $\deg \lambda$, is the size of the largest diagonal block in the Jordan canonical form of A which contains λ (i.e., the multiplicity of λ as a zero of $\psi_A(z)$, the (monic) minimal polynomial of A).

Given $A \in \mathbb{R}^{n\times n}$ and a cone $\mathcal{K} \subseteq \mathbb{R}^n$, we say that A *leaves* $\mathcal{K}$ *invariant* ($\mathcal{K}$ is A-invariant) if $A\mathcal{K} \subseteq \mathcal{K}$. If $A = [a_{ij}]$ is a matrix (in particular, a vector), we write

- $A \geq 0$ (A *nonnegative*) if $a_{ij} \geq 0$ for all i, j;
- $A > 0$ (A *nonzero nonnegative*) if $a_{ij} \geq 0$ for all i, j, and $a_{hk} > 0$ for at least one pair (h, k);
- $A \gg 0$ (A *positive*) if $a_{ij} > 0$ for all i, j.

Given any polynomial $r(z) \in \mathbb{R}[z]$, we denote by λ_R the greatest (if any) nonnegative real zero of r, namely, $\lambda_R := \max\{\lambda \in \mathbb{R}_+ : r(\lambda) = 0\}$. We say that λ_R is

- *dominant* if for any other zero of r, λ, we have $|\lambda| \leq \lambda_R$ and the multiplicity of λ is not greater than the multiplicity of λ_R as a zero of r;
- *strictly dominant* if for any other zero of r, $\lambda \neq \lambda_R$, we have $|\lambda| < \lambda_R$.

In the paper, all (discrete) sequences will be defined on the set $\mathbb{Z}_+$ of nonnegative integers. The right (forward) and the left (backward) shift operators on $(\mathbb{R}^q)^{\mathbb{Z}_+}$, the set of sequences defined on $\mathbb{Z}_+$ and taking values in $\mathbb{R}^q$, are defined as

$$\tau : (\mathbb{R}^q)^{\mathbb{Z}_+} \to (\mathbb{R}^q)^{\mathbb{Z}_+} : (\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2, \cdots) \mapsto (0, \mathbf{v}_0, \mathbf{v}_1, \cdots),$$
$$\sigma : (\mathbb{R}^q)^{\mathbb{Z}_+} \to (\mathbb{R}^q)^{\mathbb{Z}_+} : (\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2, \cdots) \mapsto (\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \cdots).$$

As we will deal with sets of sequences (our *behaviors*) which are left shift-invariant, we can restrict our attention to the left shift operator σ. For every positive integer i, the ith power of σ is naturally defined by composition as $\sigma^i = \sigma\circ\sigma\circ\cdots\circ\sigma$ (i times).

Also, we can further extend the set of shift operators. Indeed, to every polynomial matrix $R(z) = \sum_{i=0}^{L} R_i z^i \in \mathbb{R}[z]^{p\times q}$ we can associate the polynomial matrix operator $R(\sigma) = \sum_{i=0}^{L} R_i\sigma^i$ (from $(\mathbb{R}^q)^{\mathbb{Z}_+}$ to $(\mathbb{R}^p)^{\mathbb{Z}_+}$), mapping every sequence $\{\mathbf{w}(t)\}_{t\in\mathbb{Z}_+}$ into the sequence $\{R(\sigma)\mathbf{w}(t)\}_{t\in\mathbb{Z}_+}$, where $R(\sigma)\mathbf{w}(t) = R_0\mathbf{w}(t) + R_1\mathbf{w}(t+1) + \cdots + R_L\mathbf{w}(t+L)$ for every $t \in \mathbb{Z}_+$. It can be proved that $R(\sigma)$ describes an injective map if and only if R is a right prime matrix, and a surjective map if and only if R is of full row rank.

2. Identifiability issues and nonnegativity property for a complete behavior. Before proceeding, we briefly summarize some basic definitions and results about behaviors whose trajectories have support in $\mathbb{Z}_+$. Further details on the subject can be found in [15, 18, 22].

In this paper, by a *dynamic system* we mean a triple $\Sigma = (\mathbb{Z}_+, \mathbb{R}^q, \mathfrak{B})$, where $\mathbb{Z}_+$ represents the *time set*, $\mathbb{R}^q$ is the *signal alphabet*, namely, the set where the system trajectories take values, and $\mathfrak{B} \subseteq (\mathbb{R}^q)^{\mathbb{Z}_+}$ is the *behavior*, namely, the set of trajectories

which are compatible with the system laws. A behavior $\mathfrak{B} \subseteq (\mathbb{R}^q)^{\mathbb{Z}_+}$ is said to be *linear* if it is a vector subspace (over $\mathbb{R}$) of $(\mathbb{R}^q)^{\mathbb{Z}_+}$ and *left shift-invariant* if $\sigma\mathfrak{B} \subseteq \mathfrak{B}$. A linear left shift-invariant behavior $\mathfrak{B} \subseteq (\mathbb{R}^q)^{\mathbb{Z}_+}$ is *complete* if for every sequence $\tilde{\mathbf{w}} \in (\mathbb{R}^q)^{\mathbb{Z}_+}$ the condition $\tilde{\mathbf{w}}|_{\mathcal{S}} \in \mathfrak{B}|_{\mathcal{S}}$ for every finite set $\mathcal{S} \subset \mathbb{Z}_+$ implies $\tilde{\mathbf{w}} \in \mathfrak{B}$, where $\tilde{\mathbf{w}}|_{\mathcal{S}}$ denotes the restriction to $\mathcal{S}$ of the trajectory $\tilde{\mathbf{w}}$ and $\mathfrak{B}|_{\mathcal{S}}$ the set of all restrictions to $\mathcal{S}$ of behavior trajectories.

Linear left shift-invariant complete behaviors are kernels of polynomial matrices in the left shift operator σ, which amounts to saying that the trajectories $\mathbf{w} = \{\mathbf{w}(t)\}_{t\in\mathbb{Z}_+}$ of $\mathfrak{B}$ can be identified with the set of solutions in $(\mathbb{R}^q)^{\mathbb{Z}_+}$ of a system of difference equations

$$R_0\mathbf{w}(t) + R_1\mathbf{w}(t+1) + \cdots + R_L\mathbf{w}(t+L) = 0, \qquad t \in \mathbb{Z}_+, \tag{2.1}$$

with $R_i \in \mathbb{R}^{p\times q}$, and hence described by the equation

$$R(\sigma)\mathbf{w} = 0, \tag{2.2}$$

where $R(z) := \sum_{i=0}^{L} R_i z^i$ belongs to $\mathbb{R}[z]^{p\times q}$. In what follows, a behavior $\mathfrak{B}$ described as in (2.2) will be denoted, for short, as $\mathfrak{B} = \ker(R(\sigma))$. Also, we will restrict our attention to linear, left shift-invariant, and complete behaviors $\mathfrak{B} \subseteq (\mathbb{R}^q)^{\mathbb{Z}_+}$, and we will refer to them simply as *behaviors*.

DEFINITION 2.1. *A behavior $\mathfrak{B} \subseteq (\mathbb{R}^q)^{\mathbb{Z}_+}$ is said to be* autonomous *if there exists $m \in \mathbb{Z}_+$ such that if $\mathbf{w}^1, \mathbf{w}^2 \in \mathfrak{B}$, and $\mathbf{w}^1|_{[0,m]} = \mathbf{w}^2|_{[0,m]}$, then $\mathbf{w}^1 = \mathbf{w}^2$.*

As is well known, a behavior $\mathfrak{B} = \ker(R(\sigma))$ with $R \in \mathbb{R}[z]^{p\times q}$ is autonomous if and only if it is a finite-dimensional vector subspace of $(\mathbb{R}^q)^{\mathbb{Z}_+}$, or, equivalently, if and only if R has full column rank q [22]. Any autonomous behavior can always be described as the kernel of a nonsingular square matrix, which is uniquely determined up to a left unimodular factor. So, in what follows, we will steadily assume that $R(z)$ is nonsingular square. The determinant of $R(z)$ (which is, of course, independent of the specific square representation, except for a multiplicative nonzero constant) is known as the *characteristic polynomial of* $\mathfrak{B}$, and its zeros as the *characteristic values of* $\mathfrak{B}$ [17]. For the sake of simplicity, we will also assume that $\det R(z)$ is a monic polynomial.

We now address certain identifiability issues which are fundamental in order to introduce the notion of positive behavior. Such concepts are only marginally touched upon here. For further details we refer the interested reader to [9, 15].

DEFINITION 2.2. *Let $\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^m$ be m trajectories in $(\mathbb{R}^q)^{\mathbb{Z}_+}$. A behavior $\mathfrak{B} \subseteq (\mathbb{R}^q)^{\mathbb{Z}_+}$ is said to be* the most powerful unfalsified model (MPUM) explaining *$\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^m$, if the following hold.*

- *$\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^m$ belong to $\mathfrak{B}$.*
- *For any other behavior $\bar{\mathfrak{B}}$ having $\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^m$ among its trajectories, we have $\mathfrak{B} \subseteq \bar{\mathfrak{B}}$.*

For every choice of the trajectories $\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^m$, the MPUM explaining $\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^m$, denoted by $\mathfrak{B}(\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^m)$, exists and represents the smallest (linear left shift-invariant and complete) behavior including $\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^m$.

A behavior $\mathfrak{B} \subseteq (\mathbb{R}^q)^{\mathbb{Z}_+}$ is said to be *identifiable* if there exists a finite number of its trajectories, say, $\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^m$, such that $\mathfrak{B} \equiv \mathfrak{B}(\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^m)$. Under the linearity, left shift-invariance and completeness assumptions we steadily adopt, every behavior is, indeed, identifiable.

By resorting to the notion of identifiability, Nieuwenhuis proposed in [15] the following definition of nonnegative behavior.

DEFINITION 2.3. *A behavior* $\mathfrak{B} \subseteq (\mathbb{R}^q)^{\mathbb{Z}_+}$ *is said to be* nonnegative *if there exist* $m \in \mathbb{N}$ *and nonnegative trajectories* $\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^m$ *such that* $\mathfrak{B} \equiv \mathfrak{B}(\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^m)$.

3. Nonnegative autonomous behaviors. A fundamental result we will resort to in the following analysis is the fact that every autonomous behavior can be "realized" by means of a state-space model [12, 23]. Indeed, if $\mathfrak{B} \subseteq (\mathbb{R}^q)^{\mathbb{Z}_+}$ is an autonomous behavior, then there exist $n \in \mathbb{N}$ and real matrices $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{q\times n}$ such that

$$\mathfrak{B} = \left\{ \mathbf{w} \in (\mathbb{R}^q)^{\mathbb{Z}_+} : \ \exists\, \mathbf{x}(0) \text{ such that } \mathbf{x}(t+1) = A\mathbf{x}(t),\ \mathbf{w}(t) = C\mathbf{x}(t),\ t \in \mathbb{Z}_+ \right\}.$$

The pair (A, C) is an *n-dimensional realization of* $\mathfrak{B}$. Those realizations of $\mathfrak{B}$ for which n is minimal are called *minimal*. Minimal realizations of an autonomous behavior $\mathfrak{B}$ are those realizations of $\mathfrak{B}$ which are observable [20, 23].

The correspondence between kernel descriptions and state-space representations of an autonomous behavior has been explored in [12], where a certain number of algorithms have also been presented. A major role in this contribution is played by the relationship between the spectral properties of these two representations, namely, between the characteristic polynomial of $\mathfrak{B}$ and the characteristic polynomial of any matrix A appearing in the state space descriptions of $\mathfrak{B}$. Such a relationship is described in the following lemma.

LEMMA 3.1. *Let* $\mathfrak{B} = \ker(R(\sigma))$, *with* $R(z) \in \mathbb{R}[z]^{q\times q}$ *nonsingular square, be an autonomous behavior, and let* (A, C) *be an n-dimensional realization of* $\mathfrak{B}$. *Then the following hold.*

- $\det R \mid \det(zI_n - A)$, *and therefore* $\{\lambda : \det R(\lambda) = 0\} \subseteq \sigma(A)$.
- *If* (A, C) *is a minimal realization of* $\mathfrak{B}$, *then* $\det R$ *and* $\det(zI_n - A)$ *coincide, and hence* $\{\lambda : \det R(\lambda) = 0\} = \sigma(A)$.

Proof. As is well known [7, 12], if (A, C) is an n-dimensional realization of $\mathfrak{B}$ and $[\, V(z) \quad \bar{R}(z)\,]$ is a minimal left annihilator of the Popov–Belevich–Hautus (PBH) observability matrix (also known as the matrix of the Hautus observability test [10]) $\left[\begin{smallmatrix} zI_n - A \\ C \end{smallmatrix}\right]$, then (1) $\bar{R}$ is a nonsingular square matrix with $\det \bar{R} \mid \det(zI_n - A)$, and (2) $\mathfrak{B} = \ker(\bar{R}(\sigma))$. Since R and $\bar{R}$ provide two nonsingular square kernel representations of the same behavior, it follows that $\bar{R} = UR$ for some unimodular factor U. This proves the first part of the result. On the other hand, if (A, C) is a minimal, and hence observable realization, then $\left[\begin{smallmatrix} zI_n - A \\ C \end{smallmatrix}\right]$ is right prime, and hence $\det \bar{R}$ and $\det(zI_n - A)$ coincide. □

Of course, if λ is an eigenvalue of A which is not a characteristic value of $\mathfrak{B}$ (i.e., is not a zero of $\det R$), it must be an eigenvalue of the unobservable system alone [10].

Now we need to introduce a few important sets that will turn out to be useful in providing a complete characterization of both nonnegativity and, later, of nonnegative realizability. To this end, we need to consider a general (not necessarily autonomous) n-dimensional state space model with m inputs and p outputs:

$$\mathbf{x}(t+1) = A\mathbf{x}(t) + B\mathbf{u}(t), \tag{3.1}$$

$$\mathbf{w}(t) = C\mathbf{x}(t), \qquad t \in \mathbb{Z}_+. \tag{3.2}$$

Such a system is denoted, for the sake of brevity, by means of the triple (A, B, C).

DEFINITION 3.2 (see [3, 16]). *Given an n-dimensional state space model* (A, B, C) *with m inputs and p outputs, the* reachable cone *of the system (or, equivalently, of*

the pair (A, B)*) is the cone*

$$\mathcal{R}(A,B) := \overline{\text{Cone}}\{B, AB, A^2B, \ldots\}, \tag{3.3}$$

where the "upper bar" over the word "Cone" denotes the closure operation, while the observable cone *of the system (of the pair* (A, C)*) is the cone*

$$\begin{aligned} \mathcal{S}(C,A) &:= \{\mathbf{x} \in \mathbb{R}^n : CA^t\mathbf{x} \geq 0 \ \forall\ t \geq 0\} \\ &= \left(\text{Cone}\{C^T, A^TC^T, (A^T)^2C^T, \ldots\}\right)^*, \end{aligned} \tag{3.4}$$

where the symbol $*$ *denotes the "dual" of the indicated cone.*

Of course, both the reachable cone and the observable cone are always convex closed and, in general, infinitely generated. The reachable cone is solid if and only if the pair (A, B) is reachable (in the standard sense, i.e., $\text{rank}\,[\,B \quad AB \quad \ldots \quad A^{n-1}B\,] = n$); meanwhile, the observable cone is pointed if and only if the pair (A, C) is observable. Finally, the observable cone of the system (A, B, C) and the reachable cone of the dual system (A^T, C^T, B^T) are dual cones. All these results have been proved in [16] for the continuous time case, but their discrete time version is immediate.

Finally, we introduce the "positive part" of a behavior.

DEFINITION 3.3 (see [20]). *Given a behavior* $\mathfrak{B} \subseteq (\mathbb{R}^q)^{\mathbb{Z}_+}$*, we call the* positive part of $\mathfrak{B}$ *and denote it by* $\mathfrak{B}_+$*, the set of all nonnegative trajectories in* $\mathfrak{B}$*, namely,*

$$\mathfrak{B}_+ := \mathfrak{B} \cap (\mathbb{R}^q_+)^{\mathbb{Z}_+}. \tag{3.5}$$

Given any behavior $\mathfrak{B} \subseteq (\mathbb{R}^q)^{\mathbb{Z}_+}$, its positive part, $\mathfrak{B}_+$, is a convex and pointed cone in $(\mathbb{R}^q)^{\mathbb{Z}_+}$, and it is closed (with respect to the topology of the pointwise convergence). Also, $\mathfrak{B}_+$ is left shift-invariant, meaning that $\mathbf{w} \in \mathfrak{B}_+$ implies $\sigma\mathbf{w} \in \mathfrak{B}_+$.

By referring to a minimal realization of $\mathfrak{B}$, we can provide an efficient characterization of the nonnegativity property for autonomous behaviors, which refers to the aforementioned entities: the reachable cone, the observable cone, and the positive part of a behavior. The equivalence of most of the statements has been proved in [20]. There are some new statements, but their equivalence to at least one of the known statements is immediate.

THEOREM 3.4 (see [20]). *Let* $\mathfrak{B} \subseteq (\mathbb{R}^q)^{\mathbb{Z}_+}$ *be an autonomous behavior, let* $\mathfrak{B}_+$ *be its positive part, and let* (A, C) *be an* n*-dimensional and minimal realization of* $\mathfrak{B}$*. The following facts are equivalent:*

(1) $\mathfrak{B}$ *is a nonnegative behavior;*
(2) *there exists a positive integer* m *and some matrix* $X_0 \in \mathbb{R}^{n\times m}$ *such that*
  (2a) (A, X_0) *is a reachable pair, and*
  (2b) $CA^tX_0 \geq 0$ *for every* $t \geq 0$*;*
(3) *there exists a positive integer* m *and some matrix* $B \in \mathbb{R}^{n\times m}$ *such that*
  (3a) (A, B, C) *is a minimal realization of its transfer matrix* $W(z) := C(zI_n - A)^{-1}B$*, and*
  (3b) *the Markov coefficients of* $W(z)$*, i.e., the coefficients* W_t *of the power series expansion* $\sum_{t\geq 0} W_t\ z^{-t}$ *of* $W(z)$*, are all nonnegative matrices;*
(4) *there exists a positive integer* m *and some matrix* $B \in \mathbb{R}^{n\times m}$ *such that the reachable cone* $\mathcal{R}(A, B)$ *is proper and included in the observable cone* $\mathcal{S}(C, A)$*;*
(5) *there exists a proper* A*-invariant cone* $\mathcal{K} \subset \mathbb{R}^n$ *included in* $\mathcal{S}(C, A)$*;*
(6) *the observable cone* $\mathcal{S}(C, A)$ *is a proper cone;*
(7) $\mathfrak{B}$ *is the smallest (linear, left shift-invariant, and complete) behavior having* $\mathfrak{B}_+$ *as its positive part;*

(8) $\mathfrak{B}_+$ *generates an* n*-dimensional real vector space in* $(\mathbb{R}^q)^{\mathbb{Z}_+}$ *(equivalently,* $\mathfrak{B}_+$ *is a proper left shift-invariant cone in* $(\mathbb{R}^q)^{\mathbb{Z}_+}$*).*

Remarks. For a more exhaustive discussion on the previous set of characterizations, we refer the interested reader to [20]. Here we aim to introduce only a few specific comments which are relevant for the following analysis.

(i) It is well known that an autonomous behavior $\mathfrak{B} \subseteq (\mathbb{R}^q)^{\mathbb{Z}_+}$ is a finite-dimensional vector subspace of $(\mathbb{R}^q)^{\mathbb{Z}_+}$, whose dimension coincides with the dimension of a minimal realization of $\mathfrak{B}$. In fact, it is easily seen that, under the minimality (namely, observability) assumption on the pair (A, C), there exists a bijective correspondence between the set of behavior trajectories and the (vector) space $\mathbb{R}^n$ of initial conditions $\mathbf{x}(0)$. In particular, there exists a bijective correspondence between the positive part of a behavior $\mathfrak{B}$, $\mathfrak{B}_+$, and the observable cone $\mathcal{S}(C, A)$, which is just the set of initial conditions corresponding (by means of (A, C)) to the trajectories of $\mathfrak{B}_+$. Thus the above theorem states that the nonnegativity of $\mathfrak{B}$ corresponds to the fact that such a cone $\mathcal{S}(C, A)$ is a solid one, or, in a sense, is "dense" in $\mathbb{R}^n$, just as the nonnegativity of $\mathfrak{B}$ means that the set $\mathfrak{B}_+$ is rich enough to carry on all the information about $\mathfrak{B}$. In order to better understand this fact, the interested reader may refer to [20], where a few examples have been given.

(ii) Notice that for the special case of scalar autonomous behaviors, nonnegativity reduces (see Theorem 12 in [15]) to the possibility of determining one single nonnegative trajectory $w \in \mathfrak{B}$ such that $\mathfrak{B} = \mathfrak{B}(w)$. As a consequence, in the scalar case, the matrix X_0 (equivalently, the matrix B), if it exists, can always be assumed to be a simple column vector.

(iii) By Theorem 3.4, an autonomous behavior $\mathfrak{B}$, having (A, C) as an n-dimensional and minimal realization, is nonnegative if and only if there exists a proper A-invariant cone $\mathcal{K}$ included in the observable cone $\mathcal{S}(C, A)$. As is well known [2], an $n \times n$ matrix A leaves a proper cone invariant if and only if it satisfies the following two conditions (known as the *Perron–Schaefer conditions*):

(a) the spectral radius of A, $\rho(A)$, is an eigenvalue of A;

(b) any other eigenvalue λ of A whose modulus $|\lambda|$ is equal to $\rho(A)$ satisfies the inequality $\deg \lambda \leq \deg \rho(A)$ (and hence the size of the largest Jordan block corresponding to λ in the Jordan form of A is not bigger than the largest Jordan block corresponding to $\rho(A)$).

This amounts to saying that $\rho(A)$ is the greatest nonnegative real zero of $\psi_A(z)$, the minimal polynomial of A, and is dominant as a zero of ψ_A. (If A is cyclic, then $\psi_A(z) = \det(zI_n - A)$ and hence the same properties hold true in terms of the characteristic polynomial of A.)

This important characterization represents a necessary condition for an autonomous behavior to be nonnegative. In the scalar case, by resorting to different arguments, we can prove that this condition also becomes sufficient. Indeed, we have also seen that an autonomous behavior $\mathfrak{B}$, having (A, C) as a minimal (and n-dimensional) realization, is nonnegative if and only if the observable cone $\mathcal{S}(C, A)$ is proper. When $\mathfrak{B}$ is scalar autonomous, and hence C is a row vector, we can resort to (the discrete time version of) a result due to Ohta, Maeda, and Kodama (Theorem 3 in [16]) stating that, under the observability assumption on the pair (A, C), the cone $\mathcal{S}(C, A)$ is proper if and only if the $n \times n$ matrix A satisfies the Perron–Schaefer conditions. This important characterization allows us to obtain a complete spectral characterization of nonnegative scalar autonomous behaviors (which strengthens both the general result for the nonscalar case and Proposition 4.2 in [20]). It turns out that the nonnegativity

property for a scalar autonomous behavior $\mathfrak{B}$ depends only on the properties of the maximum modulus characteristic values of $\mathfrak{B}$.

PROPOSITION 3.5. *Let* $\mathfrak{B} = \ker(r(\sigma))$, *with* $r(z) = z^n + r_{n-1}z^{n-1} + \cdots + r_0 \in \mathbb{R}[z]$, *be a scalar autonomous behavior.* $\mathfrak{B}$ *is a nonnegative behavior if and only if the following two conditions hold true:*

(i) $r(z)$ *has a positive real root* λ_R *whose modulus is greater than or equal to the modulus of any other root of* $r(z)$, *namely,* $\lambda_R \geq |\lambda|$ *for any other* λ *such that* $r(\lambda) = 0$;

(ii) *any root* λ *of* $r(z)$, *with* $|\lambda| = \lambda_R$, *has multiplicity* $\mu(\lambda)$ *not greater than the multiplicity* $\mu(\lambda_R)$ *of* λ_R.

Proof. Let (A, C) be a minimal (and hence observable) realization of $\mathfrak{B}$. Since $\mathfrak{B}$ is scalar, then, by resorting to the aforementioned result by Ohta, Maeda, and Kodama, we get that $\mathfrak{B}$ is a nonnegative behavior if and only if A satisfies the Perron–Schaefer conditions. But since A is cyclic (i.e., $\psi_A(z) = \det(zI - A)$) and $r(z) = \det(zI_n - A)$, the Perron–Schaefer conditions are equivalent to (i) and (ii). □

4. Nonnegative realizability and equivalent characterizations. By following [15], we adopt the following, quite natural, definition of nonnegative realizable autonomous behavior.

DEFINITION 4.1 (see [15]). *An autonomous behavior* $\mathfrak{B} \subseteq (\mathbb{R}^q)^{\mathbb{Z}_+}$ *is said to be* nonnegative realizable *if there exists a realization of* $\mathfrak{B}$, *say,* (A_+, C_+), *with* A_+ *and* C_+ *nonnegative matrices.*

Nonnegative realizability admits a wide set of equivalent characterizations that strictly remind us of those obtained for nonnegativity and make use of the same distinguished sets: the reachable cone, the observable cone, and the positive part of the behavior.

THEOREM 4.2. *Let* $\mathfrak{B} \subseteq (\mathbb{R}^q)^{\mathbb{Z}_+}$ *be an autonomous behavior, let* $\mathfrak{B}_+$ *be its positive part, and let* (A, C) *be an* n*-dimensional and minimal realization of* $\mathfrak{B}$. *The following facts are equivalent:*

(1) $\mathfrak{B}$ *is a nonnegative realizable behavior;*

(2) *there exists a positive integer* m *and some matrix* $X_0 \in \mathbb{R}^{n \times m}$ *such that*
  (2a) *the reachable cone of the pair* (A, X_0) *is proper and polyhedral, and*
  (2b) $CA^t X_0 \geq 0$ *for every* $t \geq 0$;

(3) *there exists a positive integer* m *and some matrix* $B \in \mathbb{R}^{n \times m}$ *such that*
  (3a) (A, B, C) *is a minimal realization of its transfer matrix* $W(z) := C(zI_n - A)^{-1}B$,
  (3b) *the Markov coefficients of* $W(z)$, *i.e., the coefficients* W_t *of the power series expansion* $\sum_{t \geq 0} W_t\, z^{-t}$ *of* $W(z)$, *are all nonnegative matrices, and*
  (3c) *the reachable cone of the pair* (A, B) *is proper polyhedral;*

(4) *there exists a positive integer* m *and some matrix* $B \in \mathbb{R}^{n \times m}$ *such that the reachable cone of the pair* $\mathcal{R}(A, B)$ *is proper polyhedral and included in* $\mathcal{S}(C, A)$;

(5) *there exists an* A*-invariant proper polyhedral cone* $\mathcal{K} \subset \mathbb{R}^n$ *included in* $\mathcal{S}(C, A)$;

(6) *the set* $\mathfrak{B}_+$ *includes a proper polyhedral left shift-invariant cone.*

Proof. (1) ⇒ (2). Let (A_+, C_+) be an m-dimensional nonnegative realization of $\mathfrak{B}$, and let T be a nonsingular square matrix that reduces the pair to the standard nonobservable form [10]:

$$T^{-1}A_+T = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix}, \qquad C_+T = [\, C_1 \quad 0 \,] \tag{4.1}$$

with (A_{11}, C_1) an observable pair. The pair (A_{11}, C_1) provides an observable and hence minimal (see comments at the beginning of section 3) realization of $\mathfrak{B}$. So, it entails no loss of generality, assuming $A_{11} = A$ and $C_1 = C$. Partition the nonsingular matrix T^{-1} as

$$T^{-1} = \begin{bmatrix} T_1 \\ T_2 \end{bmatrix} \begin{matrix} \}n, \\ \}m-n. \end{matrix}$$

We aim to show that by choosing $X_0 = T_1$ we satisfy both (2a) and (2b). Of course, the reachable cone generated by the pair (A_+, I_m) coincides with $\mathbb{R}^m_+ = \text{Cone}(I_m)$ and hence is a proper polyhedral cone. Consequently, the reachable cone generated by the pair $(T^{-1}A_+T, T^{-1})$ coincides with $\text{Cone}(T^{-1})$ and is, in turn, a proper polyhedral cone. Since

$$\begin{aligned} \text{Cone}\{T^{-1}, T^{-1}A_+, T^{-1}A_+^2, \ldots\} &= \text{Cone}\left\{T^{-1}, \begin{bmatrix} A & 0 \\ A_{21} & A_{22} \end{bmatrix} T^{-1}, \begin{bmatrix} A & 0 \\ A_{21} & A_{22} \end{bmatrix}^2 T^{-1}, \ldots\right\} \\ &= \text{Cone}\left\{\begin{bmatrix} T_1 \\ T_2 \end{bmatrix}, \begin{bmatrix} A & 0 \\ A_{21} & A_{22} \end{bmatrix}\begin{bmatrix} T_1 \\ T_2 \end{bmatrix}, \begin{bmatrix} A^2 & 0 \\ * & A_{22}^2 \end{bmatrix}\begin{bmatrix} T_1 \\ T_2 \end{bmatrix}, \cdots\right\}, \end{aligned}$$

where $*$ denotes an unspecified element, it follows that

$$\text{Cone}\left\{\ T_1, AT_1, A^2T_1, \ldots\right\}$$

is polyhedral, too. But this is just the reachability cone of the pair (A, T_1), and hence (2a) holds. This first part of the proof has been inspired by the proof of Theorem 2.1 in [1]. Finally, it remains to show that (2b) holds. By the nonnegativity of the pair (A_+, C_+) we have, for every $t \geq 0$,

$$0 \leq C_+A_+^t = (C_+T)(T^{-1}A_+T)^tT^{-1} = [\,C \quad 0\,]\begin{bmatrix} A & 0 \\ A_{21} & A_{22} \end{bmatrix}^t \begin{bmatrix} T_1 \\ T_2 \end{bmatrix}.$$

This ensures that $CA^tX_0 = CA^tT_1 \geq 0$ for every $t \geq 0$.

(2) $\Leftrightarrow$ (3). As (A, C) is a minimal realization for the autonomous behavior $\mathfrak{B}$, and hence is an observable pair, the equivalence of (2) and (3) is straightforward.

(2) $\Rightarrow$ (4) $\Rightarrow$ (5). This part of the proof is obvious.

(5) $\Leftrightarrow$ (6). We have remarked earlier that $\mathfrak{B}_+$ coincides with the set of behavior trajectories generated by the state-space model

$$\mathbf{x}(t+1) = A\mathbf{x}(t), \qquad \mathbf{w}(t) = C\mathbf{x}(t), \qquad t \in \mathbb{Z}_+,$$

corresponding to initial conditions, $\mathbf{x}(0)$, belonging to $\mathcal{S}(C, A)$. Due to the minimality of the state-space representation, there exists a bijective correspondence between trajectories of $\mathfrak{B}_+$ and initial conditions in $\mathcal{S}(C, A)$. This ensures that $\mathfrak{B}_+$ includes the proper polyhedral cone generated by the m trajectories $\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^m$ if and only if $\mathcal{S}(C, A)$ includes the proper polyhedral cone generated by $\mathbf{x}_0^1, \mathbf{x}_0^2, \ldots, \mathbf{x}_0^m$, the initial conditions corresponding to $\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^m$. Finally, the A-invariance of the cone in $\mathcal{S}(C, A)$ is, of course, the corresponding property of the left shift-invariance of the cone in $\mathfrak{B}_+$.

(5) $\Rightarrow$ (1). The proof is very similar to (one part of) the proof of Theorem 2.2 in [1]. Let P be a (full row rank) matrix with n rows, such that $\mathcal{K} = \text{Cone}(P)$. The A-invariance of $\mathcal{K}$ ensures the existence of some nonnegative matrix A_+ such that $AP = PA_+$. Set $C_+ := CP$. Condition $\mathcal{K} \subseteq \mathcal{S}(C, A)$ ensures, in particular, that

$CP \geq 0$, and hence C_+ is nonnegative. We aim to show that the pair (A_+, C_+) provides a nonnegative realization of $\mathfrak{B}$. In fact, by exploiting the fact that P is of full row rank, and hence the fact that for every vector $\mathbf{x}_0$ there exists some $\bar{\mathbf{x}}_0$ such that $\mathbf{x}_0 = P\bar{\mathbf{x}}_0$, we get

$$\begin{aligned} \mathbf{w} \in \mathfrak{B} &\Leftrightarrow \exists\ \mathbf{x}_0 \text{ such that } \mathbf{w}(t) = CA^t\mathbf{x}_0\ \forall t \in \mathbb{Z}_+ \\ &\Leftrightarrow \exists\ \bar{\mathbf{x}}_0 \text{ such that } \mathbf{w}(t) = CA^tP\bar{\mathbf{x}}_0 = CPA_+^t\bar{\mathbf{x}}_0 = C_+A_+^t\bar{\mathbf{x}}_0\ \forall t \in \mathbb{Z}_+. \end{aligned}$$

This completes the proof. □

Remarks. (i) It is immediately seen from the above theorem that every nonnegative realizable behavior is nonnegative. Simple examples can be given showing that the converse is not true in general (see Example 1, below).

(ii) Conditions (4) and (5) strictly remind us of analogous characterizations obtained for the nonnegative realizability of a strictly proper rational transfer function [1, 14]. In fact, if $w(z)$ is a strictly proper rational transfer function and (A, B, C) is an n-dimensional and minimal realization of $w(z)$, then $w(z)$ admits a nonnegative realization if and only if there exists an A-invariant proper polyhedral cone $\mathcal{K}$ satisfying

$$\mathcal{R}(A, B) \subseteq \mathcal{K} \subseteq \mathcal{S}(C, A).$$

(iii) Notice that the proof of (5) ⇒ (1) does not make explicit use of the minimality assumption on the realization (A, C). In fact, it holds also for an arbitrary realization of $\mathfrak{B}$. As a matter of fact, (1) ⇒ (5) also could be proved for an arbitrary realization by suitably adapting the proof of Theorem 1 in [11]. So, the equivalence of points (1) and (5) holds true for any realization (A, C).

(iv) It is immediately apparent from the proofs of (1) ⇒ (2) and (5) ⇒ (1) that when an autonomous behavior is nonnegative realizable, then the size of its minimal nonnegative realization coincides with the minimal number of extremal edges a proper polyhedral cone $\mathcal{K} \subseteq \mathbb{R}^n$ satisfying (5) possibly exhibits. (Notice that a proper polyhedral cone is, in particular, solid, and hence the number of its extremal edges is at least n.) A more straightforward proof could be obtained by suitably adapting that of Theorem 3 in [14].

(v) If $\mathfrak{B}$ is nonnegative realizable, then, by point (5) of the previous theorem, the matrix A appearing in a minimal realization of $\mathfrak{B}$ leaves a proper polyhedral cone $\mathcal{K}$ invariant. As a consequence [19, 21], A satisfies the following three conditions:

(a) the spectral radius of A, $\rho(A)$, is an eigenvalue of A, and, when $\rho(A) \neq 0$,

(b) $\lambda \in \sigma(A)$ with $|\lambda| = \rho(A)$ implies $\deg \lambda \leq \deg \rho(A) =: m$; namely, the size of the largest Jordan block corresponding to λ in the Jordan form of A is not bigger than the largest Jordan block corresponding to $\rho(A)$;

(c) $\lambda \in \sigma(A)$ with $|\lambda| = \rho(A)$ implies that $\lambda/\rho(A)$ is a root of unity.

As it happened for the nonnegativity property, we will see that, in the scalar case, this set of necessary conditions turns out to be sufficient also and can be tested directly on the kernel representation of the behavior.

Example 1. Consider the scalar autonomous behavior $\mathfrak{B} = \ker(r(\sigma))$, where $r(z) = (z-1)(z-e^{i\theta})(z-e^{-i\theta})$ and θ/π is not rational. Of course, by Proposition 3.5, $\mathfrak{B}$ is nonnegative. A minimal realization of $\mathfrak{B}$ is given by the pair

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & -(2\cos\theta + 1) & 2\cos\theta + 1 \end{bmatrix}, \qquad C = [\,1 \quad 0 \quad 0\,].$$

So, by the previous remark, $\mathfrak{B}$ cannot be nonnegative realizable.

(vi) In Theorem 3.4, the nonnegativity property proves to be equivalent to the fact that $\mathcal{S}(C,A)$ is a proper cone or, alternatively, to the fact that $\mathfrak{B}_+$ is a proper cone in $(\mathbb{R}^q)^{\mathbb{Z}_+}$. In this respect, nonnegative realizability entails slightly weaker constraints on the sets $\mathcal{S}(C,A)$ and $\mathfrak{B}_+$ with respect to what we would expect. Indeed, it is equivalent to the fact that $\mathcal{S}(C,A)$ *includes* a proper polyhedral cone or, alternatively, to the fact that $\mathfrak{B}_+$ *includes* a proper polyhedral cone (in $(\mathbb{R}^q)^{\mathbb{Z}_+}$). As a matter of fact, we have the following result.

COROLLARY 4.3. *Given an autonomous behavior $\mathfrak{B} \subseteq (\mathbb{R}^q)^{\mathbb{Z}_+}$, let $\mathfrak{B}_+$ be its positive part, and let (A,C) be an n-dimensional and minimal realization of $\mathfrak{B}$. The following facts are equivalent:*

(i) *$\mathfrak{B}_+$ generates a proper polyhedral cone in $(\mathbb{R}^q_+)^{\mathbb{Z}_+}$;*

(ii) *the observable cone $\mathcal{S}(C,A)$ is a proper polyhedral cone.*

If either of the above equivalent conditions is satisfied, then $\mathfrak{B}$ is a nonnegative realizable behavior.

Proof. (i) $\Leftrightarrow$ (ii). See (5) $\Leftrightarrow$ (6) in the previous theorem. (Notice that, as previously remarked, $\mathcal{S}(C,A)$ is A-invariant, while $\mathfrak{B}_+$ is left shift-invariant.)

The final result follows from the fact that point (5) of Theorem 4.2 holds for $\mathcal{K} = \mathcal{S}(C,A)$ (equivalently, (6) of Theorem 4.2 holds ...), and hence $\mathfrak{B}$ is positive realizable. □

Notice that, in the scalar case, necessary and sufficient conditions for the reachable cone to be proper polyhedral have been given in [5]. So, by resorting to the duality relation existing between the reachable cone of a system and the observable cone of the dual system, we can translate the results of Theorems 1 and 2 in [5], thus obtaining a set of conditions on the matrix A for the properness and polyhedrality of the cone $\mathcal{S}(C,A)$. This way, we get a set of sufficient spectral conditions for a scalar autonomous behavior to be nonnegative realizable. Necessary *and* sufficient spectral conditions for a scalar autonomous behavior to be nonnegative realizable will be derived in the next section.

Example 2 (example on p. 373 in [5]). Consider the scalar autonomous behavior $\mathfrak{B} = \ker(r(\sigma))$, where $r(z) = (z-\lambda_1)(z-\lambda_2)(z-\lambda_3)$, and λ_1, λ_2, and λ_3 are distinct real zeros. Suppose also that λ_1 is positive and that the three zeros are ordered by decreasing order, so that $\lambda_1 > \lambda_2 > \lambda_3$. A minimal realization of $\mathfrak{B}$ is given by the pair

$$A = \begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \lambda_3 \end{bmatrix}, \qquad C = [\,1 \quad 1 \quad 1\,].$$

We consider the same cases analyzed in [5] and make use of their results as far as the polyhedrality of the interested cones is concerned:

(a) $\lambda_1 > 0$, while $\lambda_2, \lambda_3 < 0$, and $\lambda_3 \neq -\lambda_1$. Then, as stated in the example,

$$\text{Cone}\{C^T, (A^T)C^T, (A^T)^2C^T, \ldots\}$$

is proper polyhedral and hence, a fortiori,

$$\overline{\text{Cone}}\{C^T, (A^T)C^T, (A^T)^2C^T, \ldots\}$$

and its dual, $\mathcal{S}(C,A)$.

(b) $\lambda_1 > \lambda_2 > \lambda_3 > 0$. In this case, $\mathcal{S}(C,A)$ not polyhedral. However, $\mathfrak{B}$ is trivially nonnegative realizable ((A,C) is, indeed, a nonnegative realization!).

(c) $\lambda_1 > \lambda_2 > 0$, $\lambda_3 < 0$ and $\lambda_3 \neq -\lambda_1$. In this case

$$\text{Cone}\{C^T, (A^T)C^T, (A^T)^2C^T, \ldots\}$$

is not proper polyhedral, while

$$\overline{\text{Cone}}\{C^T, (A^T)C^T, (A^T)^2C^T, \ldots\},$$

and hence also its dual, $\mathcal{S}(C, A)$, are proper polyhedral cones.

So, in all the above situations, $\mathfrak{B}$ is nonnegative realizable.

To conclude the section, it may be interesting to show how the characterization given by Nieuwenhuis in Theorem 15 of [15] for the nonnegative realizability of a scalar autonomous behavior can be easily obtained as an immediate corollary of Theorem 4.2. Indeed, if $\mathfrak{B} = \ker(r(\sigma))$, with $r(z) \in \mathbb{R}[z]$, is a scalar autonomous behavior, it entails no loss of generality assuming that $r(z)$ is monic, and hence can be represented as $r(z) = z^n + r_{n-1}z^{n-1} + \cdots + r_0$. A minimal realization of $\mathfrak{B}$ is given by the (observable) pair

$$A = \begin{bmatrix} 0 & 1 & 0 & & & \\ 0 & 0 & 1 & 0 & & \\ & & & \ddots & \ddots & \\ & & & & \ddots & 0 \\ & & & & & 1 \\ -r_0 & -r_1 & -r_2 & & & -r_{n-1} \end{bmatrix}, \qquad C = [\,1 \quad 0 \quad \ldots \quad 0\,]$$

(with A in companion form [10]). The operator P introduced in [15] and acting on the vectors of $\mathbb{R}^n$ as follows,

$$\begin{aligned} P \;:\; \mathbb{R}^n &\to \mathbb{R}^n \\ : \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} &\mapsto \begin{bmatrix} v_2 \\ v_3 \\ \vdots \\ -\sum_{i=0}^{n-1} r_i v_i \end{bmatrix}, \end{aligned}$$

corresponds, in fact, to the (left) product by the matrix A, namely, $P(\mathbf{v}) = A\mathbf{v}$ for every $\mathbf{v} \in \mathbb{R}^n$. Since $C = [\,1 \quad 0 \quad \ldots \quad 0\,]$, it is easily seen that condition (5) in Theorem 4.2 becomes equivalent to the fact that there exists a proper polyhedral A-invariant cone $\mathcal{K} \subseteq \mathbb{R}^n_+$, or, in other words, a polyhedral cone $\mathcal{K} \subseteq \mathbb{R}^n_+$ such that $P(\mathcal{K}) \subseteq \mathcal{K}$. In fact, if $\mathcal{K} = \text{Cone}(P)$ is an A-invariant proper polyhedral cone in $\mathbb{R}^n$, condition $\mathcal{K} \subseteq \mathcal{S}(C, A)$ implies, in particular, that $CA^tP \geq 0$ for $t = 0, 1, \ldots, n-1$, which, for the specific choice of the pair (A, C), means

$$\begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} P \geq 0.$$

So, $P \geq 0$ and $\mathcal{K} \subseteq \mathbb{R}^n_+$. Conversely, if $\mathcal{K} = \text{Cone}(P)$ is an A-invariant proper polyhedral cone in $\mathbb{R}^n_+$, and hence $P \geq 0$, then A-invariance ensures that $AP = PA_+$ for some $A_+ \geq 0$. Consequently,

$$CA^tP = CPA_+^t = [\,1 \quad 0 \quad \ldots \quad 0\,]\,PA_+^t \geq 0 \qquad \forall\, t \geq 0.$$

This ensures that $\mathcal{K} = \text{Cone}(P) \subseteq \mathcal{S}(C, A)$. This way we have proved the following result.

PROPOSITION 4.4 (see [15]). *Let $\mathfrak{B} = \ker\,(r(\sigma))$, with $r(z) = z^n + r_{n-1}z^{n-1} + \cdots + r_0 \in \mathbb{R}[z]$, be a scalar autonomous behavior, and let P be the operator acting on the vectors of $\mathbb{R}^n$ as follows:*

$$P \;:\; \mathbb{R}^n \;\to\; \mathbb{R}^n$$

$$: \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} \mapsto \begin{bmatrix} v_2 \\ v_3 \\ \vdots \\ -\sum_{i=0}^{n-1} r_i v_i \end{bmatrix}.$$

$\mathfrak{B}$ is a nonnegative realizable behavior if and only if there exists a proper polyhedral cone $\mathcal{K} \subseteq \mathbb{R}^n_+$ such that $P(\mathcal{K}) \subseteq \mathcal{K}$.

5. The nonnegative realizability problem for scalar autonomous behaviors. In the previous section, we have provided geometric conditions which are necessary and sufficient for the nonnegative realizability of an autonomous behavior. By focusing our attention to the scalar case, we aim now to provide a complete spectral characterization of those scalar autonomous behaviors $\mathfrak{B}$ which admit a nonnegative realization. Not unexpectedly, such a characterization is, again, completely based on the properties of the maximum modulus characteristic values of $\mathfrak{B}$.

As a first step, by suitably applying both the results of the previous sections and the fundamental result (Theorem 4.1 and Corollary 4.1) obtained in [1], we can prove that nonnegative scalar autonomous behaviors, having a strictly dominant nonnegative real characteristic value, are nonnegative realizable.

THEOREM 5.1. *Let $\mathfrak{B} = \ker(r(\sigma))$, with $r(z) \in \mathbb{R}[z]$ monic of degree n, be an autonomous nonnegative behavior. If λ_R, the maximum nonnegative real characteristic value of $\mathfrak{B}$, is strictly dominant, then $\mathfrak{B}$ is nonnegative realizable.*

Proof. Let (A, C) be an n-dimensional and minimal realization of $\mathfrak{B}$. Since $\mathfrak{B}$ is nonnegative and scalar, then (see the remarks after Theorem 3.4) there exists some column vector b such that $\mathcal{R}(A, b)$ is a proper cone included in $\mathcal{S}(C, A)$, or, equivalently, there exists b such that (A, b, C) is a minimal realization of the transfer function with nonnegative Markov coefficients, $W(z) := C(zI_n - A)^{-1}b$. Notice that, since (A, b, C) is minimal, the poles of $W(z)$ coincide with the eigenvalues of A and hence (see Lemma 3.1) with the zeros of $r(z)$. So we have a strictly proper transfer function, $w(z)$, with nonnegative Markov coefficients and a positive real pole which is strictly dominant. This ensures, by Theorem 4.1 and Corollary 4.1 in [1], that $w(z)$ admits a nonnegative realization or, in other words, that there exists a proper polyhedral A-invariant cone $\mathcal{K}$, satisfying

$$\mathcal{R}(A, b) \subseteq \mathcal{K} \subseteq \mathcal{S}(C, A).$$

This ensures, in particular, that condition (5) of Theorem 4.2 is satisfied, and hence $\mathfrak{B}$ is nonnegative realizable. ∎

Notice that Theorem 5.1 assumes that the given scalar behavior is nonnegative. Such an assumption is by no means restrictive, as we have previously underlined that nonnegativity is a necessary condition for an autonomous behavior to be nonnegative realizable. On the other hand, in the specific scalar case we are addressing, such a property can be tested, due to the spectral characterization given in Proposition 3.5.

We aim now to provide a way for testing whether an arbitrary scalar autonomous nonnegative behavior is nonnegative realizable or not. Our proof has been inspired by [4, 11] and is based on the following technical lemma, which is easily proved along the same lines of the proof of Lemma 2 in [11].

LEMMA 5.2 (see [11]). *Let A and C be an $n\times n$ real matrix and a $q\times n$ real matrix, respectively, and let p be any positive integer. There exists a proper (polyhedral) A-invariant cone included in the observable cone $\mathcal{S}(C,A)$ if and only if there exists a proper (polyhedral) A^p-invariant cone included in the observable cone $\mathcal{S}(C,A^p)$.*

THEOREM 5.3. *Let $\mathfrak{B} = \ker(r(\sigma))$, with $r(z) \in \mathbb{R}[z]$ monic of degree n, be an autonomous nonnegative behavior. $\mathfrak{B}$ is nonnegative realizable if and only if the following conditions hold true:*

(i) *$r(z)$ has a positive real dominant root λ_R, which amounts to saying that $\lambda_R \geq |\lambda|$ for any other λ such that $r(\lambda) = 0$ and any root λ, with $|\lambda| = \lambda_R$, has multiplicity $\mu(\lambda)$ not greater than the multiplicity $\mu(\lambda_R)$ of λ_R;*

(ii) *for any root λ of $r(z)$ with $|\lambda| = \lambda_R$, we have that λ/λ_R is a root of unity.*

Proof. Let (A, C) be an n-dimensional and minimal realization of $\mathfrak{B}$. If $\mathfrak{B}$ is nonnegative realizable, then, as we previously remarked, A leaves a proper polyhedral cone invariant and hence satisfies the following three conditions:

(a) the spectral radius of A, $\rho(A)$, is an eigenvalue of A,

and, when $\rho(A) \neq 0$,

(b) $\lambda \in \sigma(A)$ with $|\lambda| = \rho(A)$ implies $\deg \lambda \leq \deg \rho(A) =: m$, namely, the size of the largest Jordan block corresponding to λ in the Jordan form of A is not bigger than the largest Jordan block corresponding to $\rho(A)$;

(c) $\lambda \in \sigma(A)$ with $|\lambda| = \rho(A)$ implies $\lambda/\rho(A)$ is a root of unity.

Since the given realization is a minimal one (and hence $\det(zI_n - A) = r(z)$) and, being in the scalar case, the matrix A is necessarily cyclic (namely, it exhibits one Jordan block for every eigenvalue), the above conditions on the Jordan form of A can be easily translated in terms of the polynomial $r(z)$, thus getting (i) and (ii).

Conversely, suppose that conditions (i) and (ii) hold, and let p be a positive integer such that for any root λ of $r(z)$ with $|\lambda| = \lambda_R$, λ/λ_R is a pth root of unity. Let $\mathfrak{B}_p$ be the scalar autonomous behavior having the pair (A^p, C) as an (n-dimensional) realization. We make the following observations:

- As $\mathfrak{B}$ is nonnegative, there exists a proper A-invariant cone included in $\mathcal{S}(C,A)$. By the previous lemma, then, there exists a proper A^p-invariant cone included in the observable cone $\mathcal{S}(C,A^p)$, and hence $\mathfrak{B}_p$ is nonnegative.
- By the assumption on p, the matrix A^p has a strictly dominant nonnegative real eigenvalue.
- Since the pair (A, C) is minimal and hence observable, then the pair (A^p, C) is either observable or, if not, $\rho(A)^p$ cannot be the eigenvalue of the unobservable system alone. In fact, by the observability of the pair (A, C) there exists a nonzero vector $\mathbf{v}$ such that $A\mathbf{v} = \rho(A)\mathbf{v}$ while $C\mathbf{v} \neq 0$. But then, the state-space model (A^p, C) corresponding to the initial condition $\mathbf{x}(0) = \mathbf{v}$ provides the (nontrivial) output trajectory

$$\mathbf{w}(t) = C(A^p)^t\mathbf{v} = CA^{pt}\mathbf{v} = \rho(A)^{pt}C\mathbf{v} \ \in \mathfrak{B}_p.$$

This ensures that $\rho(A)^p$ cannot be the eigenvalue of the unobservable system alone, and hence if we let $r_p(z)$ be the monic polynomial providing a kernel description of $\mathfrak{B}_p$, we have that $r_p(z)$ has a zero of maximum modulus in $\rho(A)^p$.

So we have that $\mathfrak{B}_p := \ker(r_p(\sigma))$ is nonnegative, and its characteristic polynomial, $r_p(z)$, has $\rho(A)^p$ as a strictly dominant nonnegative real zero. This ensures, by Theorem 5.1, that $\mathfrak{B}_p$ is nonnegative realizable. So, finally, by putting together Theorem 4.2 and Lemma 5.2, we obtain that $\mathfrak{B}$ is nonnegative realizable too. □

Remark. Notice that, with respect to the traditional nonnegative realization problem, here we have obtained a much more favorable situation. In fact, the solutions presently available of the nonnegative realization problem for proper rational transfer functions [1, 4, 8, 11] assume, as a steady assumption, the nonnegativity of the impulse response of the system. This is, of course, a necessary condition for the problem solution; however, up to now, no algorithm has been obtained to check this condition. The nonnegativity constraint on the impulse response is here replaced by the nonnegativity assumption on the scalar autonomous behavior. This property, however, in the scalar case, has been completely captured in terms of spectral conditions. Indeed, as stated in Proposition 3.5, $\mathfrak{B}$ is a nonnegative behavior if and only if the so-called "extended Perron–Schaefer conditions" hold true. Even more, nonnegative realizability has, in turn, obtained a complete spectral characterization in the scalar case. So, it seems that, at least in the scalar autonomous case, the properties here considered can be practically checked.

As we did for nonnegativity in [20], given a scalar autonomous behavior $\mathfrak{B}$, we can find the largest behavior $\mathfrak{B}^*$ included in $\mathfrak{B}$ which is nonnegative realizable. In fact, assume that $\mathfrak{B} = \ker(r(\sigma))$, with $r(z)$ (monic) in $\mathbb{R}[z]$, is a scalar autonomous behavior. Consider the set $\{\lambda \in \mathbb{R}_+ : r(\lambda) = 0\}$ and, in case it is nonempty, set

$$\lambda_R := \max\{\lambda \in \mathbb{R}_+ : r(\lambda) = 0\},$$

and let $\mu(\lambda)$ denote the multiplicity of λ as a zero of $r(z)$. Set, also,

$$p_1(z) := \prod_{\substack{\lambda: r(\lambda)=0 \\ |\lambda|>\lambda_R}} (z-\lambda)^{\mu(\lambda)},$$

$$p_2(z) := \prod_{\substack{\lambda: r(\lambda)=0, |\lambda|=\lambda_R \\ |\lambda|/\lambda_R \text{ a root of unity, } \mu(\lambda)>\mu(\lambda_R)}} (z-\lambda)^{\mu(\lambda)-\mu(\lambda_R)},$$

$$p_3(z) := \prod_{\substack{\lambda: r(\lambda)=0, |\lambda|=\lambda_R \\ |\lambda|/\lambda_R \text{ not a root of unity}}} (z-\lambda)^{\mu(\lambda)},$$

and correspondingly define

$$r^*(z) := \frac{r(z)}{p_1(z)p_2(z)p_3(z)}.$$

Then $\mathfrak{B}^* := \ker(r^*(\sigma))$ is the largest nonnegative realizable behavior included in $\mathfrak{B}$. For the sake of brevity, we skip the boring details of the proof.

6. Comparisons with the *K*-realizability notion. To conclude the paper, it is worthwhile to further deepen our analysis and make a quick comparison with a set of nice results presented in a celebrated paper by Maeda and Kodama. In [14] a different notion of realizability, called K-realizability, has been analyzed and characterized for scalar autonomous behaviors (as a matter of fact, for the set of solutions of a scalar linear time-invariant and homogeneous difference equation. The

paper was written in 1981, before the publication of Willems' papers). This was the only type of nonnegative realizability addressed in the literature until the paper by Nieuwenhuis [15].

Let $\mathfrak{B}$ be a scalar autonomous behavior, and let $\mathfrak{B}_+$ be its positive part. $\mathfrak{B}$ is said to be *K-realizable* if there exists a nonnegative autonomous state-space model

$$\begin{aligned} \mathbf{x}(t+1) &= A\mathbf{x}(t), \\ w(t) &= C\mathbf{x}(t), \qquad t \in \mathbb{Z}_+ \end{aligned}$$

(which is not necessarily a realization of $\mathfrak{B}$!!!), such that $\mathfrak{B}_+$ coincides with the set of all (nonnegative) trajectories produced by the nonnegative state-space model (A, C), corresponding to *nonnegative initial conditions*. In formulas

$$\mathfrak{B}_+ \equiv \{w \in \mathfrak{B} : \exists\ \mathbf{x}(0) \geq 0 \text{ such that } w(t) = CA^t\mathbf{x}(0)\ \forall\ t \geq 0\}. \tag{6.1}$$

In the general case, a K-realizable behavior is not necessarily nonnegative realizable.

Example 3. Consider the scalar autonomous behavior $\mathfrak{B} = \ker(r(\sigma))$, with $r(z) = (z+2)(z-1)$. Of course, by Proposition 3.5, $\mathfrak{B}$ is not nonnegative, as the maximum modulus zero of r is a negative one, and henceforth is not even nonnegative realizable. The positive part of $\mathfrak{B}_+$, by Theorem 4.3 in [20], coincides with the positive part of $\mathfrak{B}^* := \ker(\sigma - 1)$ and hence is

$$\mathfrak{B}_+ = \Big\{\{w(t)\}_{t\geq 0} = \{c\}_{t\geq 0} : c \geq 0\Big\}.$$

It is immediately apparent that the nonnegative system $A = C = 1$ makes condition (6.1) satisfied, and hence $\mathfrak{B}$ is K-realizable.

Under the nonnegativity assumption, however, K-realizable behaviors are always nonnegative realizable. In fact, upon resorting to the results obtained in [20] and in this paper, we can suitably translate into "behavioral terms" the results of Theorem 2 in [14].

THEOREM 2 (from [14]). *Suppose that $\mathfrak{B}$ is a scalar autonomous behavior and that $\mathfrak{B}_+$ generates an n-dimensional vector space in $\mathbb{R}^{\mathbb{Z}_+}$ (equivalently, by Theorem* 3.4, *$\mathfrak{B}$ is a nonnegative behavior). Let (A, C) be a minimal (and n-dimensional) realization of $\mathfrak{B}$. Then $\mathfrak{B}$ is K-realizable if and only if the observable cone $\mathcal{S}(C, A)$ is proper polyhedral.*

Notice that, in the original statement, $\mathcal{S}(C, A)$ is required only to be polyhedral. However, the nonnegativity assumption on $\mathfrak{B}$ ensures (see, again, Theorem 3.4) that the observable cone is proper. So, by comparing Corollary 4.3 with the previous result, we obtain that under the nonnegativity assumption on the scalar autonomous behavior (which, by the way, is a necessary condition for nonnegative realizability), K-realizability is equivalent to the polyhedrality of $\mathcal{S}(C, A)$ and hence implies nonnegative realizability. The converse, however, as it has been shown in Example 2, is not true.

REFERENCES

[1] B. D. O. ANDERSON, M. DEISTLER, L. FARINA, AND L. BENVENUTI, *Nonnegative realization of a linear system with nonnegative impulse response*, IEEE Trans. Circuits Systems I Fund. Theory Appl., CAS-43 (1996), pp. 134–142.

[2] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

[3] P. G. Coxson and H. Shapiro, *Positive reachability and controllability of positive systems*, Linear Algebra Appl., 94 (1987), pp. 35–53.
[4] L. Farina, *On the existence of a positive realization*, Systems Control Lett., 28 (1996), pp. 219–226.
[5] L. Farina and L. Benvenuti, *Polyhedral reachable set with positive controls*, Math. Control Signals Systems, 10 (1997), pp. 364–380.
[6] L. Farina and S. Rinaldi, *Positive Linear Systems: Theory and Applications*, Pure Appl. Math., Wiley-Interscience, New York, 2000.
[7] G. D. Forney, Jr., *Minimal bases of rational vector spaces, with applications to multivariable linear systems*, SIAM J. Control, 13 (1975), pp. 493–520.
[8] K.-H. Forster and B. Nagy, *Nonnegative realizations of matrix transfer functions*, Linear Algebra Appl., 311 (2000), pp. 107–129.
[9] C. Heij, *Exact modelling and identifiability of linear systems*, Automatica J. IFAC, 28 (1992), pp. 325–344.
[10] T. Kailath, *Linear Systems*, Prentice Hall, Englewood Cliffs, NJ, 1980.
[11] T. Kitano and H. Maeda, *Positive realization of discrete-time systems by geometric approach*, IEEE Trans. Automat. Control, AC-45 (1998), pp. 308–311.
[12] M. Kuijper, *First Order Representations of Linear Systems*, Birkhäuser Boston, Boston, 1994.
[13] D. G. Luenberger, *Introduction to Dynamical Systems*, John Wiley and Sons, New York, 1979.
[14] H. Maeda and S. Kodama, *Positive realization of difference equations*, IEEE Trans. Circuits Systems I Fund. Theory Appl., CAS-28 (1981), pp. 39–47.
[15] J. W. Nieuwenhuis, *When to call a linear system nonnegative*, Linear Algebra Appl., 281 (1998), pp. 43–58.
[16] Y. Ohta, H. Maeda, and S. Kodama, *Reachability, observability, and realizability of continuous-time positive systems*, SIAM J. Control Optim., 22 (1984), pp. 171–80.
[17] J. W. Polderman and J. C. Willems, *Introduction to Mathematical Systems Theory: A Behavioral Approach*, Springer-Verlag, New York, 1997.
[18] J. Rosenthal, J. M. Schumacher, and E. V. York, *On behaviors and convolutional codes*, IEEE Trans. Inform. Theory, IT-42 (1996), pp. 1881–1891.
[19] B. S. Tam and H. Schneider, *On the core of a cone-preserving map*, Trans. Amer. Math. Soc., 343 (1994), pp. 479–524.
[20] M. E. Valcher, *Nonnegative linear systems in the behavioral approach: The autonomous case*, Linear Algebra Appl., 319 (2000), pp. 147–162.
[21] M. E. Valcher and L. Farina, *An algebraic approach to the construction of polyhedral invariant cones*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 453–471.
[22] M. E. Valcher and J. C. Willems, *Dead beat observer synthesis*, Systems Control Lett., 37 (1999), pp. 285–292.
[23] J. C. Willems, *From time series to linear system, part* I: *Finite dimensional linear time invariant systems*, Automatica J. IFAC, 22 (1986), pp. 561–580.
[24] J. C. Willems, *From time series to linear system, part* II: *Exact modelling*, Automatica J. IFAC, 22 (1986), pp. 675–694.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 2, pp. 557–576

STOCHASTIC FREQUENCY CHARACTERISTICS*

HANZHONG WU† AND XUN YU ZHOU‡

Abstract. This paper is concerned with stochastic linear-quadratic (LQ) problems in infinite time horizon with control-dependent diffusions and indefinite cost weighting matrices. The classical approach of frequency domain is employed to tackle the problem, starting with the introduction of a frequency characteristic. The equivalence is established among the unique solvability of the LQ problem, the solvability of the associated stochastic Riccati equation, the coercivity of some bilinear form in the Hilbert space of the admissible controls, and the uniformly positive definiteness of the frequency characteristic matrix.

Key words. indefinite stochastic linear-quadratic control, frequency characteristic, stochastic Riccati equation, bilinear form, stability, solvability

AMS subject classifications. 93E20, 93C80, 93D15

PII. S0363012900373756

1. Introduction.

Linear-quadratic (LQ) control, pioneered by Kalman [9], is one of the most fundamental and useful tools in modern engineering and has developed into a major research field in control theory. In the LQ literature it is typically assumed (for a minimization problem) that the cost function has positive semidefinite weighting matrices for the control and the state. In fact, the positive semidefiniteness of the control cost matrix is *necessary* for the well-posedness of the *deterministic* LQ problem (see, e.g., [21, Chapter 6, Proposition 2.4]).

However, it was found in [4] for the first time that a stochastic LQ problem with an *indefinite* control cost may still be well-posed. This phenomenon, which occurs only when the diffusion term depends on the control, has to do with the deep nature of the uncertainty involved. By and large, in a stochastic environment the uncertainty or risk is *costly*, and this uncertainty/risk cost, which can be evaluated precisely via a stochastic Riccati equation, must be taken into consideration when one exercises the control. This, in turn, gives rise to a meaningful or well-posed LQ problem even when the control cost matrix is indefinite (and, in particular, negative definite).

Starting from [4], there have been extensive research efforts devoted to indefinite stochastic LQ control, and a systematic theory is being established; refer to [1, 2, 5, 6, 11, 20, 21]. In addition, computational algorithms to solve the problem were developed in [1, 20]. On the other hand, the theory provides a nice framework for many applications, especially in finance, as many finance problems can be formulated as stochastic LQ problems which are inherently indefinite [10, 22].

In the research on indefinite LQ problems so far, however, there is one important and deep issue that has not been addressed, i.e., what are the precise relations among the parameters of the problem (namely, the linear coefficient matrices in the dynamics

*Received by the editors June 15, 2000; accepted for publication (in revised form) March 9, 2001; published electronically July 25, 2001.

http://www.siam.org/journals/sicon/40-2/37375.html

†Department of Mathematics, Fudan University, Shanghai 200433, P. R. China (hzwu@fudan.edu.cn). This research was carried out during the author's visit to the Department of System Engineering and Engineering Management, The Chinese University of Hong Kong. The research of this author was supported by the NSF of China 19901030.

‡Department of System Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, Hong Kong (xyzhou@se.cuhk.edu.hk). The research of this author was supported by the RGC earmarked grants CUHK 4124/97E and CUHK 4054/98E.

and the weighting matrices in the cost) so as to make a well-posed and solvable LQ problem? In [4], it was shown that the well-posedness and solvability of an LQ control problem boil down to certain conditions in terms of a Riccati equation, but it is hard to translate the conditions into ones that *directly* relate the parameters of the original LQ problem, due to the difficulty and complexity of the Riccati equation. For example, in [4] the following question was posed: if the control cost is allowed to be negative, then how negative can it be so that the LQ problem is still well-posed and solvable? An answer was given in [4], nevertheless only implicitly via the solution to the Riccati equation, which was not really an "answer" in that it was not easily verifiable. Another possible way of addressing the issue is to reformulate the stochastic LQ problem as a (minimizing) optimization problem of a bilinear form in a Hilbert space—the space of all admissible controls. From this perspective, whether or not the control cost is positive is not important in order for the LQ problem to be well-posed and solvable; what is important is that all the parameters must be given in such a way that the resulting infinite dimensional optimization problem is convex (see, e.g., [21, Chapter 6, Theorem 4.2]). This approach gives a deeper and more abstract view of the original LQ problem, but again the difficulty lies in interpreting the abstract convexity condition by the parameters of the LQ problem.

In the classical control literature there is a so-called frequency domain approach that describes the input-output or control-state relation via the Fourier transformation and uses some frequency characteristics to characterize the underlying LQ problem. For the deterministic finite-dimensional systems, Yakubovich [17] systematically established the frequency theory in both nondegenerate and degenerate forms. These results were later extended to infinite dimensions in [18, 19, 15, 16]. For the stochastic case, Dokuchaev [7] introduced a frequency characteristic for a certain (definite) LQ problem under the solvability of a linear matrix equation. However, in [7] the diffusion coefficient does not depend on the control variable, and the controls are taken as deterministic functions; hence the results and their derivations are very much parallel to the deterministic case. On the other hand, Ugrinovskii [14] established, via the frequency domain approach, the equivalence between the solvability of a matrix inequality and the nonnegativity of the aforementioned bilinear form in the control space. However, no frequency characteristic that can be computed explicitly through the parameter matrices of the LQ problem has been given in [14] to characterize the well-posedness/solvability of the original stochastic LQ problem. (For more details see Remark 7.3 below.)

In this paper, we introduce a new frequency characteristic for the indefinite stochastic LQ problem. This characteristic, which reduces to the one studied in [7] when the diffusion is control-independent, is a complex matrix that can be calculated directly from the given parameters of the LQ problem. Then we establish the equivalence of the following statements: (1) the LQ problem is solvable and has a unique optimal control; (2) the stochastic Riccati equation has a unique solution such that the induced feedback control is stabilizing; (3) the associated optimization problem of the bilinear form is coercive; and (4) the frequency characteristic matrix is uniformly positive definite. This way we completely characterize the solvability of the indefinite LQ problem via the frequency characteristic introduced, and we establish a grand unification of different approaches in dealing with the LQ problem. One of the implications of the above equivalence is that it is the overall coordination of *all* the parameters of the problem, rather than the positive definiteness of individual cost matrices, that is essential to the solvability of the LQ problem. This is the underlying

reason for the seemingly surprising phenomenon that one may solve a stochastic LQ problem even when the cost matrices are indefinite. An interesting by-product is that, by comparing and equating the frequency characteristics, the stochastic LQ problem is shown to be equivalent to a problem without a diffusion term but with different cost weighting matrices. In other words, the diffusion part of the problem can be transferred to the overall cost, which in turn implies that the uncertainty is nothing but a *part* of the cost. One can then clearly figure out how negative a control or state cost can be so as to make a meaningful LQ problem.

The remainder of the paper is organized as follows. In section 2 the indefinite stochastic LQ problem is formulated, and some preliminaries, including the important issue of stabilizability, are presented. In section 3 a frequency characteristic is introduced. Section 4 gives a link between the frequency characteristic and the bilinear form. Section 5 is devoted to the relationship between the stochastic Riccati equation and the LQ problem, while section 6 is devoted to that between the frequency characteristic and the LQ problem. In section 7 all the results obtained are unified to establish a grand equivalence theorem. Finally, section 8 concludes the paper.

2. Problem formulation and preliminaries. Let $(\Omega, \mathcal{F}, \mathcal{P}; \mathcal{F}_t)$ be a given standard filtered probability space with a standard scalar Brownian motion $w(t)$ on $[0, +\infty)$ (with $w(0) = 0$). The Brownian motion is assumed to be one-dimensional only for simplicity; there is no essential difficulty with the multidimensional case. Consider the controlled Itô differential equation

$$\begin{cases} dx(t) = [Ax(t) + Bu(t)]dt + [Cx(t) + Du(t)]dw(t), \\ x(0) = x_0 \in \mathbb{R}^n, \end{cases} \tag{2.1}$$

where A, B, C, and D are real matrices of sizes $n \times n$, $n \times m$, $n \times n$, and $n \times m$, respectively. The associated cost functional is

$$J(x_0; u(\cdot)) = E \int_0^{+\infty} [x(t)^T Q x(t) + 2u(t)^T S x(t) + u(t)^T R u(t)]\, dt, \tag{2.2}$$

where Q and R are real symmetric matrices and S is a real matrix of appropriate sizes.

Throughout this paper, the superscript "T" denotes the transpose of a matrix, while "*" denotes the adjoint of a (complex) matrix (i.e., the complex conjugate of the transpose). For a matrix or vector $X = (x_{ij})$, we define its norm by $|X| = (\sum_{i,j} |x_{ij}|^2)^{\frac{1}{2}}$. Moreover, we denote by $\|\cdot\|_H$ the underlying norm in a Hilbert space H. Set

$$L^2_{\mathcal{F}}(\mathbb{R}^k) \triangleq \begin{cases} \phi(\cdot) : [0, +\infty) \times \Omega \mapsto \mathbb{R}^k \mid \phi(\cdot) \text{ is } \mathcal{F}_t\text{-adapted, measurable,} \\ \text{and } E \displaystyle\int_0^{+\infty} |\phi(t, \omega)|^2\, dt < +\infty, \end{cases} \tag{2.3}$$

which is a Hilbert space with the inner product $E \int_0^{+\infty} \phi(t)^T \psi(t)\, dt$ for $\phi(\cdot)$, $\psi(\cdot) \in L^2_{\mathcal{F}}(\mathbb{R}^k)$. Define $\mathcal{U}(x_0) \subset L^2_{\mathcal{F}}(\mathbb{R}^m)$, the set of admissible controls (at x_0), as the collection of such $u(\cdot) \in L^2_{\mathcal{F}}(\mathbb{R}^m)$ that the corresponding solution $x(\cdot) \equiv x(\cdot; x_0, u(\cdot))$ of (2.1) satisfies $x(\cdot) \in L^2_{\mathcal{F}}(\mathbb{R}^n)$. In this case, $(u(\cdot), x(\cdot))$ is called an admissible pair (at x_0).

The (indefinite) stochastic LQ optimal control problem can be stated as follows.

Problem (LQ). For a given $x_0 \in \mathbb{R}^n$, find $u(\cdot) \in \mathcal{U}(x_0)$ so that the cost functional (2.2) is minimized.

We call the LQ problem under consideration in this paper *indefinite* in the sense that we do not impose any positive/positive semi- definiteness for the cost matrices Q and R. However, as it turns out, being indefinite does not rule out the possibility that the bilinear form of the stochastic LQ problem is "positive definite" so that the problem admits an optimal control (see Theorem 7.1). In fact, this is exactly the distinctive feature of the stochastic problem that motivates this paper to characterize the solvability of an indefinite stochastic LQ problem.

Since now the problem is indefinite, the infimum of the cost functional (2.2) could be negatively infinite. If it holds that

$$\inf_{u(\cdot)\in\mathcal{U}(x_0)} J(x_0; u(\cdot)) > -\infty, \tag{2.4}$$

then we say that Problem (LQ) is well-posed at $x_0 \in \mathbb{R}^n$. If there exists a $\bar{u}(\cdot) \in \mathcal{U}(x_0)$ such that

$$J(x_0; \bar{u}(\cdot)) = \inf_{u(\cdot)\in\mathcal{U}(x_0)} J(x_0; u(\cdot)) > -\infty, \tag{2.5}$$

then we say that Problem (LQ) is solvable at $x_0 \in \mathbb{R}^n$. In this case, we call the control $\bar{u}(\cdot)$ an optimal control and the pair $(\bar{u}(\cdot), \bar{x}(\cdot))$ an optimal pair. If there is only one optimal control satisfying (2.5), then Problem (LQ) is called uniquely solvable at $x_0 \in \mathbb{R}^n$.

Compared with optimality, an almost equally important term, due to the infinite time horizon, is the stability/stabilizability.

Definition 2.1. *The system* (2.1) *with a given admissible control* $u(\cdot)$ *is called (mean-square) stable at a given initial state* $x_0 \in \mathbb{R}^n$, *if* $\lim_{t\to+\infty} E|x(t)|^2 = 0$, *where* $x(\cdot)$ *is the corresponding state trajectory. The system* (2.1) *is called (mean-square) stabilizable if there exists a feedback control* $u(\cdot) = Kx(\cdot)$ *with a constant matrix* K, *such that the corresponding solution* $x(\cdot)$ *of the system* (2.1), *for any initial state* $x_0 \in \mathbb{R}^n$, *satisfies* $\lim_{t\to+\infty} E|x(t)|^2 = 0$. *In this case, the matrix* K *is called a (mean-square) stabilizing feedback operator, and the feedback control* $u(\cdot) = Kx(\cdot)$ *is called a (mean-square) stabilizing control.*

The following standard assumption is imposed throughout this paper.

Assumption (A1). The system (2.1) is stabilizable.

Remark 2.1. It is easy to see that Assumption (A1) implies the stabilizability of the pair (A, B) in the deterministic sense.

Remark 2.2. It follows from Theorems 4.1 and 4.2 of [13] that Assumption (A1) is equivalent to the nonemptiness of the admissible control set $\mathcal{U}(x_0)$ at *any* x_0. This is a very deep result, as the nonemptiness of $\mathcal{U}(x_0)$ is a rather weak statement. It also shows that (A1) is almost a minimum assumption.

Under Assumption (A1), we can further assume, without loss of generality, that the uncontrolled system of (2.1) (i.e., the system (2.1) with $u(t) \equiv 0$) is (mean-square) stable at any initial x_0. Indeed, let K be a stabilizing feedback operator, and put $u(\cdot) = Kx(\cdot) + v(\cdot)$ in (2.1). Then (2.1) is turned to

$$dx(t) = [A_1x(t) + Bv(t)]dt + [C_1x(t) + Dv(t)]dw(t), \quad x(0) = x_0, \tag{2.6}$$

where $A_1 = A + BK$ and $C_1 = C + DK$. So the system (2.6) with $v(t) \equiv 0$ is stable at any x_0.

Based on the above argument, we assume the following throughout this paper.

Assumption (A2). The uncontrolled system of (2.1) is stable at any initial x_0.

Remark 2.3. Similar to Remark 2.1, Assumption (A2) implies that the matrix A is Hurwitz (i.e., the real parts of all the eigenvalues of A are strictly negative).

LEMMA 2.1. *For any* $u(\cdot) \in \mathcal{U}(x_0)$, $E|x(t)|^2$ *is uniformly continuous on* $[0, +\infty)$ *and* $\lim_{t\to+\infty} E|x(t)|^2 = 0$.

Proof. By Itô's formula, $X(t) \triangleq E[x(t)x(t)^T]$ satisfies the differential equation

$$\frac{d}{dt}X(t) = AX(t) + X(t)A^T + CX(t)C^T + Y(t), \tag{2.7}$$

where

$$\begin{aligned} Y(t) \triangleq E[&Bu(t)x(t)^T + x(t)u(t)^T B^T + Du(t)x(t)^T C^T \\ &+ Cx(t)u(t)^T D^T + Du(t)u(t)^T D^T]. \end{aligned} \tag{2.8}$$

It follows from $u(\cdot) \in \mathcal{U}(x_0)$ that, for some constant $k > 0$,

$$\int_0^\infty |Y(t)|\, dt \le k[\|u(\cdot)\|^2_{L^2_{\mathcal{F}}(\mathbb{R}^m)} + \|x(\cdot)\|^2_{L^2_{\mathcal{F}}(\mathbb{R}^n)}] < +\infty. \tag{2.9}$$

On the other hand, by Assumption (A2), the solution to the homogeneous version of (2.7),

$$\frac{d}{dt}X(t) = AX(t) + X(t)A^T + CX(t)C^T, \tag{2.10}$$

is exponentially stable, i.e., there exist numbers $c, \varepsilon > 0$ such that $|X(t)| \le c|x_0|^2 e^{-\varepsilon t}$. Hence it can be easily verified by the variation-of-constant formula along with (2.9) that $E[x(t)x(t)^T] = X(t)$ is uniformly continuous. This also leads to $\lim_{t\to+\infty} E[x(t)x(t)^T] = 0$ since $x(\cdot) \in L^2_{\mathcal{F}}(\mathbb{R}^n)$. □

LEMMA 2.2. $\mathcal{U}(x_0) = L^2_{\mathcal{F}}(\mathbb{R}^m)$ *for any* $x_0 \in \mathbb{R}^n$.

Proof. It suffices to prove that for any $u(\cdot) \in L^2_{\mathcal{F}}(\mathbb{R}^m)$, the corresponding solution $x(\cdot)$ of (2.1) satisfies $x(\cdot) \in L^2_{\mathcal{F}}(\mathbb{R}^n)$. By Itô's formula, $X(t) \triangleq E[x(t)x(t)^T]$ satisfies the matrix-valued differential equation (2.7). It has been shown in the proof of Lemma 2.1 that there exist $c_1, \varepsilon_1 > 0$ such that the fundamental solution matrix (with a size of $n^2 \times n^2$) of the linear ODE (2.10), denoted by $M(t) \triangleq (m_{ij}(t))$, satisfies

$$|M(t)| = \left(\sum_{i,j} |m_{ij}(t)|^2\right)^{\frac{1}{2}} \le c_1 e^{-\varepsilon_1 t} \quad \forall t \in [0, +\infty). \tag{2.11}$$

By Young's inequality there exists a constant $c_2 > 0$ (independent of $t \in [0, +\infty)$) such that $Y(t)$ defined by (2.8) satisfies

$$|Y(t)| \le \frac{\varepsilon_1}{2c_1}|X(t)| + c_2 E|u(t)|^2 \quad \forall t \in [0, +\infty). \tag{2.12}$$

Hence, using the variation-of-constant formula, it follows from (2.11) and (2.12) that

the solution $X(\cdot)$ to (2.7) satisfies

$$\begin{aligned}|X(t)| &\le c_1 e^{-\varepsilon_1 t}|X(0)| + c_1 \int_0^t e^{-\varepsilon_1(t-s)}|Y(s)|\,ds \\ &\le c_1 e^{-\varepsilon_1 t}|X(0)| + c_1 c_2 \int_0^t e^{-\varepsilon_1(t-s)} E|u(s)|^2\,ds \\ &\quad + \frac{\varepsilon_1}{2}\int_0^t e^{-\varepsilon_1(t-s)}|X(s)|\,ds.\end{aligned}$$

Integrating from 0 to $j \in [0, +\infty)$ and employing Fubini's theorem, we obtain

$$\begin{aligned}\int_0^j |X(t)|\,dt &\le \frac{c_1}{\varepsilon_1}|X(0)| + c_1 c_2 \int_0^j \int_0^t e^{-\varepsilon_1(t-s)} E|u(s)|^2\,ds\,dt \\ &\quad + \frac{\varepsilon_1}{2}\int_0^j \int_0^t e^{-\varepsilon_1(t-s)}|X(s)|\,ds\,dt \\ &\le \frac{c_1}{\varepsilon_1}\left[|X(0)| + c_2 \int_0^{+\infty} E|u(t)|^2\,dt\right] + \frac{1}{2}\int_0^j |X(t)|\,dt.\end{aligned}$$

Noticing the arbitrariness of j and the fact that

$$\|x(\cdot)\|^2_{L^2_{\mathcal{F}}(\mathbb{R}^n)} = \int_0^{+\infty} \operatorname{Tr} X(t)\,dt,$$

we have, for some constant $k > 0$,

$$\|x(\cdot)\|_{L^2_{\mathcal{F}}(\mathbb{R}^n)} \le k[|x_0| + \|u(\cdot)\|_{L^2_{\mathcal{F}}(\mathbb{R}^m)}]. \tag{2.13}$$

This completes the proof. □

In view of Lemma 2.2, from now on we shall use $\mathcal{U}(x_0)$ and $L^2_{\mathcal{F}}(\mathbb{R}^m)$ interchangeably.

3. Frequency characteristic. Since A is a Hurwitz matrix (see Remark 2.3), the complex matrix-valued function

$$g(\lambda) \triangleq (i\lambda I - A)^{-1} \tag{3.1}$$

is well defined for any $\lambda \in \mathbb{R}$.

Now introduce the following matrix equation with the unknown Θ being a symmetric matrix:

$$\Theta = Q + \frac{1}{2\pi}\int_{-\infty}^{+\infty} C^T g(-\lambda)^T \Theta g(\lambda) C\,d\lambda. \tag{3.2}$$

This is a linear algebraic equation whose solvability is interesting in its own right. We shall study the solvability of this equation at the end of this section.

Suppose that (3.2) is solvable with a solution Θ. Define a complex matrix-valued function

$$\begin{aligned}\hat{\Phi}(\lambda) \triangleq & R + \frac{1}{2\pi} D^T \int_{-\infty}^{+\infty} g(-\lambda)^T \Theta g(\lambda)\,d\lambda D + B^T g(-\lambda)^T \Theta g(\lambda) B \\ & + B^T g(-\lambda)^T \left[S^T + \frac{1}{2\pi} C^T \int_{-\infty}^{+\infty} g(-\lambda)^T \Theta g(\lambda)\,d\lambda D\right] \\ & + \left[S + \frac{1}{2\pi} D^T \int_{-\infty}^{+\infty} g(-\lambda)^T \Theta g(\lambda)\,d\lambda C\right] g(\lambda) B \quad \forall \lambda \in \mathbb{R}.\end{aligned} \tag{3.3}$$

The above $\hat{\Phi}(\lambda)$ is called the frequency characteristic of Problem (LQ), which can be computed explicitly through the given parameters of the original stochastic LQ problem. Note that when $C = 0$ and $D = 0$, by Proposition 3.1(i) below, this frequency characteristic coincides with that in the deterministic case (see [16],[17],[18],[19]):

$$\hat{\Phi}(\lambda) \triangleq R + B^T g(-\lambda)^T Q g(\lambda) B + B^T g(-\lambda)^T S^T + S g(\lambda) B \quad \forall \lambda \in \mathbb{R}. \tag{3.4}$$

Let us now turn to the solvability of (3.2). First we investigate the scalar case.

Example 3.1. Consider Problem (LQ), where the state and control variables are both scalar. It is easy to verify that the solution $x(\cdot)$ to the uncontrolled system (2.1) satisfies $Ex(t)^2 = e^{(2A+C^2)t}x_0^2$. By the stability assumption (A2), we must have $2A + C^2 < 0$. On the other hand, (3.2) reduces to

$$\Theta = Q + \frac{1}{2\pi}\int_{-\infty}^{+\infty} \frac{C^2}{\lambda^2 + A^2} d\lambda \cdot \Theta = Q - \frac{C^2}{2A}\Theta.$$

This equation has a unique solution

$$\Theta = \frac{Q}{1 + \frac{C^2}{2A}} = \frac{2AQ}{2A + C^2}.$$

It is interesting to note that Θ and Q always have the same signs. Moreover, $\Theta \geq Q$ if and only if $Q \geq 0$.

For the general multidimensional case, we are now giving equivalent algebraic conditions for (3.2) to have solutions and a unique solution, respectively. First, with an $n \times n$ matrix $X = (x_{ij})$ we associate the n^2-dimensional column vector vec X defined by

$$\text{vec } X \triangleq (x_{11}, \dots, x_{n1}, x_{12}, \dots, x_{n2}, \dots, x_{1n}, \dots, x_{nn})^T.$$

Next, recall that for two $n \times n$ matrices $X = (x_{ij})$ and $Y = (y_{ij})$ we define the Kronecker product, denoted by $X \otimes Y$, by

$$X \otimes Y \triangleq \begin{pmatrix} x_{11}Y & \cdots & x_{1n}Y \\ \vdots & \cdots & \vdots \\ x_{n1}Y & \cdots & x_{nn}Y \end{pmatrix}.$$

THEOREM 3.1. *Equation* (3.2) *admits solutions if and only if*

$$\text{rank } (\Gamma \ \text{ vec } Q) = \text{rank } \Gamma,$$

where the $n^2 \times n^2$ *matrix* Γ *is given by*

$$\Gamma \triangleq I - \frac{1}{2\pi}\int_{-\infty}^{+\infty} (C^T g(\lambda)^T) \otimes (C^T g(-\lambda)^T) d\lambda.$$

Moreover, (3.2) *admits a unique solution if and only if* Γ *is nonsingular.*

Proof. Vectorizing both sides of (3.2) and noting the general formula vec $(XYZ) = (Z^T \otimes X)$vec Y (see p. 254, Lemma 4.3.1 of [8]), we have

$$\text{vec } \Theta = \text{vec } Q + \frac{1}{2\pi}\int_{-\infty}^{+\infty} (C^T g(\lambda)^T) \otimes (C^T g(-\lambda)^T) d\lambda \text{ vec } \Theta.$$

The desired results follow immediately. □

The following proposition further gives two easily verifiable *sufficient* conditions for the unique solvability of (3.2).

PROPOSITION 3.1. *Equation* (3.2) *admits a unique solution* Θ *in either of the following two cases.*

(i) $C = 0$ *(in which case* $\Theta = Q$*).*

(ii) A *and* C *satisfy*

$$\frac{1}{2\pi}\int_{-\infty}^{+\infty} |C^T g(-\lambda)^T||g(\lambda)C|\, d\lambda < 1.$$

Proof. Case (i) is straightforward. For (ii), denote by $\mathcal{S}^n$ the space of $n \times n$ symmetric matrices. Define a mapping $\mathcal{T} : \mathcal{S}^n \mapsto \mathcal{S}^n$ as follows:

$$\mathcal{T}\Theta \triangleq Q + \frac{1}{2\pi}\int_{-\infty}^{+\infty} C^T g(-\lambda)^T \Theta g(\lambda) C\, d\lambda \quad \forall \Theta \in \mathcal{S}^n.$$

The conclusion then follows immediately from the contraction mapping theorem. □

It should be noted that the case $C = 0$ is typically encountered in financial applications (see, e.g., [10, 22]).

4. Frequency characteristic and bilinear form. It is well known that an optimal LQ problem can be formulated as an optimization problem of a bilinear form in a Hilbert space. This formulation, in turn, gives insight on the LQ problem from the viewpoint of convex analysis in infinite dimensions. This section establishes the relation between the frequency characteristic and the bilinear form.

Define the bounded bilinear operators Φ and Γ on $L^2_{\mathcal{F}}(\mathbb{R}^m)$ and $\mathbb{R}^n$, respectively, as

$$\begin{aligned}\Phi(u(\cdot), v(\cdot)) \triangleq & E\int_0^{+\infty} [x(t;0,u(\cdot))^T Q x(t;0,v(\cdot)) + u(t)^T S x(t;0,v(\cdot)) \\ & + x(t;0,u(\cdot))^T S^T v(t) + u(t)^T R v(t)]\, dt \quad \forall u(\cdot), v(\cdot) \in \mathcal{U}(0),\end{aligned} \tag{4.1}$$

and

$$\Gamma(x_0) \triangleq E\int_0^{+\infty} x(t;x_0,0)^T Q x(t;x_0,0)\, dt. \tag{4.2}$$

Further, define the bounded bilinear operator Ψ on $L^2_{\mathcal{F}}(\mathbb{R}^m) \times \mathbb{R}^n$ by

$$\Psi(u(\cdot), x_0) \triangleq E\int_0^{+\infty} [x(t;0,u(\cdot))^T Q x(t;x_0,0) + u(t)^T S x(t;x_0,0)]\, dt. \tag{4.3}$$

The boundedness of the operators Φ, Γ, and Ψ follows from (2.13).

Then, the cost functional (2.2) can be represented as

$$J(x_0; u(\cdot)) = \Phi(u(\cdot), u(\cdot)) + 2\Psi(u(\cdot), x_0) + \Gamma(x_0). \tag{4.4}$$

Clearly, the first term on the right-hand side of (4.4), namely, the bilinear operator Φ, is crucial in determining the well-posedness and solvability of the optimization problem. To link the bilinear form introduced above to the frequency characteristic, we need to first recall the Fourier transformation on Hilbert spaces.

Let $Z \triangleq L^2(\Omega, \mathcal{F}, \mathcal{P}; \mathbb{R}^k)$ be the Hilbert space equipped with the inner product $\int_\Omega \phi(\omega)^T \psi(\omega) \mathcal{P}(\, d\omega)$ for any $\phi, \psi \in Z$. For $\phi(\cdot) \in L^2(-\infty, +\infty; Z)$, define its Fourier transformation as

$$\tilde{\phi}(\lambda) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-i\lambda t} \phi(t)\, dt \quad \forall \lambda \in \mathbb{R}. \tag{4.5}$$

If we denote

$$\tilde{\phi}_\alpha(\lambda) = \frac{1}{\sqrt{2\pi}} \int_{-\alpha}^{\alpha} e^{-i\lambda t} \phi(t)\, dt \quad \forall \lambda \in \mathbb{R}, \tag{4.6}$$

then obviously we have

$$\|\tilde{\phi}_\alpha(\cdot) - \tilde{\phi}(\cdot)\|_{L^2(-\infty, +\infty; Z)} \to 0 \tag{4.7}$$

as $\alpha \to +\infty$.

Now we turn to our system (2.1). For any $x(\cdot) \in L^2_{\mathcal{F}}(\mathbb{R}^n)$, we set $x(t) = 0$ for $t < 0$. We augment $u(\cdot) \in \mathcal{U}(0)$ in a similar way. In other words, we have the natural embedding $L^2_{\mathcal{F}}(\mathbb{R}^n) \subset L^2(-\infty, \infty; L^2(\Omega, \mathcal{F}, \mathcal{P}; \mathbb{R}^n))$ and $\mathcal{U}(0) \subset L^2(-\infty, \infty; L^2(\Omega, \mathcal{F}, \mathcal{P}; \mathbb{R}^m))$.

The following result, which represents the bilinear form through the frequency characteristic, plays a central role in this paper.

PROPOSITION 4.1. *Assume that the matrix equation* (3.2) *is solvable with a solution* Θ. *Then*

$$\Phi(u(\cdot), u(\cdot)) = E \int_{-\infty}^{+\infty} \tilde{u}(\lambda)^* \hat{\Phi}(\lambda) \tilde{u}(\lambda)\, d\lambda, \tag{4.8}$$

where Φ *and* $\hat{\Phi}(\lambda)$, $\lambda \in \mathbb{R}$, *are defined by* (4.1) *and* (3.3), *respectively.*

Proof. For $u(\cdot) \in \mathcal{U}(0)$, the system (2.1) with $x_0 = 0$ yields

$$\begin{aligned} &\int_0^\alpha e^{-i\lambda t}[Ax(t) + Bu(t)]\, dt + \int_0^\alpha e^{-i\lambda t}[Cx(t) + Du(t)]\, dw(t) \\ &= \int_0^\alpha e^{-i\lambda t}\, dx(t) = i\lambda \int_0^\alpha e^{-i\lambda t} x(t)\, dt + e^{-i\lambda\alpha} x(\alpha). \end{aligned}$$

It follows that

$$\tilde{x}_\alpha(\lambda) = (i\lambda - A)^{-1} \left[B\tilde{u}_\alpha(\lambda) + f_\alpha(\lambda) - \frac{1}{\sqrt{2\pi}} e^{-i\lambda\alpha} x(\alpha) \right], \tag{4.9}$$

where

$$f_\alpha(\lambda) = \frac{1}{\sqrt{2\pi}} \int_0^\alpha e^{-i\lambda t}[Cx(t) + Du(t)]\, dw(t), \quad g(\lambda) = (i\lambda I - A)^{-1}.$$

Now, for any symmetric matrix H,

$$\begin{aligned} E[\tilde{x}_\alpha(\lambda)^* H \tilde{x}_\alpha(\lambda)] = E \Big\{ &[g(\lambda) B\tilde{u}_\alpha(\lambda)]^* H g(\lambda) B\tilde{u}_\alpha(\lambda) \\ &- \frac{2}{\sqrt{2\pi}} \operatorname{Re}\{[g(\lambda) B\tilde{u}_\alpha(\lambda)]^* H g(\lambda) e^{-i\lambda\alpha} x(\alpha)\} \\ &+ \frac{1}{2\pi} [g(\lambda) e^{-i\lambda\alpha} x(\alpha)]^* H g(\lambda) e^{-i\lambda\alpha} x(\alpha) \\ &+ [g(\lambda) f_\alpha(\lambda)]^* H g(\lambda) f_\alpha(\lambda) \Big\}. \end{aligned} \tag{4.10}$$

Hence, for $\beta > 0$,

$$\begin{aligned}\int_{-\beta}^{\beta} E[\tilde{x}_\alpha(\lambda)^* H \tilde{x}_\alpha(\lambda)]\, d\lambda &= \int_{-\beta}^{\beta} E[\tilde{u}_\alpha(\lambda)^* B^T g(-\lambda)^T H g(\lambda) B \tilde{u}_\alpha(\lambda)]\, d\lambda \\ &\quad + \int_{-\beta}^{\beta} E\{[g(\lambda) f_\alpha(\lambda)]^* H g(\lambda) f_\alpha(\lambda)\}\, d\lambda + \rho(\alpha, \beta),\end{aligned} \tag{4.11}$$

where $\rho(\alpha, \beta)$ includes all the remainder terms. Noting that $\|x(\alpha)\|_{L^2(\Omega, \mathcal{F}, \mathcal{P}; \mathbb{R}^n)} \to 0$ as $\alpha \to +\infty$ due to Lemma 2.1, that $|g(\lambda)|$ is uniformly bounded due to the stability of (2.1), and that $|g(\lambda) B \tilde{u}_\alpha(\lambda)|$ is uniformly bounded due to (4.7), we conclude that $\rho(\alpha, \beta)$ converges to zero uniformly in β as $\alpha \to +\infty$.

By Itô's formula, we have

$$\begin{aligned}&\int_{-\beta}^{\beta} E\{[g(\lambda) f_\alpha(\lambda)]^* H g(\lambda) f_\alpha(\lambda)\}\, d\lambda \\ &= \int_{-\alpha}^{\alpha} E\left\{ [Cx(t) + Du(t)]^T \left[\frac{1}{2\pi} \int_{-\beta}^{\beta} g(-\lambda)^T H g(\lambda)\, d\lambda \right] [Cx(t) + Du(t)] \right\} dt.\end{aligned} \tag{4.12}$$

Letting $\alpha \to +\infty$ and $\beta \to +\infty$ in (4.11) and (4.12) and appealing to Parseval's equality, we obtain

$$\begin{aligned}&E \int_0^{+\infty} x(t)^T H x(t)\, dt = \frac{1}{2\pi} E \int_{-\infty}^{+\infty} \tilde{x}(\lambda)^* H \tilde{x}(\lambda)\, d\lambda \\ &= \frac{1}{2\pi} E \int_{-\infty}^{+\infty} \tilde{u}(\lambda)^* B^T g(-\lambda)^T H g(\lambda) B \tilde{u}(\lambda)\, d\lambda \\ &\quad + \frac{1}{2\pi} E \int_0^{+\infty} [Cx(t) + Du(t)]^T \left(\int_{-\infty}^{+\infty} g(-\lambda)^T H g(\lambda)\, d\lambda \right) [Cx(t) + Du(t)] dt.\end{aligned} \tag{4.13}$$

Similarly, it follows from (4.7), (4.9), and Lemma 2.1 that

$$E \int_0^{+\infty} u(t)^T S x(t)\, dt = \frac{1}{2\pi} E \int_0^{+\infty} \tilde{u}(\lambda)^* S g(\lambda) B \tilde{u}(\lambda)\, d\lambda. \tag{4.14}$$

Finally, Parseval's equality yields

$$E \int_0^{+\infty} u(t)^T R u(t)\, dt = \frac{1}{2\pi} E \int_{-\infty}^{+\infty} \tilde{u}(\lambda)^* R \tilde{u}(\lambda)\, d\lambda. \tag{4.15}$$

Combining (4.13), (4.14), and (4.15) and noting that Θ satisfies (3.2), we arrive at (4.8). □

THEOREM 4.1. *If there exists some constant $\sigma > 0$ such that*

$$\hat{\Phi}(\lambda) \geq \sigma I \quad \forall \lambda \in \mathbb{R}, \tag{4.16}$$

then

$$\Phi(u(\cdot), u(\cdot)) \geq \sigma E \int_0^{+\infty} |u(t)|^2\, dt \qquad \forall u(\cdot) \in L^2_{\mathcal{F}}(\mathbb{R}^m). \tag{4.17}$$

As a consequence, Problem (LQ) *is uniquely solvable at any initial state* $x_0 \in \mathbb{R}^n$.

Proof. It follows from (4.8) that

$$\Phi(u(\cdot),u(\cdot)) \geq \sigma E \int_{-\infty}^{+\infty} \tilde{u}(\lambda)^* \tilde{u}(\lambda)\, d\lambda = \sigma \|u(\cdot)\|^2_{L^2_{\mathcal{F}}(\mathbb{R}^m)}.$$

This coercivity of the bilinear form Φ further implies (see, e.g., [18, Lemma 1] or [21, Chapter 6, Theorem 4.2]) that Problem (LQ) is uniquely solvable at any initial state x_0. □

5. Riccati equation and LQ control. The Riccati equation is the primary approach in dealing with LQ problems. In this section we present the equivalent relation between the unique solvability of Problem (LQ) and the solvability of the following stochastic algebraic Riccati equation (SARE) introduced in [1]:

$$\begin{cases} A^TP + PA + C^TPC + Q \\ \qquad - (S^T + PB + C^TPD)(R + D^TPD)^{-1}(S + B^TP + D^TPC) = 0, \\ R + D^TPD > 0. \end{cases} \tag{5.1}$$

It should be noted that the second (positive definiteness) constraint in (5.1) is *part* of the equation and must be satisfied by any solution. This in turn gives rise to additional difficulty in solving the equation.

To cope with the possible singularity of $R + D^TPD$, we need to employ the notion of the pseudoinverse of a matrix. To be specific, for any matrix M, there exists a unique matrix $M^\dagger$ satisfying the following properties:

$$MM^\dagger M = M,\ \ M^\dagger MM^\dagger = M^\dagger,\ \ (M^\dagger M)^T = M^\dagger M,\ \ (MM^\dagger)^T = MM^\dagger. \tag{5.2}$$

$M^\dagger$ is called the Moore–Penrose pseudoinverse (see [12]) of M. When M is nonsingular, the pseudoinverse coincides with the usual inverse, i.e., $M^\dagger = M^{-1}$.

The generalized version of the SARE (5.1), when the matrix $R+D^TPD$ is allowed to be singular, is the following generalized algebraic Riccati equation (GARE), which was first introduced in [2]:

$$\begin{cases} \mathcal{M}(P) - \mathcal{L}(P)^T \mathcal{N}(P)^\dagger \mathcal{L}(P) = 0, \\ [I - \mathcal{N}(P)^\dagger \mathcal{N}(P)]\mathcal{L}(P) = 0, \\ \mathcal{N}(P) \geq 0, \end{cases} \tag{5.3}$$

where

$$\begin{cases} \mathcal{M}(P) \triangleq A^TP + PA + C^TPC + Q, \\ \mathcal{N}(P) \triangleq R + D^TPD, \\ \mathcal{L}(P) \triangleq S + B^TP + D^TPC. \end{cases} \tag{5.4}$$

LEMMA 5.1. *If Problem* (LQ) *is solvable at any* $x_0 \in \mathbb{R}$, *then the GARE* (5.3) *has a solution* $P = P^T$ *satisfying*

$$\inf_{u(\cdot)\in\mathcal{U}(x_0)} J(x_0;u(\cdot)) = x_0^TPx_0. \tag{5.5}$$

Proof. This is shown in the proof of [2, Theorem 4.1]. □

THEOREM 5.1. *Problem* (LQ) *is uniquely solvable at any initial state* $x_0 \in \mathbb{R}^n$ *if and only if the SARE* (5.1) *has one and only one such solution* $P = P^T$ *that the associated feedback operator*

$$\widetilde{K} \triangleq -(R + D^T P D)^{-1}[S + B^T P + D^T P C] \tag{5.6}$$

is (mean-square) stabilizing.

Proof. The "if" part is an immediate consequence of [2, Theorems 2.1, 4.1]. We now prove the "only if" part. By Lemma 5.1, the GARE (5.3) admits a solution $P = P^T$ satisfying (5.5). Let $u(\cdot) \in \mathcal{U}(x_0)$ be any admissible control, and let $x(\cdot) \in L^2_{\mathcal{F}}(\mathbb{R}^n)$ be the corresponding state trajectory. By applying Itô's formula to $x(t)^T P x(t)$ and integrating from 0 to $+\infty$, and then by combining it with the cost functional (2.2), we obtain

$$\begin{aligned} J(x_0; u(\cdot)) =& x_0^T P x_0 \\ &+ E \int_0^{+\infty} \{x(t)^T \mathcal{M}(P) x(t) + 2u(t)^T \mathcal{L}(P) x(t) + u(t)^T \mathcal{N}(P) u(t)\}\, dt. \end{aligned} \tag{5.7}$$

First, we show that

$$\mathcal{N}(P) \neq 0. \tag{5.8}$$

In fact, if $\mathcal{N}(P) = 0$, then the GARE (5.3) is reduced to

$$\mathcal{M}(P) = 0,\ \mathcal{L}(P) = 0,\ \mathcal{N}(P) = 0.$$

Hence it follows from (5.7) that

$$J(x_0; u(\cdot)) = 0 \qquad \forall u(\cdot) \in L^2_{\mathcal{F}}(\mathbb{R}^m),$$

which implies that any $u(\cdot) \in L^2_{\mathcal{F}}(\mathbb{R}^m)$ is optimal. This contradicts the assumption of the unique solvability of Problem (LQ), which proves (5.8).

Next we prove that $\mathcal{N}(P)$ is nonsingular. Indeed, if $\mathcal{N}(P) \neq 0$ is singular, then there is a symmetric matrix V with $VV^T = V^T V = I$ such that

$$\mathcal{N}(P) = V \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} V^T, \tag{5.9}$$

where Σ is a positive definite diagonal matrix of order r with $0 < r < m$. Moreover,

$$\mathcal{N}(P)^\dagger = V \begin{pmatrix} \Sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} V^T. \tag{5.10}$$

Define $v(\cdot) \triangleq V^T u(\cdot)$. It follows from (5.7) and the GARE (5.3) that

$$\begin{aligned} J(x_0; u(\cdot)) &= x_0^T P x_0 + E \int_0^{+\infty} \left\{ x(t)^T \mathcal{L}(P)^T V \begin{pmatrix} \Sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} V^T \mathcal{L}(P) x(t) \right. \\ &\qquad \left. + 2v(t)^T V^T \mathcal{L}(P) x(t) + v(t)^T \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} v(t) \right\} dt \\ &= x_0^T P x_0 + E \int_0^{+\infty} |(\Sigma^{\frac{1}{2}} \ \ 0)_{r\times m} v(t) + (\Sigma^{-\frac{1}{2}} \ \ 0)_{r\times m} V^T \mathcal{L}(P) x(t)|^2\, dt. \end{aligned} \tag{5.11}$$

Note that in equating the cross term of $v(t)$ and $x(t)$ in the last equality above, we have used the second equality of (5.3) as well as (5.9)–(5.10). In view of (5.5), $u(\cdot) \in L^2_{\mathcal{F}}(\mathbb{R}^m)$ is an optimal control of Problem (LQ) at $x_0 \in \mathbb{R}^n$ if and only if the integrand on the right-hand side of (5.11) is zero, i.e.,

$$u(\cdot) = Vv(\cdot) \equiv V\begin{pmatrix} v_r(\cdot) \\ v_{m-r}(\cdot) \end{pmatrix} = -V\begin{pmatrix} (\Sigma^{-1} \ \ 0)_{r\times m} V^T \mathcal{L}(P)x(\cdot) \\ v_{m-r}(\cdot) \end{pmatrix}, \tag{5.12}$$

where $v_{m-r}(\cdot) \in L^2_{\mathcal{F}}(\mathbb{R}^{m-r})$. With the above control, the dynamics (2.1) is reduced to

$$\begin{cases} dx(t) = [A_1 x(t) + B_1 v_{m-r}(t)]dt + [C_1 x(t) + D_1 v_{m-r}(t)]dw(t), \\ x(0) = x_0 \in \mathbb{R}^n, \end{cases} \tag{5.13}$$

where

$$\begin{cases} A_1 \triangleq A - B\mathcal{N}(P)^\dagger \mathcal{L}(P), \quad C_1 \triangleq C - D\mathcal{N}(P)^\dagger \mathcal{L}(P), \\ B_1 \triangleq -BV\begin{pmatrix} 0_{r\times(m-r)} \\ I_{(m-r)\times(m-r)} \end{pmatrix}, \quad D_1 \triangleq -DV\begin{pmatrix} 0_{r\times(m-r)} \\ I_{(m-r)\times(m-r)} \end{pmatrix}. \end{cases}$$

The above argument shows that any $v_{m-r}(\cdot) \in L^2_{\mathcal{F}}(\mathbb{R}^{m-r})$ applied to the system (5.13), as long as the corresponding state $x(\cdot) \in L^2_{\mathcal{F}}(\mathbb{R}^n)$, provides an optimal control for the original Problem (LQ) via (5.12). Now we are going to show that there are more than one such $v_{m-r}(\cdot)$. First note that at this point we *cannot* directly conclude the stabilizability of system (5.13) from that of system (2.1) because (5.13) is "smaller" than (2.1) in the sense that the admissible controls for (5.13) are confined to the form specified by (5.12). To get around this, let $(\bar{u}(\cdot), \bar{x}(\cdot)) \in L^2_{\mathcal{F}}(\mathbb{R}^m) \times L^2_{\mathcal{F}}(\mathbb{R}^n)$ be an optimal pair for the original Problem (LQ) at an initial x_0, which exists by assumption, and let $\bar{v}(\cdot) \equiv \binom{\bar{v}_r(\cdot)}{\bar{v}_{m-r}(\cdot)} = V^T \bar{u}(\cdot)$. Then it follows from (5.12) that

$$\bar{u}(\cdot) = -V\begin{pmatrix} (\Sigma^{-1} \ \ 0)_{r\times m} V^T \mathcal{L}(P)^T \bar{x}(\cdot) \\ \bar{v}_{m-r}(\cdot) \end{pmatrix},$$

where $\bar{v}_{m-r}(\cdot) \in L^2_{\mathcal{F}}(\mathbb{R}^{m-r})$. This implies that $\bar{x}(\cdot) \in L^2_{\mathcal{F}}(\mathbb{R}^n)$ is the trajectory to (5.13) under the admissible control $\bar{v}_{m-r}(\cdot) \in L^2_{\mathcal{F}}(\mathbb{R}^{m-r})$. Therefore, the set of admissible controls for (5.13) is nonempty at any initial state x_0. By Theorems 4.1 and 4.2 in [13], the SARE

$$\begin{aligned} &A_1^T P + PA_1 + C_1^T PC_1 + I \\ &\quad - (PB_1 + C_1^T PD_1)(I + D_1^T PD_1)^{-1}(B_1^T P + D_1^T PC_1) = 0 \end{aligned}$$

admits a solution $P_1 = P_1^T$ such that

$$\widetilde{K} \triangleq -(I + D_1^T P_1 D_1)^{-1}[B_1^T P_1 + D_1^T P_1 C_1]$$

is a stabilizing feedback operator. In other words, system (5.13) is stabilizable. Using the same argument as that in the proof of Lemma 2.2, we can show that the state $x(\cdot)$ of (5.13) corresponding to the control $v_{m-r}(\cdot) = Kx(\cdot) + \xi(\cdot)$, where $\xi(\cdot) \in L^2_{\mathcal{F}}(\mathbb{R}^{m-r})$ is arbitrary, must satisfy $x(\cdot) \in L^2_{\mathcal{F}}(\mathbb{R}^n)$. This implies that the original Problem (LQ) has more than one optimal control, which is a contradiction. Therefore, $\mathcal{N}(P)$ must be nonsingular. The desired results then follow from the fact that in this case the two Riccati equations, SARE (5.1) and GARE (5.3), coincide, and the optimal feedback operator given in (5.12) degenerates to (5.6). Finally, the uniqueness of such a solution P to (5.1) that (5.6) is stabilizing follows from [2, Theorem 2.3]. □

6. Frequency characteristic and LQ control. In this section, to close the loop of equivalence we establish the relationship between the frequency characteristic and LQ control. Throughout the section we assume that the matrix equation (3.2) is solvable.

PROPOSITION 6.1. *The cost functional* (2.2) *can be represented as*

$$J(x_0;u(\cdot)) = E\int_{-\infty}^{+\infty}\{\tilde{u}(\lambda)^*\hat{\Phi}(\lambda)\tilde{u}(\lambda) + 2\,\mathrm{Re}[\tilde{u}(\lambda)^*\hat{\Psi}(\lambda)x_0]\}\,d\lambda + \Gamma(x_0), \tag{6.1}$$

where

$$\hat{\Psi}(\lambda) \triangleq \left[B^T g(-\lambda)^T\Theta + \frac{1}{2\pi}D^T\int_{-\infty}^{+\infty} g(-\lambda)^T\Theta g(\lambda)\,d\lambda C + S\right] g(\lambda). \tag{6.2}$$

Proof. Similar to the argument in section 4, it follows from (4.7), (4.9), and Lemma 2.1 that

$$\begin{aligned} E\int_0^{+\infty} x(t;0,u(\cdot))^T\Theta x(t;x_0,0)\,dt &= E\int_{-\infty}^{+\infty}\tilde{u}(\lambda)^*B^Tg(-\lambda)^T\Theta g(\lambda)x_0\,d\lambda \\ &+ \frac{1}{2\pi}E\int_0^{+\infty}[Cx(t;0,u(\cdot)) + Du(t)]^T\int_{-\infty}^{+\infty} g(-\lambda)^T\Theta g(\lambda)\,d\lambda Cx(t;x_0,0)\,dt. \end{aligned}$$

Moreover, for any matrix $H \in \mathbb{R}^{m\times n}$,

$$E\int_0^{+\infty} u(t)^THx(t;x_0,0)\,dt = E\int_{-\infty}^{+\infty}\tilde{u}(\lambda)^*Hg(\lambda)x_0\,d\lambda.$$

Therefore,

$$\begin{aligned} &E\int_0^{+\infty} x(t;0,u)^TQx(t;x_0,0)\,dt \\ &= E\int_{-\infty}^{+\infty}\tilde{u}(\lambda)^*\left[B^Tg(-\lambda)^T\Theta + \frac{1}{2\pi}D^T\int_{-\infty}^{+\infty} g(-\lambda)^T\Theta g(\lambda)\,d\lambda C\right] g(\lambda)x_0\,d\lambda. \end{aligned}$$

We can then conclude from (4.3) that

$$\begin{aligned} &\Psi(u(\cdot),x_0) \\ &= E\int_{-\infty}^{+\infty}\tilde{u}(\lambda)^*\left[B^Tg(-\lambda)^T\Theta + \frac{1}{2\pi}D^T\int_{-\infty}^{+\infty} g(-\lambda)^T\Theta g(\lambda)\,d\lambda C + S\right] g(\lambda)x_0\,d\lambda. \end{aligned} \tag{6.3}$$

The desired result therefore follows from (4.1)–(4.4), (4.8), and (6.3). ∎

Next, based on the frequency characteristic, we construct an auxiliary stochastic LQ problem that does not have a diffusion part but is equivalent to the original stochastic LQ problem. To do this, define

$$\begin{aligned} &Q_1 \triangleq \Theta, \qquad S_1 \triangleq S + \frac{1}{2\pi}D^T\int_{-\infty}^{+\infty} g(-\lambda)^T\Theta g(\lambda)\,d\lambda C, \\ &R_1 \triangleq R + \frac{1}{2\pi}D^T\int_{-\infty}^{+\infty} g(-\lambda)^T\Theta g(\lambda)\,d\lambda D. \end{aligned} \tag{6.4}$$

Construct the following LQ problem with the cost functional

$$\bar{J}(x_0;u(\cdot)) = E\int_0^{+\infty} [x(t)^T Q_1 x(t) + 2u(t)^T S_1 x(t) + u(t)^T R_1 u(t)]\,dt, \tag{6.5}$$

subject to the dynamics

$$\begin{cases} \dfrac{d}{dt}x(t) = Ax(t) + Bu(t), \\ x(0) = x_0 \in \mathbb{R}^n, \end{cases} \tag{6.6}$$

where the admissible control set is $L^2_{\mathcal{F}}(\mathbb{R}^m)$.

From the way we constructed the LQ problem (6.5)–(6.6), it is clear that the problem has the same frequency characteristic as that of the original Problem (LQ). By an analogous argument to those in section 4 and the above in this section, it holds that

$$\bar{J}(x_0;u(\cdot)) = E\int_{-\infty}^{+\infty} \{\tilde{u}(\lambda)^*\hat{\Phi}(\lambda)\tilde{u}(\lambda) + 2\,\mathrm{Re}[\tilde{u}(\lambda)^*\hat{\Psi}(\lambda)x_0]\}\,d\lambda + \Gamma_1(x_0), \tag{6.7}$$

where

$$\Gamma_1(x_0) \triangleq E\int_0^{+\infty} x(t;x_0,0)^T Q_1 x(t;x_0,0)\,dt \tag{6.8}$$

for the solution $x(\cdot;x_0,0)$ of (6.6) with $u(\cdot)=0$.

LEMMA 6.1. *The LQ problem* (2.1)–(2.2) *is uniquely solvable at any* x_0 *if and only if the LQ problem* (6.5)–(6.6) *is uniquely solvable at any* x_0.

Proof. In view of (6.1) and (6.7), $J(x_0;u(\cdot))$ and $\bar{J}(x_0;u(\cdot))$ differ by a term which does not depend on the control variable $u(\cdot)$. The result then follows. □

Remark 6.1. Lemma 6.1 is of significant implications on its own. It suggests that the original Problem (LQ) is in fact equivalent to the auxiliary LQ problem (6.5)–(6.6) whose diffusion part in the dynamics is absent while the cost parameters are modified. More specifically, one can transfer the diffusion part (corresponding to the uncertainty/risk) of a stochastic system to the cost part. The new cost can be precisely calculated according to (6.4), which in turn draws the boundary of the possible indefiniteness of the original LQ problem. See Example 7.1 below for more details.

LEMMA 6.2. *If the LQ problem* (6.5)–(6.6) *is uniquely solvable at any initial state* $x_0 \in \mathbb{R}^n$, *then* $R_1 > 0$, *and the Riccati equation*

$$A^T P + PA + Q_1 - [B^T P + S_1]^T R_1^{-1}[B^T P + S_1] = 0 \tag{6.9}$$

admits a solution P_1 *with the matrix*

$$A - BR_1^{-1}[B^T P_1 + S_1] \tag{6.10}$$

being Hurwitz.

Proof. This follows from Theorem 5.1, in view of the fact that the LQ problem (6.5)–(6.6) is a special case of the original Problem (LQ). □

LEMMA 6.3. *Assume that the LQ problem* (6.5)–(6.6) *is uniquely solvable at any initial state* $x_0 \in \mathbb{R}^n$. *Then*

$$\hat{\Phi}(\lambda) = [I + E(\lambda)^*]R_1[I + E(\lambda)], \tag{6.11}$$

where

$$E(\lambda) \triangleq R_1^{-1}[B^T P_1 + S_1](i\lambda - A)^{-1}B. \tag{6.12}$$

Moreover, it holds that

$$I + E(\lambda) = [I - G(\lambda)B]^{-1} \tag{6.13}$$

with

$$G(\lambda) \triangleq R_1^{-1}[B^T P_1 + S_1]\{i\lambda - A + BR_1^{-1}[B^T P_1 + S_1]\}^{-1}. \tag{6.14}$$

Proof. It follows from (3.3) and (6.4) that

$$\begin{aligned}\hat{\Phi}(\lambda) = {} & R_1 + B^T(-i\lambda - A^T)^{-1}Q_1(i\lambda - A)^{-1}B \\ & + B^T(-i\lambda - A^T)^{-1}S_1^T + S_1(i\lambda - A)^{-1}B \quad \forall \lambda \in \mathbb{R}.\end{aligned} \tag{6.15}$$

By Lemma 6.2, the Riccati equation (6.9) admits a unique solution P_1. First we calculate

$$\begin{aligned}& -(-i\lambda - A^T)^{-1}A^T P_1(i\lambda - A)^{-1} - (-i\lambda - A^T)^{-1}P_1 A(i\lambda - A)^{-1} \\ & = P_1(i\lambda - A)^{-1} + i\lambda(-i\lambda - A^T)^{-1}P_1(i\lambda - A)^{-1} \\ & \quad + (-i\lambda - A^T)^{-1}P_1 - i\lambda(-i\lambda - A^T)^{-1}P_1(i\lambda - A)^{-1} \\ & = P_1(i\lambda - A)^{-1} + (-i\lambda - A^T)^{-1}P_1.\end{aligned}$$

Consequently,

$$\begin{aligned}& (-i\lambda - A^T)^{-1}Q_1(i\lambda - A)^{-1} \\ & = (-i\lambda - A^T)^{-1}(B^T P_1 + S_1)^T R_1^{-1}(B^T P_1 + S_1)(i\lambda - A)^{-1} \\ & \quad - (-i\lambda - A^T)^{-1}(A^T P_1 + P_1 A)(i\lambda - A)^{-1} \\ & = (-i\lambda - A^T)^{-1}(B^T P_1 + S_1)^T R_1^{-1}(B^T P_1 + S_1)(i\lambda - A)^{-1} \\ & \quad + P_1(i\lambda - A)^{-1} + (-i\lambda - A^T)^{-1}P_1.\end{aligned} \tag{6.16}$$

Plugging (6.16) in (6.15), we get (6.11).

By Lemma 6.2, the matrix

$$\pi(\lambda) \triangleq i\lambda - A + BR_1^{-1}(B^T P_1 + S_1)$$

is nonsingular for all $\lambda \in \mathbb{R}$, which in turn implies that $G(\lambda)$ in (6.14) is well defined. Now we compute

$$\begin{aligned}& E(\lambda)G(\lambda)B \\ & = R_1^{-1}(B^T P_1 + S_1)(i\lambda - A)^{-1}BR_1^{-1}(B^T P_1 + S_1)\pi(\lambda)^{-1}B \\ & = R_1^{-1}(B^T P_1 + S_1)(i\lambda - A)^{-1}[i\lambda - A + BR_1^{-1}(B^T P_1 + S_1)]\pi(\lambda)^{-1}B \\ & \quad - R_1^{-1}(B^T P_1 + S_1)\pi(\lambda)^{-1}B \\ & = E(\lambda) - G(\lambda)B.\end{aligned}$$

Thus

$$[I + E(\lambda)][I - G(\lambda)B] = I + E(\lambda) - G(\lambda)B - E(\lambda) + G(\lambda)B = I,$$

which yields (6.13). □

THEOREM 6.1. *If Problem* (LQ) *is uniquely solvable at any* x_0*, then there exists some constant* $\sigma > 0$ *such that*

$$\hat{\Phi}(\lambda) \geq \sigma I \quad \forall \lambda \in \mathbb{R}. \tag{6.17}$$

Proof. It follows from (6.10) that there exists a constant $k > 0$ satisfying

$$|[i\lambda - A + BR_1^{-1}(B^T P_1 + S_1)]^{-1}| \leq k \quad \forall \lambda \in \mathbb{R}.$$

Hence there exists a $k_1 > 0$ such that

$$|G(\lambda)| \leq k_1 \quad \forall \lambda \in \mathbb{R}.$$

For any $u \in \mathbb{C}^m$ (the set of m-dimensional complex vectors), let

$$v(\lambda) \triangleq [I + E(\lambda)]u = [I - G(\lambda)B]^{-1}u \quad \forall \lambda \in \mathbb{R}.$$

Then it holds that

$$|u| = |[I - G(\lambda)B]v(\lambda)| \leq (1 + k_1|B|)|v(\lambda)|.$$

Hence it follows from (6.11) that

$$\begin{aligned} u^* \hat{\Phi}(\lambda) u &= |R_1^{\frac{1}{2}}[I + E(\lambda)]u|^2 = |R_1^{\frac{1}{2}} v(\lambda)|^2 \\ &\geq k_2^2 |v(\lambda)|^2 \geq \left(\frac{k_2}{1 + k_1|B|}\right)^2 u^* u \quad \forall \lambda \in \mathbb{R}, \quad \forall u \in \mathbb{C}^m, \end{aligned}$$

where $k_2 > 0$ is the minimum eigenvalue of the positive-definite matrix $R_1^{\frac{1}{2}}$. This completes the proof. □

7. Synthesis. In this section, based on the results in the previous sections, we establish a grand unification of four statements for Problem (LQ) from different aspects: the LQ problem itself, the Riccati equation, the bilinear form, and the frequency characteristic.

THEOREM 7.1. *Assume that the matrix equation* (3.2) *is solvable. Then the following statements are equivalent.*

(I) *Problem* (LQ) *is uniquely solvable at any initial state* $x_0 \in \mathbb{R}^n$.

(II) *The SARE* (5.1) *admits only one such solution* P *such that* $P^T = P$ *and*

$$\widetilde{K} \triangleq -(R + D^T P D)^{-1}[S + B^T P + D^T P C] \tag{7.1}$$

is a (mean-square) stabilizing feedback operator.

(III) *There exists a constant* $\sigma > 0$ *such that*

$$\Phi(u(\cdot), u(\cdot)) \geq \sigma E \int_0^{+\infty} |u(t)|^2 \, dt \qquad \forall u(\cdot) \in L^2_{\mathcal{F}}(\mathbb{R}^m). \tag{7.2}$$

(IV) *There exists a constant* $\sigma > 0$ *such that*

$$\hat{\Phi}(\lambda) \geq \sigma I \quad \forall \lambda \in \mathbb{R}. \tag{7.3}$$

Proof. We have the following loop of implications.

(I) $\Longrightarrow$ (IV): Theorem 6.1; (IV) $\Longrightarrow$ (III) $\Longrightarrow$ (I): Theorem 3.1; (I) $\Longleftrightarrow$ (II): Theorem 5.1. □

Remark 7.1. The assumption of the solvability of the matrix equation (3.2) is not necessary for the equivalence between (I) and (II).

Remark 7.2. When (I) or (II) holds true, the unique optimal feedback control of Problem (LQ) is

$$\bar{u}(t) = -(R + D^T P D)^{-1}[S + B^T P + D^T P C]\bar{x}(t), \quad t \geq 0.$$

Remark 7.3. In [14], it was proved that the bilinear form $\Phi \geq 0$ if and only if there exists a symmetric matrix H such that the following inequality holds:

$$2\,\mathrm{Re}\, x^T H(Ax + Bu) + (Cx + Du)^T H(Cx + Du) + F(x, u) \geq 0 \quad \forall u \in \mathbb{R}^m, x \in \mathbb{R}^n,$$

where $F(x, u) = x^T Q x + 2u^T S x + u^T R u$. It was further mentioned in [14, Remark 2] that Φ can be estimated by a certain frequency-type inequality derived in [3], which is precisely as follows. Let $h_1(t)$ and $h_2(t)$ be functions such that $\int_0^{+\infty} |h_k(t)|^2\, dt < +\infty$ and $h_k(t) = 0$ for $t < 0$ $(k = 1, 2)$. The frequency inequality in [3] is the estimate

$$\begin{aligned}\int_0^{+\infty} E\left[\int_0^t h_1(t-\tau)u(\tau)\,d\tau + \int_0^t h_2(t-\tau)u(\tau)dw(\tau)\right] u(t)\,dt \\ \geq \int_{-\infty}^{+\infty} [\mathrm{Re}\, H_1(i\lambda) - G]E|\hat{u}(i\lambda)|^2\,d\lambda \;\forall u(\cdot) \in L^2_{\mathcal{F}}(\mathbb{R}^m),\end{aligned}$$

where $H_1(i\lambda) \triangleq \int_0^{+\infty} h_1(t)e^{i\lambda t}\,dt$ and $\hat{u}(i\lambda) \triangleq \int_0^{+\infty} u(t)e^{i\lambda t}\,dt$. Clearly, the above inequality is hard to verify via the original parameters of Problem (LQ).

Example 7.1. Continue with Example 3.1, where it has been shown that $\Theta = \frac{2AQ}{2A+C^2}$. We can then calculate the parameters in (6.4) as

$$Q_1 = \Theta = \frac{2AQ}{2A + C^2},$$

$$R_1 = R + \frac{1}{2\pi}\int_{-\infty}^{+\infty} \frac{D^2\Theta}{\lambda^2 + A^2} d\lambda = R - \frac{D^2\Theta}{2A} = R - \frac{D^2 Q}{2A + C^2},$$

$$S_1 = S - \frac{CDQ}{2A + C^2}.$$

The frequency characteristic is

$$\hat{\Phi}(\lambda) = \frac{2A}{\lambda^2 + A^2}\left[\frac{B^2 Q}{2A + C^2} - B\left(S - \frac{CDQ}{2A + C^2}\right)\right] + R - \frac{D^2 Q}{2A + C^2}.$$

Consequently, the underlying LQ problem is uniquely solvable if and only if the parameters satisfy the following:

$$\begin{cases} R - \dfrac{D^2 Q}{2A + C^2} > 0 & \text{if } \dfrac{B^2 Q}{2A + C^2} - B\left(S - \dfrac{CDQ}{2A + C^2}\right) \leq 0, \\ \dfrac{2}{A}\left[\dfrac{B^2 Q}{2A + C^2} - B\left(S - \dfrac{CDQ}{2A + C^2}\right)\right] + R - \dfrac{D^2 Q}{2A + C^2} > 0 \\ \qquad\qquad \text{if } \dfrac{B^2 Q}{2A + C^2} - B\left(S - \dfrac{CDQ}{2A + C^2}\right) > 0. \end{cases}$$

Next let us look at a two-dimensional example.

Example 7.2. Consider a two-dimensional LQ problem with the following data in the system dynamics,

$$A = \begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad C = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad D = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},$$

and with the cost weighting matrices

$$Q = \begin{pmatrix} 4 & 0 \\ 0 & -1 \end{pmatrix}, \quad S = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad R = \begin{pmatrix} -1 & 0 \\ 0 & 3 \end{pmatrix}.$$

Note that both Q and R are indefinite in this example. It is clear that the uncontrolled system (2.1) is mean-square stable. Moreover, $\Theta = Q$, and the corresponding frequency characteristic is

$$\begin{aligned} \hat{\Phi}(\lambda) &= \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix} + \frac{1}{(1+\lambda^2)^2} \begin{pmatrix} 4(1+\lambda^2) & 4(1-i\lambda) \\ 4(1+i\lambda) & 4-(1+\lambda^2) \end{pmatrix} \\ &\geq I \quad \forall \lambda \in \mathbb{R}, \end{aligned}$$

where the last inequality is obtained by directly evaluating the eigenvalues. Therefore, it follows from Theorem 7.1 that this LQ problem is uniquely solvable at any initial state, and the SARE (5.1) admits a unique solution.

8. Concluding remarks. In this paper we applied the frequency domain approach to the indefinite stochastic LQ problems. We introduced a new frequency characteristic and explored its links to the underlying LQ problem, the stochastic Riccati equation, and the bilinear form. It turned out that there are intrinsic equivalence relations among them. The frequency characteristic introduced is expressed explicitly through the parameters of the LQ problem and is easy to compute. More importantly, it gives new insights into the internal structure of an LQ problem and explains the fundamental reason why a stochastic LQ control problem can be indefinite.

Acknowledgment. The authors would like to thank the two anonymous referees for their careful reading and constructive comments that led to an improved version of the paper.

REFERENCES

[1] M. Ait Rami and X. Y. Zhou, *Linear matrix inequalities, Riccati equations, and indefinite stochastic linear-quadratic controls*, IEEE Trans. Automat. Control, 45 (2000), pp. 1131–1143.

[2] M. Ait Rami, X. Y. Zhou, and J. B. Moore, *Well-posedness and attainability of indefinite stochastic linear-quadratic control in infinite time horizon*, Systems Control Lett., 41 (2000), pp. 123–133.

[3] V. A. Brusin, *Global stability and dichotomy of a class of nonlinear systems with random parameters*, Sibirsk. Mat. Zh., 22 (1981), pp. 210–222.

[4] S. Chen, X. Li, and X. Y. Zhou, *Stochastic linear quadratic regulators with indefinite control weight costs*, SIAM J. Control Optim., 36 (1998), pp. 1685–1702.

[5] S. Chen and J. Yong, *Stochastic linear quadratic optimal control problems with random coefficients*, Chinese Ann. Math. Ser. B, 21 (2000), pp. 323–338.

[6] S. Chen and X. Y. Zhou, *Stochastic linear quadratic regulators with indefinite control weight costs.* II, SIAM J. Control Optim., 39 (2000), pp. 1065–1081.

[7] N. G. Dokuchaev, *A frequency criterion for the existence of an optimal control for Itô equation*, Vestnik Leningrad Univ. Math., 16 (1984), pp. 41–47.

[8] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.

[9] R. E. Kalman, *Contribution to the theory of optimal control*, Bol. Soc. Mat. Mexicana (3), 5 (1960), pp. 102–119.

[10] M. Kohlmann and X. Y. Zhou, *Relationship between backward stochastic differential equations and stochastic controls: A linear-quadratic approach*, SIAM J. Control Optim., 38 (2000), pp. 1392–1407.

[11] A. E. B. Lim and X. Y. Zhou, *Stochastic optimal LQR control with integral quadratic constraints and indefinite control weights*, IEEE Trans. Automat. Control, 44 (1999), pp. 359–369.

[12] R. Penrose, *A generalized inverse of matrices*, Proc. Cambridge Philos. Soc., 51 (1955), pp. 406–413.

[13] G. Tessitore, *Some remarks on the Riccati equation arising in an optimal control problem with state- and control-dependent noise*, SIAM J. Control Optim., 30 (1992), pp. 717–744.

[14] V. A. Ugrinovskii, *Stochastic analog of frequency theorem*, Izv. Vyssh. Uchebn. Zaved. Mat., 10 (1987), pp. 37–43.

[15] H. Wu, *Some equivalent conditions for exponential stabilization of linear systems with unbounded control*, Sci. China Ser. E, 42 (1999), pp. 252–259.

[16] H. Wu and X. Li, *An Infinite Horizon Linear–Quadratic Problem with Unbounded Controls in Hilbert Space*, working paper, Department of Mathematics, Fudan University, Shanghai, P. R. China, 2000.

[17] V. A. Yakubovich, *The frequency theorem in control theory*, Siberian Math. J., 14 (1973), pp. 265–289.

[18] V. A. Yakubovich, *The frequency theorem for the case in which the state space and the control space are Hilbert spaces and its application in certain problems in the synthesis of optimal control.* I, Siberian Math. J., 15 (1974), pp. 457–476.

[19] V. A. Yakubovich, *The frequency theorem for the case in which the state space and the control space are Hilbert spaces and its application in certain problems in the synthesis of optimal control.* II, Siberian Math. J., 16 (1975), pp. 828–845.

[20] D. D. Yao, S. Zhang, and X. Y. Zhou, *Stochastic LQ control via semidefinite programming*, SIAM J. Control. Optim., to appear.

[21] J. Yong and X. Y. Zhou, *Stochastic Controls: Hamiltonian Systems and HJB Equations*, Springer-Verlag, New York, 1999.

[22] X. Y. Zhou and D. Li, *Continuous-time mean-variance portfolio selection: A stochastic LQ framework*, Appl. Math. Optim., 42 (2000), pp. 19–33.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 2, pp. 577–609

NEW SECOND-ORDER OPTIMALITY CONDITIONS FOR VARIATIONAL PROBLEMS WITH C^2-HAMILTONIANS*

VERA ZEIDAN†

Abstract. The generalized problem of Bolza with a C^2-Hamiltonian is considered. Necessary and sufficient conditions are obtained, respectively, in terms of the accessory problem, the existence of conjoined basis, and the existence of a solution to a Riccati equation with boundary conditions. No strengthened Legendre–Clebsch condition is required. When applied to a general optimal control problem, these results include and generalize known results.

Key words. generalized Bolza problem, optimal control, accessory problem, conjoined basis, Riccati equation, general boundary conditions, optimality criteria, strong, $W^{1,s}$- and L^s-weak local minima

AMS subject classifications. 49K05, 49K15

PII. S0363012999358725

1. Introduction.

Consider the generalized problem of Bolza,

$$\text{(P)} \qquad \text{minimize } J(x) = \ell(x(0), x(1)) + \int_0^1 L(t, x(t), \dot{x}(t))dt,$$

over all absolutely continuous functions (arcs)$x(\cdot) \in W^{1,1}[0,1]$ satisfying

$$\phi(x(0), x(1)) = 0, \tag{1.1}$$

where $\ell : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, $L : [0,1] \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, and $\phi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^r$, $r \leq 2n$. The norm in $W^{1,s}[0,1]$ is $\|x\|_{1,s} := |x(0)| + \|\dot{x}(\cdot)\|_s$. An arc is said to be *admissible* if it satisfies (1.1).

The Hamiltonian corresponding to the problem is

$$H(t,x,p) := \sup\{\langle p, v\rangle - L(t,x,v) : v \in \mathbb{R}^n\}. \tag{1.2}$$

The problem (P) was introduced by R. T. Rockafellar. It is more general than it appears to be, due to the fact that L is allowed to take the value $+\infty$. While this problem subsumes diverse constrained problems (e.g., calculus of variations and optimal control problems), it is distinguished from the classical setting by the lack of regularity of L. Nevertheless, the Hamiltonian itself may actually be well behaved for certain classes of problems. This was the reason for the program of studying the problem (P) from the point of view of the Hamiltonian H. In fact, existence theory involving conditions on H was obtained by Rockafellar [9], and first-order necessary conditions in terms of the Hamiltonian inclusions were developed by Clarke in [3] and [4]. Sufficiency criteria for optimality of zero, first and second order were derived in [14] and [16] for the case where $H(t,\cdot,\cdot)$ is not necessarily concave-convex. There, one state endpoint is assumed to be fixed, and the second-order sufficiency criterion

*Received by the editors July 9, 1999; accepted for publication (in revised form) February 9, 2001; published electronically August 29, 2001.

http://www.siam.org/journals/sicon/40-2/35872.html

†Department of Mathematics, Michigan State University, East Lansing, MI 48824-1027 (zeidan@math.msu.edu). A part of this research was supported by the National Science Foundation under grant DMS-0072598.

assumes that $H(t,\cdot,\cdot)$ is $C^{1+}(C^{1,1})$, that is, has a Lipschitz gradient, and is phrased in terms of a certain Riccati inequality. However, the question of second-order necessary conditions in terms of the Hamiltonian remains an open question. By analogy with the classical calculus of variations setting, these conditions are expected to be expressed in terms of

(i) the accessory problem,
(ii) the coupled point theory,
(iii) a conjoined basis with appropriate boundary conditions, and
(iv) a certain Riccati-type equation with appropriate boundary conditions.

As opposed to the classical setting, these conditions must be phrased in terms of the Hamiltonian, which can be "nice" even when L is "bad."

Now consider the *optimal control* problem

$$\text{(C)} \qquad \text{minimize } J(x,u) := \ell(x(0),x(1)) + \int_0^1 g(t,x(t),u(t))dt \text{ subject to}$$

$$\text{(1.3)} \qquad \begin{cases} \dot{x}(t) = f(t,x(t),u(t)) \text{ almost everywhere (a.e.)}, \\ u(t) \in U \text{ a.e.}, \\ \phi(x(0),x(1)) = 0, \end{cases}$$

where $x(\cdot) \in W^{1,1}[0,1]$ and $u(\cdot)$ is measurable, ℓ and ϕ are as in the problem (P), $f : [0,1] \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, and $g : [0,1] \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$. A pair (x,u) is called *admissible* if $x(\cdot) \in W^{1,1}[0,1]$, $u(\cdot)$ is measurable and (x,u) satisfies (1.3). An arc x is called *admissible* if there exists a measurable u for which (x,u) is admissible.

We shall assume the following.

$$\text{(A)} \qquad \begin{cases} \bullet\ f \text{ and } g \text{ are } (\mathcal{L} \times \mathcal{B})\text{-measurable and continuous in } (x,u), \\ \bullet\ U \text{ is closed}, \\ \bullet\ \ell \text{ is lower semicontinuous}. \end{cases}$$

The Hamiltonian associated with (C) is

$$\text{(1.4)} \qquad \mathcal{H}(t,x,p) := \sup\{p \cdot f(t,x,u) - g(t,x,u) \mid u \in U\}.$$

The program of studying the problem (C) from the point of view of the "true" Hamiltonian $\mathcal{H}$ is one step further along than that for (P). In fact, in addition to the existence theory of [9], the first-order necessary conditions in [4], and the sufficiency criteria up to the second order in [14] and [15] for the case where only one state endpoint varies, necessary conditions for (C) were obtained in [2, Theorem 5.1] in terms of the Riccati equation $(\mathcal{R})$ and (3.17). However, these conditions were derived for $W^{1,1}$-weak minimality and for the case where the initial state variable is *fixed* and the final one is free. There, a triplet $(\bar{x},\bar{u},\bar{p})$ satisfying the first-order necessary conditions is considered. It is assumed that $\nabla^2\ell$ is Lipschitz and $\nabla^2_{(x,p)}\mathcal{H}$ is Lipschitz at $(\bar{x},\bar{p})$ from L^∞ to L^1 (see Definition 2.2). The result states that if the Riccati equation $(\mathcal{R})$ and the boundary condition (3.17) have on $]t_c,t]$ for some $t_c \in (0,1)$, a solution that does not extend to t_c, and if the *strengthened Legendre–Clebsch* condition $(S\mathcal{L})$ holds near t_c and for $t < t_c$,

$$(S\mathcal{L}) \qquad \text{for some } \lambda > 0, \quad \mathcal{H}_{pp}(t,\bar{x}(t),\bar{p}(t)) \geq \lambda I \quad \text{a.e.},$$

and if

$$\sup_{t\in[t_c-\lambda,t_c]} \left|\nabla^2_{(x,p)}\mathcal{H}(t,\bar{x}(t),\bar{p}(t))\right| < \infty, \tag{1.5}$$

then $(\bar{x},\bar{u})$ is not a $W^{1,1}$-weak local minimum for (C). These assumptions (see Lemma 5.4 and Remark 5.2) imply that the problem (C) resembles the classical calculus of variations setting, which is quite restrictive. Also the question of necessary conditions of types (i)–(iii) remains open.

The aim of this paper is to *complete* the program of studying the problems (P) and (C) from the point of view of the "true" Hamiltonian. This is accomplished by answering the open necessity and sufficiency questions of types (i), (iii), and (iv). Conditions of type (ii) will be handled in a separate paper.

The paper is divided as follows. In section 2, preliminary results are derived concerning the $W^{1,s}$- and L^s-weak local optimality notions used in the manuscript and the connection with the classical ones. In the same section we recall known results on the connection between (P) and (C) and on the notion of M-controllability.

For the problem (P), necessary conditions in terms of the Hamiltonian H are obtained in section 3 for $W^{1,s}$-local optimality. The first condition takes the form of the accessory problem. The novelty of its proof lies in associating to (P) an optimal control problem ($\mathrm{C}_{\mathcal{H}}$) that inherits the optimality of (P). The accessory problem for ($\mathrm{C}_{\mathcal{H}}$) is what we call the accessory problem of (P). As a consequence and by means of the results in [17], we obtain necessary conditions of types (iii) and (iv) for the case when the endpoint costs and constraints are separable. In section 4 we develop a sufficiency optimality criterion in (P) for general endpoints conditions and another one when the endpoints cost and constraint are separable. This latter enjoys the distinction that it is as close as possible to the necessary conditions developed in section 3.

In section 5 we apply the results of sections 3 and 4 to the optimal control problem (C) to obtain necessary and sufficient conditions. Indeed, the necessary conditions are derived under two different sets of conditions depending on whether we write (C) in the framework of (P) or ($\mathrm{C}_{\mathcal{H}}$). Neither of the two sets requires the strengthened Legendre–Clebsch conditions ($S\mathcal{L}$). Hence, when specialized to the case considered in [2], our results are much more general. More specifically, part (1)(ii) of Theorem 5.2 of this paper extends [2, Theorem 5.1] to the case where the Hessians $\nabla^2\ell$ and $\nabla^2\mathcal{H}$ are not necessarily Lipschitz, (1.5) is not satisfied, or the strengthened Legendre–Clebsch ($S\mathcal{L}$) does not hold. This latter is replaced by a normality condition which is automatically satisfied when ($S\mathcal{L}$) holds. Example 5.1 shows that the normality condition can be satisfied when ($S\mathcal{L}$) fails, and hence it illustrates the utility of our generalization. Furthermore, part (1) of Theorem 5.2 applies equally to all $W^{1,s}$-weak local minimum for $s\in[1,\infty]$. Moreover, part (2) of Theorem 5.2 applies to all L^s-weak local minimum for $s\in[1,\infty]$. In addition to necessary conditions in terms of the Riccati equation ($\mathcal{R}$), Theorem 5.2 develops necessary conditions in terms of the accessory problem (Theorem 3.3) and a conjoint basis for (3.14)–(3.15) (Theorem 3.9).

In this paper we assume that $H(t,\cdot,\cdot)$ and $\mathcal{H}(t,\cdot,\cdot)$ are C^2. This assumption is appropriate for certain classes of optimal control problems (see, e.g., [2]). Furthermore, the results of this paper will serve as a principal stepping stone for a subsequent paper where this regularity assumption on H and $\mathcal{H}$ is weakened.

2. Preliminary results. The following three notions of local optimality apply to both problems (P) and (C).

DEFINITION 2.1. *In* (P) *(resp., in* (C)*), an admissible arc* $\bar{x}$ *is said to be*

(i) *a strong local minimum if there exists* $\varepsilon > 0$ *such that for any admissible arc* x *satisfying* $\|x - \bar{x}\|_\infty < \varepsilon$ *we have* $J(x) \geq J(\bar{x})$ *(resp.,* $J(x,u) \geq J(\bar{x},\bar{u})$*),*

(ii) *a* $W^{1,s}$*-weak local minimum for* $1 \leq s \leq \infty$ *if there exists* $\varepsilon > 0$ *such that for any admissible arc* x *for which* $\|x - \bar{x}\|_{1,s}$ *is defined and* $\|x - \bar{x}\|_{1,s} < \varepsilon$*, we have* $J(x) \geq J(\bar{x})$ *(resp.,* $J(x,u) \geq J(\bar{x},\bar{u})$*).*

Clearly, if $\bar{x}$ is a strong local minimum, then $\bar{x}$ is a $W^{1,s}$-weak local minimum for all $s \in [1,\infty]$. If for $s \in [1,\infty]$ $\bar{x}$ is a $W^{1,s}$-weak local minimum, then it is a $W^{1,\infty}$-weak local minimum.

In the control context the notion of strong local minimality coincides with the classical notion. On the other hand, the following *classical* notions of L^s-weak local minima are known for (C).

DEFINITION 2.2. *Let* $s \in [1,\infty]$*. An admissible pair* $(\bar{x},\bar{u})$ *for* (C) *is a classically* L^s*-weak local minimum if for some* $\varepsilon > 0$

$$J(x,u) \geq J(\bar{x},\bar{u})$$

for all admissible pairs (x,u) *satisfying*

$$\|x - \bar{x}\|_\infty + \|u - \bar{u}\|_s < \varepsilon.$$

We shall need the following notions. A function K on $[0,1] \times \mathbb{R}^k$ is $(\mathcal{L} \times \mathcal{B})$-*measurable* if it is measurable with respect to the σ-algebra generated by products of Lebesgue sets in $[0,1]$ and Borel sets in $\mathbb{R}^k$.

For $\bar{z} \in L^\infty[0,1]$ we give these definitions.

DEFINITION 2.3. *Let* $K : [0,1] \times \mathbb{R}^k \to \mathbb{R}$ *be* $(\mathcal{L} \times \mathcal{B})$*-measurable. For* $1 \leq s \leq \infty$*, we say that* K *is continuous at* $\bar{z}$ *from* L^∞ *to* L^s *if the map* $z(\cdot) \to K(\cdot, z(\cdot))$ *from* L^∞ *to* L^s *is continuous at* $\bar{z}$*. The function* K *is said to be Lipschitz near* $\bar{z}$ *from* L^∞ *to* L^s *if the map* $z(\cdot) \to K(\cdot, z(\cdot))$ *is Lipschitz near* $\bar{z}$ *from* L^∞ *to* L^s*.*

DEFINITION 2.4. *Let* $\bar{x} \in L^\infty$*,* $\bar{y} \in L^s$ *for* $s \in [1,\infty]$*, and let* $K : [0,1] \times \mathbb{R}^r \times \mathbb{R}^s \to \mathbb{R}$*. The function* K *is said to be continuous at* $(\bar{x},\bar{y})$ *from* $L^\infty \times L^s$ *to* L^s *if the map* $(x(\cdot), y(\cdot)) \to K(\cdot, x(\cdot), y(\cdot))$ *is continuous at* $(\bar{x},\bar{y})$ *from* $L^\infty \times L^s \to L^s$*.*

We say that K *is Lipschitz near* $(\bar{x},\bar{y})$ *from* $L^\infty \times L^s$ *to* L^s *if the map* $(x(\cdot), y(\cdot)) \to K(\cdot, x(\cdot), y(\cdot))$ *from* $L^\infty \times L^s$ *to* L^s *is Lipschitz near* $(\bar{x},\bar{y})$*.*

It is clear that the Lipschitz property of K yields the corresponding continuity property of K. However, as we shall see below, the continuity of the gradient $\nabla_z K$ yields the Lipschitz condition on K.

LEMMA 2.5. *Let* $K(t,\cdot)$ *be* C^1*, and let* $\bar{z} \in L^\infty$*. Assume that* $\nabla_z K(\cdot, \bar{z}(\cdot))$ *is* L^s *for some* s *in* $[1,\infty]$*. If* $\nabla_z K$ *is continuous at* $\bar{z}$ *from* L^∞ *to* L^s*, then* K *is Lipschitz near* $\bar{z}$ *from* L^∞ *to* L^s*.*

Proof. The continuity at $\bar{z}$ of $\nabla_z K$ from L^∞ to L^s yields the existence of $\delta > 0$ such that for all $z : \|z - \bar{z}\|_\infty < \delta$ we have

$$\|\nabla_z K(\cdot, z(\cdot)) - \nabla_z K(\cdot, \bar{z}(\cdot))\|_s < 1.$$

Then, for all z and z' with $\|z - \bar{z}\|_\infty < \delta$, $\|z' - \bar{z}\|_\infty < \delta$, the mean value theorem implies that for almost all t there exists $\tilde{z}_t$ satisfying $|\tilde{z}_t - \bar{z}(t)| < \delta$ and

$$\begin{aligned} |K(t,z(t)) - K(t,z'(t))| &\leq |\nabla_z K(t,\tilde{z}_t)| \cdot |z(t) - z'(t)| \\ &\leq |\nabla_z K(t,\tilde{z}_t) - \nabla_z K(t,\bar{z}(t))| \cdot \|z - z'\|_\infty \\ &\quad + |\nabla_z K(t,\bar{z}(t))| \cdot \|z - z'\|_\infty. \end{aligned}$$

Since $\nabla_z K(\cdot, \bar{z}(\cdot))$ is L^s, by the Minkowski inequality we get

$$\|K(\cdot, z(\cdot)) - K(\cdot, z'(\cdot))\|_s \leq (1 + \|\nabla_z K(\cdot, \bar{z}(\cdot))\|_s)\|z - z'\|_\infty.$$

Therefore, the result follows. □

The next result shows that, under natural hypotheses on the dynamics f, the $W^{1,s}$-weak local minimality in (C) implies the *classical* L^s-weak local minimality. This implication will play an important role in developing the results of this paper. The converse is shown to hold under more restrictive hypotheses on f and is not needed for this paper.

LEMMA 2.6. *Let* (A) *be satisfied, and let* $(\bar{x}, \bar{u})$ *be an admissible pair for* (C) *such that* $\bar{u} \in L^s$ *for* $s \in [1, \infty]$.

(a) *Assume that* $(\bar{x}, \bar{u})$ *is a* $W^{1,s}$*-weak local minimum for* (C). *If* f *is continuous at* $(\bar{x}, \bar{u})$ *from* $L^\infty \times L^\infty$ *to* L^s, *then* $(\bar{x}, \bar{u})$ *is a classically* L^∞*-weak local minimum for* (C). *If* f *is continuous at* $(\bar{x}, \bar{u})$ *from* $L^\infty \times L^s$ *to* L^s, *then* $(\bar{x}, \bar{u})$ *is a classically* L^s*-weak local minimum for* (C).

(b) *Conversely, let* $(\bar{x}, \bar{u})$ *be an* L^s*-weak local minimum for* (C). *Assume that the map* $x(\cdot) \to f(\cdot, x(\cdot), u(\cdot))$ *is continuous at* $\bar{x}(\cdot)$ *from* L^∞ *to* L^s *uniformly in* $u(\cdot) \in L^s$ *with* $u(t) \in U$ *a.e., and that there exists* $m > 0$ *such that for all* $u(\cdot) \in L^s$ *such that* $u(t) \in U$ *a.e.,*

$$\|f(\cdot, \bar{x}(\cdot), u(\cdot)) - f(\cdot, \bar{x}(\cdot), \bar{u}(\cdot))\|_s \geq m\|u - \bar{u}\|_s.$$

Then $(\bar{x}, \bar{u})$ *is a* $W^{1,s}$*-weak local minimum for* (C).

Proof. For (a), let $\varepsilon > 0$ be such that for any admissible pair (x, u) for which $\|x - \bar{x}\|_{1,s}$ is defined and $\|x - \bar{x}\|_{1,s} < \varepsilon$, we have $J(x, u) \geq J(\bar{x}, \bar{u})$. If f is continuous at $(\bar{x}, \bar{u})$ from $L^\infty \times L^\infty$ to L^s, there exists $\delta_0 > 0$ such that for $\|(x, u) - (\bar{x}, \bar{u})\|_\infty < \delta_0$ we have $\|f(\cdot, x(\cdot), u(\cdot)) - f(\cdot, \bar{x}(\cdot), \bar{u}(\cdot))\|_s < \frac{\varepsilon}{2}$. Set

$$\bar{\varepsilon} := \min\left\{\frac{\varepsilon}{2}, \delta_0\right\}.$$

Now let (x, u) be admissible for (C) with

$$\|x - \bar{x}\|_\infty + \|u - \bar{u}\|_\infty < \bar{\varepsilon};$$

it follows that

$$\|x - \bar{x}\|_{1,s} \leq |x(0) - \bar{x}(0)| + \|f(\cdot, x(\cdot), u(\cdot)) - f(\cdot, \bar{x}(\cdot), \bar{u}(\cdot))\|_s < \varepsilon,$$

whence

$$J(x, u) \geq J(\bar{x}, \bar{u}),$$

proving that $(\bar{x}, \bar{u})$ is *classically* L^∞-weak local minimum. On the other hand, if f is continuous at $(\bar{x}, \bar{u})$ from $L^\infty \times L^s$ to L^s, then there exists $\delta_0 > 0$ such that for $\|x - \bar{x}\|_\infty < \delta_0$ and $\|u - \bar{u}\|_s < \delta_0$ we have $\|f(\cdot, x(\cdot), u(\cdot)) - f(\cdot, \bar{x}(\cdot), \bar{u}(\cdot))\|_s < \frac{\varepsilon}{2}$. Set $\bar{\varepsilon} := \min\{\frac{\varepsilon}{2}, \delta_0\}$, and let (x, u) be admissible for (C) with

$$\|x - \bar{x}\|_\infty + \|u - \bar{u}\|_s < \bar{\varepsilon}.$$

Then $\|x - \bar{x}\|_{1,s} < \varepsilon$. Thus $J(x, u) \geq J(\bar{x}, \bar{u})$; that is, $(\bar{x}, \bar{u})$ is *a classically* L^s-weak local minimum.

For (b), the L^s-weak local minimality of $(\bar{x},\bar{u})$ for (C) yields the existence of $\varepsilon>0$ such that $J(x,u)\geq J(\bar{x},\bar{u})$ for all admissible pairs (x,u) satisfying

$$\|x-\bar{x}\|_\infty+\|u-\bar{u}\|_s<\varepsilon.$$

By the continuity assumption on f, there exists $\delta_0>0$ such that for $\|x-\bar{x}\|_\infty<\delta_0$,

$$\|f(\cdot,x(\cdot),u(\cdot))-f(\cdot,\bar{x}(\cdot),u(\cdot))\|_s<\frac{\varepsilon m}{4}$$

for all $u(\cdot)\in L^s$ with $u(t)\in U$ a.e., where m is the constant in the assumption.

Define $\bar{\varepsilon}=\min\{\frac{m\varepsilon}{4},\frac{\varepsilon}{2},\delta_0\}$. Let (x,u) be admissible for (C) with $\|x-\bar{x}\|_{1,s}<\bar{\varepsilon}$. Then $\|x-\bar{x}\|_\infty<\frac{\varepsilon}{2}$, and

$$\|f(\cdot,x(\cdot),u(\cdot))-f(\cdot,\bar{x}(\cdot),\bar{u}(\cdot))\|_s<\frac{m\varepsilon}{4}.$$

It results that

$$\begin{aligned}\frac{m\varepsilon}{4}&>\|f(\cdot,\bar{x}(\cdot),u(\cdot))-f(\cdot,\bar{x}(\cdot),\bar{u}(\cdot))+f(\cdot,x(\cdot),u(\cdot))-f(\cdot,\bar{x}(\cdot),u(\cdot))\|_s\\&\geq\|f(\cdot,\bar{x}(\cdot),u(\cdot))-f(\cdot,\bar{x}(\cdot),\bar{u}(\cdot))\|_s-\|f(\cdot,x(\cdot),u(\cdot))-f(\cdot,\bar{x}(\cdot),u(\cdot))\|_s\\&>m\|u-\bar{u}\|_s-\frac{m\varepsilon}{4}.\end{aligned}$$

Thus $\|u-\bar{u}\|_s<\frac{\varepsilon}{2}$, whence $J(x,u)\geq J(\bar{x},\bar{u})$, and, therefore, $(\bar{x},\bar{u})$ is a $W^{1,s}$-weak local minimum for (C). □

Let $\bar{x}\in C[0,1]$ and $\varepsilon>0$ be given. Define

$$T(\bar{x};\varepsilon):=\{(t,x)\in[0,1]\times\mathbb{R}^n:|x-\bar{x}(t)|<\varepsilon\}. \tag{2.1}$$

We say that an arc $x(\cdot)\in T(\bar{x};\varepsilon)$ when $(t,x(t))\in T(\bar{x};\varepsilon)$ for all t.

There are several ways to write (C) in the form of (P). Below we exhibit a general one. Set

$$L_C(t,x,v):=\begin{cases}\inf\{g(t,x,u)\;\big|\;v=f(t,x,u)\text{ and }u\in U\}&\text{if }(t,x)\in T(\bar{x};\varepsilon),\\+\infty&\text{otherwise .}\end{cases} \tag{2.2}$$

Clearly, $L_C(t,x,v)=+\infty$ when $\{u:v=f(t,x,u),u\in U\}=\emptyset$. Furthermore, even if $L_C(t,x,v)$ is finite, it can be discontinuous or nonsmooth. To problem (C) we can associate the following *generalized Bolza problem*:

$$(\mathrm{P_C})\qquad\text{minimize }J_C(x):=\ell(x(0),x(1))+\int_0^1 L_C(t,x(t),\dot{x}(t))dt\text{ subject to}$$

$$\phi(x(0),x(1))=0.$$

The Hamiltonian corresponding to $(\mathrm{P_C})$ on $T(\bar{x};\varepsilon)\times\mathbb{R}^n$ is

$$\begin{aligned}H_C(t,x,p)&:=\sup\{\langle p,v\rangle-L_C(t,x,v):v\in\mathbb{R}^n\}\\&=\sup_{u\in U}\{\langle p,f(t,x,u)\rangle-g(t,x,u)\}\\&=\mathcal{H}(t,x,p),\end{aligned}$$

the Hamiltonian corresponding to (C).

Remark 2.1. The first two assumptions in (A) imply that L_C, defined by (2.2), is $(\mathcal{L} \times \mathcal{B})$-measurable. Indeed, define a set-valued map by

$$\Gamma(t,x,v) := \{u \in U : v = f(t,x,u)\}. \tag{2.3}$$

By [3, Proposition 3.1.2], $\Gamma(\cdot)$ is measurable and closed. Using [1, Theorem 8.2.11], it follows that L_C is $(\mathcal{L} \times \mathcal{B})$-measurable.

Remark 2.2. Since $\int_0^1 L_C(t,x(t),\dot{x}(t))dt \le \int_0^1 g(t,x(t),u(t))dt$ for all (x,u) admissible for (C), it results that if $\bar{x}$ is a strong or a $W^{1,s}$-weak local minimum for $(\mathrm{P_C})$, and if there exists $\bar{u}$ such that $(\bar{x},\bar{u})$ is admissible for (C) and

$$L_C(t,\bar{x}(t),\dot{\bar{x}}(t)) = g(t,\bar{x}(t),\bar{u}(t)) \text{ a.e.},$$

then it follows that $(\bar{x},\bar{u})$ is, respectively, a strong, $W^{1,s}$-weak local minimum for (C).

For the converse, an extra hypothesis is needed. Assume the following.

$$(\bar{\mathrm{A}}) \qquad \begin{cases} \bullet \text{ For } (t,x) \in T(\bar{x};\varepsilon) \text{ and } v \in f(t,x,U), \\ \qquad \text{the infimum in (2.2) is attained,} \\ \bullet \ L_C \text{ is lower semicontinuous in } (x,v). \end{cases}$$

With minor modification of the proof of the equivalence theorem in [9], we obtain the following.

THEOREM 2.7 (see [9]). *If assumptions* (A) *and* $(\bar{\mathrm{A}})$ *are satisfied, then for all arcs* $x(\cdot)$ *in* $T(\bar{x};\varepsilon)$ *that are admissible for* $(\mathrm{P_C})$ *we have that* $J_C(x)$ *is well defined. Furthermore, if* $J_C(x) < +\infty$, *then*

$$J_C(x) = \min_{u(\cdot)\ measurable} \{J(x,u) \ \mid \ \dot{x}(t) = f(t,x(t),u(t)) \ a.e., \ and \ u(t) \in U \ a.e.\}.$$

That is, $\bar{x}$ *is strong, or, for some* $s \in [1,\infty]$, *a* $W^{1,s}$*-weak local minimum for* $(\mathrm{P_C})$ *iff there exists a control* $\bar{u}$ *corresponding to* $\bar{x}$ *such that* $(\bar{x},\bar{u})$ *is, respectively, strong, or a* $W^{1,s}$*-weak local minimum for* (C).

In applications it is important to have more verifiable conditions guaranteeing $(\bar{\mathrm{A}})$. We shall use the following assumption.

$(\mathcal{B})$ For every *bounded* set $V \subseteq \mathbb{R}^n$, the following set is bounded: $\{u \in U \mid \exists\, (t,x,v) \in T(\bar{x};\varepsilon) \times V : \ v = f(t,x,u)\}$.

Remark 2.3. Assumptions (A) and $(\mathcal{B})$ yield assumption $(\bar{\mathrm{A}})$(i). Furthermore, we have

$$L_C(t,x,v) = \begin{cases} \inf\{g(t,x,u) : u \in \Gamma(t,x,v)\} & \text{if } \ (t,x) \in T(\bar{x};\varepsilon) \text{ and } \Gamma(t,x,v) \neq \phi, \\ +\infty & \text{otherwise,} \end{cases}$$

where Γ is given by (2.3). From (A) it follows that $\Gamma(t,\cdot,\cdot)$ is upper semicontinuous and from (A) and $(\mathcal{B})$ it results that $\Gamma(t,x,v)$ is compact. Thus, using [1, Theorem 1.4.16(ii)] we obtain that $L_C(t,\cdot,\cdot)$ is lower semicontinuous. Therefore, (A) and $(\mathcal{B})$ imply $(\bar{\mathrm{A}})$. When U is compact, $(\mathcal{B})$ holds automatically.

In the rest of this section we present notions of controllability and normality that we shall use throughout this paper. For given L^1-matrix functions $A(\cdot) : [0,1] \to \mathbb{R}^{n\times n}$ and $R(\cdot) : [0,1] \to \mathbb{R}^{n\times m}$, consider the linear system

$$\dot{\eta}(t) = A(t)\eta(t) + R(t)v(t) \qquad \text{a.e.}, \tag{2.4}$$

where $v(\cdot) \in L^\infty[0,1]$.

DEFINITION 2.8. *Let M be in $\mathbb{R}^{r\times 2n}$, and let $[a,b]$ be an interval in $[0,1]$. The system (2.4) is M-controllable on $[a,b]$ when for all $\beta \in \mathbb{R}^r$ there exist $v_\beta(\cdot) \in L^\infty[a,b]$ and $\eta(\cdot) \in W^{1,1}[a,b]$ such that (η, v_β) solves (2.4) on $[a,b]$ and satisfies*

$$M \begin{pmatrix} \eta(a) \\ \eta(b) \end{pmatrix} = \beta.$$

(When $r = 0$, this condition is satisfied trivially.)

The M-controllability on $[a,b]$ amounts to saying that

$$\operatorname{Im} \Lambda = \mathbb{R}^r,$$

where

$$\Lambda \begin{pmatrix} \alpha \\ v(\cdot) \end{pmatrix} := M \begin{bmatrix} I \\ Z_b(a) \end{bmatrix} \alpha + M \begin{bmatrix} 0 \\ I \end{bmatrix} \int_a^b Z_b(\tau) R(\tau) v(\tau) d\tau,$$

and $Z_b(\cdot)$ is defined by $-\dot{Z}(t) = Z(t)A(t)$ a.e., $Z(b) = I$; here I is the $n \times n$-identity matrix.

When $b = 1$, we denote Z_1 by Z.

Remark 2.4. The M-controllability on $[a,b]$ of the system (2.4) is known to be equivalent to its *M-normality*; that is, $(y,\gamma) \equiv 0$ is the only solution on $[a,b]$ to

$$\text{(2.5)} \qquad \begin{cases} \dot{y}(t) = -A^T(t)y(t) & \text{a.e.,} \\ \begin{pmatrix} y(a) \\ -y(b) \end{pmatrix} = M^T\gamma, \\ R^T(t)y(t) = 0 & \text{a.e.} \end{cases}$$

When the endpoints of η in (2.4) are separated, then for some $r_0 \times n$-matrix M_0 and $r_1 \times n$-matrix M_1 we have $M = \left[\begin{smallmatrix} M_0 & 0 \\ 0 & M_1 \end{smallmatrix}\right]$. In this case, the M-normality is denoted by $(M_0 : M_1)$-normality and the vector γ in (2.5) takes the form $\gamma = \left(\begin{smallmatrix} \gamma_0 \\ \gamma_1 \end{smallmatrix}\right)$, where $\gamma_0 \in \mathbb{R}^{r_0}$ and $\gamma_1 \in \mathbb{R}^{r_1}$.

3. Necessary conditions for (P). In this section we present the first part of the main results of this paper. It concerns deriving three types of second-order necessary conditions for (weak or strong) optimality in the generalized problem of Bolza (P). The key feature resides in the fact that all these conditions are phrased in terms of the Hamiltonian H given in (1.2). The first condition takes the form of the accessory problem, that is, the nonnegativity of a certain quadratic functional. It plays a fundamental role in obtaining the other two.

Let $\bar{z} := (\bar{x}, \bar{p}) \in W^{1,1}[0,1]$ be given. Throughout the paper we denote by $\bar{\Phi}(t)$ the evaluation $\Phi(t, \bar{z}(t))$ and by Φ_x and Φ_p the partial derivatives of Φ with respect to x and p, respectively. Elements in $\mathbb{R}^n$ are considered to be column vectors, and gradients of real-valued maps evaluated at a point are considered row vectors. The notation $B_n(x_0;\varepsilon)$ refers to the ball in $\mathbb{R}^n$ centered at x_0 of radius ε. When $x_0 = 0$, we write $B_n(\varepsilon)$.

In this section we assume that ℓ and ϕ are C^1 on $B_{2n}((\bar{x}(0), \bar{x}(1)); \varepsilon)$ and L is $(\mathcal{L} \times \mathcal{B})$-measurable. Set

$$\bar{M} := \nabla\phi(\bar{x}(0), \bar{x}(1)).$$

In stating our main results, we will refer to the following assumptions concerning (P).

There exists $\varepsilon > 0$ such that the following hold.

($\mathcal{L}$) $L(t,x,\cdot)$ is convex and lower semicontinuous for $(t,x) \in T(\bar{x};\varepsilon)$.

(H_1)
$$\begin{cases} \bullet\ \forall\, t,\ H(t,\cdot) \text{ is } C^1 \text{ on } B_{2n}(\bar{z}(t);\varepsilon), \\ \bullet\ \bar{H}(\cdot) \text{ and } \overline{\nabla_z H}(\cdot) \text{ are integrable}, \\ \bullet\ \nabla_z H \text{ is continuous at } \bar{z} \text{ from } L^\infty \text{ to } L^1. \end{cases}$$

(H_p^s) H_p is continuous at $\bar{z}$ from L^∞ to L^s.

(H_2)
$$\begin{cases} \bullet\ \forall\, t,\ H(t,\cdot) \text{ is } C^2 \text{ on } B_{2n}(\bar{z}(t);\varepsilon), \\ \bullet\ \bar{H}(\cdot),\ \overline{\nabla_z H}(\cdot),\ \overline{\nabla_z^2 H}(\cdot) \text{ are integrable}, \\ \bullet\ \nabla_z^2 H \text{ is continuous at } \bar{z} \text{ from } L^\infty \text{ to } L^1, \\ \bullet\ \ell \text{ and } \phi \text{ are } C^2 \text{ on } B_{2n}((\bar{x}(0),\bar{x}(1));\varepsilon), \\ \bullet\ \bar{M} \text{ is of full rank.} \end{cases}$$

When $s = 1$, (H_p^s) is included in (H_1)(iii). From Lemma 2.5, (H_2)(iii) yields (H_1) (iii). Hence (H_2) implies (H_1). Also, if we have that $\nabla_z H_p$ is continuous from L^∞ to L^s and $\overline{\nabla_z H_p}(\cdot) \in L^s$, by Lemma 2.5 we get that (H_p^s) holds true.

The last condition in (H_2) holds trivially in the free-endpoints case, where $r = 0$.

DEFINITION 3.1. *Let $\bar{z} := (\bar{x},\bar{p}) \in W^{1,1}[0,1]$ at which (H_1) holds. The arc $\bar{z}$ is called a normal extremal for* (P) *if for some $\bar{\gamma} \in \mathbb{R}^r$, it satisfies the Hamiltonian equations and the transversality conditions*

(3.1)
$$\begin{cases} -\dot{\bar{p}}^T(t) = \bar{H}_x(t) \quad a.e., \\ \dot{\bar{x}}^T(t) = \bar{H}_p(t) \quad a.e., \\ (\bar{p}^T(0), -\bar{p}^T(1)) = \nabla \ell(\bar{x}(0),\bar{x}(1)) + \bar{\gamma}^T \bar{M}. \end{cases}$$

Note that when the matrix $\bar{M}$ is of full rank, at most one vector $\bar{\gamma} \in \mathbb{R}^r$ can satisfy (3.1).

Clarke showed in [3] that when problem (P) is calm, $L(t,\cdot,\cdot)$ satisfies certain assumptions including ($\mathcal{L}$), and when H satisfies the strong Lipschitz condition

$\exists\ K(\cdot) \in L^1[0,1],\ \exists\ \varepsilon > 0 : \forall\, p \in \mathbb{R}^n,\ \forall\,(t,y_1)$, and $(t,y_2) \in T(\bar{x};\varepsilon)$, one has $|H(t,y_1,p) - H(t,y_2,p)| \le K(t)(1+|p|)|y_1 - y_2|$,

then a necessary condition for the *strong* minimality of $\bar{x}$ is the existence of $\bar{p} \in W^{1,1}[0,1]$ satisfying the Hamiltonian inclusions and the transversality conditions, and thus, if $H(t,\cdot,\cdot)$ is C^1 near $\bar{z} := (\bar{x},\bar{p})$, then $\bar{z}$ satisfies (3.1). The result in [3] was generalized by Clarke in [4] and also by Loewen and Rockafellar in [6]. When $\bar{x}$ is an $W^{1,\infty}$-weak local minimum for (P), it is proved that there exists an arc $\bar{p} \in W^{1,1}[0,1]$ satisfying the *Euler equations* and the *transversality conditions*. Under extra hypotheses, and if $H(t,\cdot,\cdot)$ is C^1 near $\bar{z} := (\bar{x},\bar{p})$, these conditions are equivalent to (3.1), as was proved by Rockafellar in [10].

DEFINITION 3.2. *For a given normal extremal* $(\bar{x}, \bar{p})$ *with associated* $\bar{\gamma}$*, the accessory problem corresponding to* (P) *at* $(\bar{x}, \bar{p}, \bar{\gamma})$ *is defined to be*

$$\text{(AP)} \quad \textit{minimize} \;\; J_2(\eta, v) := \frac{1}{2}\left\langle \Gamma \begin{pmatrix} \eta(0) \\ \eta(1) \end{pmatrix}, \begin{pmatrix} \eta(0) \\ \eta(1) \end{pmatrix} \right\rangle + \frac{1}{2}\int_0^1 \left\{ \langle \bar{H}_{pp}(t)v(t), v(t)\rangle - \langle \bar{H}_{xx}(t)\eta(t), \eta(t)\rangle \right\} dt$$

over $(\eta, v) \in W^{1,1}[0,1] \times L^\infty[0,1]$ *solution of*

$$\dot{\eta}(t) = \bar{H}_{px}(t)\eta(t) + \bar{H}_{pp}(t)v(t) \quad \textit{a.e.}, \tag{3.2}$$

$$\bar{M}\begin{pmatrix} \eta(0) \\ \eta(1) \end{pmatrix} = 0,$$

where $\Gamma := \nabla^2(\ell + \phi^T\bar{\gamma})|_{(\bar{x}(0),\bar{x}(1))}$.

Necessary conditions for (P) involving the accessory problem are given by the following result.

THEOREM 3.3. *Assume that* $\bar{z} := (\bar{x}, \bar{p}) \in W^{1,1}[0,1]$ *satisfy* (H_2) *and* $(\mathcal{L})$*, and form together with* $\bar{\gamma} \in \mathbb{R}^r$ *a normal extremal for* (P)*. Suppose that for* $s \in [1, \infty]$*,* $\bar{x}$ *is a* $W^{1,s}$*-weak local minimum with* (H_p^s) *satisfied. Then either the minimum value of the accessory problem is zero or else the* $\bar{M}$*-controllability of* (3.2) *fails on* $[0,1]$*.*

Let $Z : [0,1] \to \mathbb{R}^{n\times n}$ *denote the fundamental matrix of the system*

$$\begin{cases} -\dot{Z}(t) = Z(t)\bar{H}_{px}(t) & \textit{a.e.}, \\ Z(1) = I = \textit{identity matrix in } \mathbb{R}^{n\times n}. \end{cases} \tag{3.3}$$

Remark 3.1. From Definition 2.8, system (3.2) is *not* $\bar{M}$-controllable on $[0,1]$, meaning that Im Λ is a proper linear subspace of $\mathbb{R}^r$, where $A(t) := \bar{H}_{px}(t)$, $R(t) := \bar{H}_{pp}(t)$, and $Z_b := Z$. Hence there exists $\gamma \in \mathbb{R}^r$, $\gamma \neq 0$ satisfying

$$\gamma^T\bar{M}\begin{bmatrix} I \\ Z(0) \end{bmatrix} = 0$$

and

$$\gamma^T\bar{M}\begin{bmatrix} 0 \\ I \end{bmatrix} Z(t)\bar{H}_{pp}(t) = 0, \quad t \in [0,1] \quad \text{a.e.}$$

This is equivalent to saying that there exists $(y, \gamma) \not\equiv 0$ a solution on $[0,1]$ of (2.5), where the equivalence is revealed by setting $y(t) := Z^T(t)\bar{M}_R^T\gamma$. $\bar{M}_R$ is the $r \times n$-matrix in the partition $\bar{M} = (\bar{M}_L\ \bar{M}_R)$. Hence, if the system (3.2) is $\bar{M}$-normal, then the accessory problem must have a zero minimum value. This is the case, for instance, if we have that $\bar{H}_{pp}(t)$ is positive definite and $\bar{M}$ is of full rank, since in this case $(y, \gamma) \equiv 0$ is the only solution to the corresponding system (2.5).

The proof of Theorem 3.3, which we present later, will involve the following variational problem.

$$(\mathrm{C}_{\mathcal{H}}) \qquad \text{minimize } J_C(x,p) := \ell(x(0),x(1)) + \int_0^1 \{\langle p(t), \dot{x}(t)\rangle - H(t,x(t),p(t))\}dt$$

subject to

$$\dot{x}^T(t) = H_p(t,x(t),p(t)) \quad \text{a.e.},$$
$$\phi(x(0),x(1)) = 0,$$

where $x(\cdot) \in W^{1,1}[0,1]$, $p(\cdot) \in L^\infty[0,1]$, and H is the Hamiltonian given by (1.2).

Note that the problem $(\mathrm{C}_{\mathcal{H}})$ is an *optimal control* problem, where $x(\cdot)$ is the state and $p(\cdot)$ is the control. Hence the optimality notions of strong and $W^{1,s}$-weak local minima introduced in section 2 and the corresponding classical notions given in Definition 2.2 apply to $(\mathrm{C}_{\mathcal{H}})$.

It is known that the objective function $J_C(x,p)$, when unconstrained, has no local minima nor maxima. However, as we shall see below, when constrained as in $(\mathrm{C}_{\mathcal{H}})$, $J_C(x,p)$ admits a (local) minimum whenever the generalized Bolza problem (P) does.

PROPOSITION 3.4. *Assume that there exists* $\bar{z} := (\bar{x},\bar{p}) \in W^{1,1}[0,1] \times L^\infty[0,1]$ *that satisfy* $(\mathcal{L})$, (H_1), *and* (3.1)(ii). *Then, for* x *near* $\bar{x}$, $J(x)$ *is well defined (possibly* $+\infty$*), and if* $\bar{x}$ *is a strong,* $W^{1,s}$*-weak local minimum for* (P) *for some* $s \in [1,\infty]$, *then* $(\bar{x},\bar{p})$ *is, respectively, a strong, or* $W^{1,s}$*-weak local minimum for* $(\mathrm{C}_{\mathcal{H}})$.

Proof. From the definition of H in (1.2) and hypothesis (H_1)(i) it follows that $H(t,x,\cdot)$ is convex and lower semicontinuous for all $(t,x) \in T(\bar{x};\varepsilon)$. Then, by $(\mathcal{L})$ we obtain, for $(t,x,v) \in T(\bar{x};\varepsilon) \times \mathbb{R}^n$,

$$L(t,x,v) = \sup\{\langle p,v\rangle - H(t,x,p) \mid p \in \mathbb{R}^n\}, \tag{3.4}$$

whence, by (H_1)(ii)–(iii) and Lemma 2.5 it results that there exist $\bar{\delta} > 0$ ($\bar{\delta} < \varepsilon$) and γ such that for $\|z - \bar{z}\|_\infty < \bar{\delta}$,

$$\int_0^1 |H(t,z(t))|dt \le \gamma. \tag{3.5}$$

Then by the $(\mathcal{L} \times \mathcal{B})$-measurability of L we get that for $x \in W^{1,1}[0,1]$ satisfying $\|x - \bar{x}\|_\infty < \bar{\delta}$

$$\int_0^1 L(t,x(t),\dot{x}(t))dt \ge \int_0^1 \langle \bar{p}(t),\dot{x}(t)\rangle dt - \gamma,$$

and thus $J(x)$ is well defined (possibly $+\infty$) near $\bar{x}$. Using (3.4) again, it follows that for such x

$$\begin{aligned} J(x) &:= \ell(x(0),x(1)) + \int_0^1 L(t,x(t),\dot{x}(t))dt \\ &= \ell(x(0),x(1)) + \int_0^1 \sup_{p\in\mathbb{R}^n} \{\langle p,\dot{x}(t)\rangle - H(t,x(t),p)\}dt. \end{aligned}$$

The convexity of $H(t,x,\cdot)$ and (3.1)(ii) yield that

$$\begin{aligned} J(\bar{x}) &= \ell(\bar{x}(0),\bar{x}(1)) + \int_0^1 \{\langle \bar{p}(t),\dot{\bar{x}}(t)\rangle - \bar{H}(t)\}dt \\ &= J_C(\bar{x},\bar{p}). \end{aligned}$$

Furthermore, for any (x,p) that is admissible for $(\mathrm{C}_{\mathcal{H}})$ with $\|x-\bar{x}\|_\infty < \bar{\delta}$ we have

$$J(x) = J_C(x,p).$$

Therefore, assume there exists $\varepsilon_0 > 0$ such that

$$J(x) \geq J(\bar{x})$$

for all arcs $x(\cdot)$ that are admissible for (P) and satisfying $\|x-\bar{x}\|_\infty < \varepsilon_0$, or $\|x-\bar{x}\|_{1,s} < \varepsilon_0$, for some $s \in [1,\infty]$. Then, for all (x,p) admissible in $(\mathrm{C}_{\mathcal{H}})$ with $\|x-\bar{x}\|_\infty < \bar{\varepsilon}$, or $\|x-\bar{x}\|_{1,s} < \bar{\varepsilon}$, respectively, we have

$$J_C(x,p) = J(x) \geq J(\bar{x}) = J_C(\bar{x},\bar{p}),$$

where $\bar{\varepsilon} = \min\{\bar{\delta}, \varepsilon_0\}$. □

Combining Lemma 2.6 and Proposition 3.4, we obtain the connection between optimality in (P) and *classical* optimality in $(\mathrm{C}_{\mathcal{H}})$.

COROLLARY 3.5. *Assume that $\bar{z} := (\bar{x},\bar{p})$ satisfy the conditions of Proposition* 3.1. *Then the following hold.*

(i) *If $\bar{x}$ is a strong local minimum for* (P), *then $(\bar{x},\bar{p})$ is a classically L^∞-weak local minimum for* $(\mathrm{C}_{\mathcal{H}})$. *If, in addition, for all t, $H(t,\cdot)$ is C^1 on $B_n(\bar{x}(t);\varepsilon)\times\mathbb{R}^n$, then $(\bar{x},\bar{p})$ is a strong local minimum for* $(\mathrm{C}_{\mathcal{H}})$.

(ii) *If, for $s \in [1,\infty]$, $\bar{x}$ is a $W^{1,s}$-weak local minimum for* (P) *and* (H_p^s) *holds, then $(\bar{x},\bar{p})$ is a classically L^∞-weak local minimum for* $(\mathrm{C}_{\mathcal{H}})$. *If, in addition, $H(t,\cdot)$ is C^1 on $B_n(\bar{x}(t);\varepsilon)\times\mathbb{R}^n$ and H_p is continuous at $(\bar{x},\bar{p})$ from $L^\infty\times L^s$ to L^s, then $(\bar{x},\bar{p})$ is a classically L^s-weak local minimum for* $(\mathrm{C}_{\mathcal{H}})$.

The next result shows that if system (3.2) is $\bar{M}$-controllable (see Definition 2.8) and $(\bar{x},\bar{p})$ is a classically L^∞-weak local minimum for the optimal control problem $(\mathrm{C}_{\mathcal{H}})$, then the adjoint variable $q(\cdot)$ corresponding to $(\bar{x},\bar{p})$ in the weak version of the Pontryagin maximum principle applied to $(\mathrm{C}_{\mathcal{H}})$ coincides with $\bar{p}$.

LEMMA 3.6. *Assume that $(\bar{x},\bar{p},q,\gamma)$ satisfy the weak version of the Pontryagin maximum principle applied to* $(\mathrm{C}_{\mathcal{H}})$. *If the system* (3.2) *is $\bar{M}$-controllable, $\bar{M}$ is of full rank, and $(\bar{x},\bar{p})$ satisfy* (3.1), *then the problem* $(\mathrm{C}_{\mathcal{H}})$ *is normal and $(q,\gamma) = (\bar{p},\bar{\gamma})$, up to scalar multiplication.*

Proof. Let $(\bar{x},\bar{p},q,\gamma)$ satisfy the weak version of the maximum principle corresponding to $(\mathrm{C}_{\mathcal{H}})$. Then there exists $\lambda_0 \geq 0$ such that λ_0,γ, and q are not all zero, and

$$-\dot{q}(t) = \bar{H}_{xp}(t)q(t) - \lambda_0[\bar{H}_{xp}(t)\bar{p}(t) - \bar{H}_x(t)] \quad \text{a.e.},$$

$$\bar{H}_{pp}(t)[q(t) - \lambda_0\bar{p}(t)] = 0 \quad \text{a.e.},$$

$$(q^T(0), -q^T(1)) = \lambda_0\nabla\ell(\bar{x}(0),\bar{x}(1)) + \gamma^T\bar{M}.$$

If $\lambda_0 = 0$, then (q,γ) satisfies system (2.5), where $A = \bar{H}_{xp}$ and $R = \bar{H}_{pp}$. By the $\bar{M}$-controllability of the system (3.2) and Remark 2.4 we get that $q \equiv 0$ and $\gamma = 0$. Since λ_0,γ, and q are not all zero, we conclude that $\lambda_0 \neq 0$.

Now, assume $\lambda_0 = 1$, and use in the above equations that $(\bar{x},\bar{p})$ satisfy (3.1); we obtain

$$-(\dot{q}(t) - \dot{\bar{p}}(t)) = \bar{H}_{xp}(t)(q(t) - \bar{p}(t)) \quad \text{a.e.},$$

$$\bar{H}_{pp}(t)(q(t) - \bar{p}(t)) = 0 \quad \text{a.e.},$$

$$(q^T(0) - \bar{p}^T(0), -(q^T(1) - \bar{p}^T(1))) = (\gamma^T - \bar{\gamma}^T)\bar{M}.$$

Using the $\bar{M}$-controllability, it follows that $q \equiv \bar{p}$. The full rank condition on $\bar{M}$ yields that $\gamma = \bar{\gamma}$. □

In order to prove Theorem 3.3, we shall establish the following result, which says that the necessary conditions given by Theorem 3.3 are in fact necessary for the L^∞-weak local minimality of $(\bar{x}, \bar{p})$ in $(\mathrm{C}_\mathcal{H})$.

THEOREM 3.7. *Assume that* $(\bar{x}, \bar{p}) \in W^{1,1}[0,1]$ *satisfies* (H_2). *If* $(\bar{x}, \bar{p})$ *is a classically* L^∞*-weak local minimum for* $(\mathrm{C}_\mathcal{H})$, *then the conclusions of Theorem* 3.3 *are satisfied.*

To prove Theorem 3.7 we intend to calculate the second variation of the optimal control problem $(\mathrm{C}_\mathcal{H})$. For achieving this goal, one could attempt to apply known second-order necessary conditions. However, all known results (see [8], [12], and the references therein) require the dynamics and the integrand to be at least *twice* directionally *differentiable*. Since $H_p(t,\cdot,\cdot)$ is only C^1 and is involved in both the dynamics and the integrand of $(\mathrm{C}_\mathcal{H})$, those results cannot be applied here. Nevertheless, as we shall see below in the proof of Theorem 3.7, the second variation of $(\mathrm{C}_\mathcal{H})$ can be derived by taking its special structure into consideration. We shall need the following result, which says that the $\bar{M}$-controllability of system (3.2) allows us to associate with each pair (η, v) solving that system, an admissible family $(x(\cdot,\alpha), p(\cdot,\alpha))_{\alpha\in(-\delta,\delta)}$ for the problem $(\mathrm{C}_\mathcal{H})$ which includes $(\bar{x}, \bar{p})$ at $\alpha = 0$ and has a first derivative with respect to α at $\alpha = 0$ equal to (η, v).

PROPOSITION 3.8. *Assume* (H_2) *to hold at the admissible pair* $\bar{z} = (\bar{x}, \bar{p})$ *and that the system* (3.2) *is* $\bar{M}$*-controllable. Then, for every* $(\eta, v) \in W^{1,1}[0,1] \times L^\infty[0,1]$ *satisfying system* (3.2) *and its boundary conditions, there exists an admissible family* $(x(\cdot;\alpha), p(\cdot;\alpha))_{\alpha\in(-\delta,\delta)}$ *for* $(\mathrm{C}_\mathcal{H})$ *such that* $(x(t;\cdot),\ p(t;\cdot))$ *is* C^1, $(\frac{\partial x}{\partial\alpha}, \frac{\partial p}{\partial\alpha})$ *is continuous at* 0 *from* $\mathbb{R}$ *to* L^∞, *and*

(i) $x(t;0) = \bar{x}(t),\ p(t;0) = \bar{p}(t) \quad \forall\, t \in [0,1]$,

(ii) $\frac{\partial x}{\partial\alpha}(t;0) = \eta(t),\ \frac{\partial p}{\partial\alpha}(t;0) = v(t) \quad \forall\, t \in [0,1]$.

Proof. Let $\beta_1, \ldots, \beta_r$ be r linearly independent vectors in $\mathbb{R}^r$. By the $\bar{M}$-controllability of system (3.2), there exist functions $v_1, \ldots, v_r \in L^\infty[0,1]$ and $\eta_1, \ldots, \eta_r$ in $W^{1,1}[0,1]$ satisfying

$$\left\{\begin{array}{l} \dot{\eta}_j(t) = \bar{H}_{px}(t)\eta_j(t) + \bar{H}_{pp}(t)v_j(t) \quad \text{a.e.,} \\ \bar{M}\begin{pmatrix} \eta_j(0) \\ \eta_j(1) \end{pmatrix} = \beta_j. \end{array}\right. \tag{3.6}$$

Define on $[0,1] \times \mathbb{R} \times \mathbb{R}^r$

$$p(t;\alpha,\lambda) := \bar{p}(t) + \alpha v(t) + \sum_{j=1}^{r} \lambda_j v_j(t). \tag{3.7}$$

Consider the system

$$\left\{\begin{array}{l} \dot{x}^T(t) = H_p(t, x(t), p(t,\alpha,\lambda)), \\ x(0) = \bar{x}(0) + \alpha\eta(0) + \sum_{j=1}^{r} \lambda_j \eta_j(0). \end{array}\right. \tag{3.8}$$

Clearly, when $(\alpha,\lambda) = (0,0)$, system (3.8) has $\bar{x}$ as the solution. Hypothesis (H_2) and Lemma 2.5 yield that for

$$F(t,x,\alpha,\lambda) := H_p(t,x,p(t,\alpha,\lambda))$$

F and $\nabla_{(x,\alpha,\lambda)}F$ are Carathéodory on some tube $T(\bar{x},(0,0);\varepsilon_0)$ for $\varepsilon_0 \leq \varepsilon$. Hence, by the embedding theorem of differential equations [11, II.4.11], it results that there exists $\bar{\delta} > 0$ ($\bar{\delta} < \varepsilon_0$) such that for $(\alpha,\lambda) \in B_{r+1}(\bar{\delta})$ and for $x_0 \in B_n(\bar{x}(0);\bar{\delta})$, the system (3.8)(i) has a unique solution $x(\cdot;x_0,\alpha,\lambda)$ in $W^{1,1}[0,1]$ satisfying $x(0;x_0,\alpha,\lambda) = x_0$. Furthermore, on $B_n(\bar{x}(0);\bar{\delta}) \times B_{r+1}(\bar{\delta})$, $x(t,\cdot,\cdot,\cdot)$ has a derivative that is continuous, L^∞-uniformly in t. However, by (3.8)(ii), there exists $\delta_0 > 0$ ($\delta_0 \leq \bar{\delta}$) such that for $(\alpha,\lambda) \in B_{r+1}(\delta_0)$, the unique solution to (3.8) $x(t;\alpha,\lambda) := x(t;x_{\alpha,\lambda},\alpha,\lambda)$, where $x_{\alpha,\lambda} := \bar{x}(0) + \alpha\eta(0) + \sum_{j=1}^r \lambda_j\eta_j(0)$, has a derivative with respect to (α,λ) that is continuous at 0 from $\mathbb{R}^{r+1}$ to L^∞.

Differentiate the system (3.8) with respect to (α,λ) and use (3.2), (3.6), and (3.7) to get

$$\text{(3.9)} \qquad \begin{cases} \dfrac{\partial x}{\partial \alpha}(\cdot;0,0) = \eta(\cdot), & \dfrac{\partial p}{\partial \alpha}(\cdot;0,0) = v(\cdot), \\ \dfrac{\partial x}{\partial \lambda_j}(\cdot;0,0) = \eta_j(\cdot), & \dfrac{\partial p}{\partial \lambda_j}(\cdot;0,0) = v_j(\cdot) \quad \forall\, j = 1,\dots,r. \end{cases}$$

Define $G : B(\delta_0) \times B_r(\delta_0) \to \mathbb{R}^r$ by

$$(\alpha,\lambda) \to G(\alpha,\lambda) := \phi(x(0;\alpha,\lambda), x(1;\alpha,\lambda)).$$

We have $G(0,0) = 0$, $\frac{\partial G(0,0)}{\partial \lambda_j} = \bar{M}\binom{\eta_j(0)}{\eta_j(1)} = \beta_j$, and thus $\frac{\partial G}{\partial \lambda}(0,0)$ is nonsingular. By the implicit function theorem, there exist $\delta < \delta_0$ and a C^1-function $\lambda(\cdot) : B(\delta) \to \mathbb{R}^r$ with $\lambda(0) = 0$, $\lambda(B(\delta)) \subset B(\delta_0)$, and

$$\text{(3.10)} \qquad \phi(x(0;\alpha,\lambda(\alpha)), x(1;\alpha,\lambda(\alpha))) = 0.$$

Differentiating (3.10) with respect to α at $\alpha = 0$ and using (3.9), (3.6), and the fact the (η, v) satisfies (3.2), we get $\sum_{j=1}^r \dot{\lambda}_j(0)\beta_j = 0$, which yields $\dot{\lambda}_j(0) = 0$ for all j.

Now, set on $[0,1] \times (-\delta,\delta)$

$$p(t;\alpha) := p(t;\alpha,\lambda(\alpha)),$$

$$x(t;\alpha) := x(t;\alpha,\lambda(\alpha)).$$

Using (3.7) and (3.9), we obtain the result of this proposition. □

We are now ready to prove Theorem 3.7. As we shall see in the proof, the map

$$\alpha \to J_C(x(\cdot;\alpha), p(\cdot;\alpha))$$

is twice differentiable at 0, even though the functions defining $J_C(x,p)$ are not themselves twice differentiable.

Proof of Theorem 3.7. Let $(\bar{x},\bar{p})$ satisfy (3.1).

Assume that system (3.2) is $\bar{M}$-controllable and that $(\bar{x},\bar{p})$ provides a classically L^∞-weak local minimum for $(C_\mathcal{H})$. Thus $(\bar{x},\bar{p})$ must satisfy the weak version of the Pontryagin maximum principle. By Lemma 3.6, we deduce that $(C_\mathcal{H})$ is normal, and the adjoint variable corresponding to $(\bar{x},\bar{p})$ is $\bar{p}$.

Let (η, v) satisfy system (3.2) and its boundary conditions. Then, by Proposition 3.8, there exists a family $(x(\cdot;\alpha), p(\cdot;\alpha))_{\alpha\in(-\delta,\delta)}$ that is admissible for $(C_\mathcal{H})$ and satisfies the conditions stated in Proposition 3.8. In particular, using Lemma 2.5, it

follows that there exist $\gamma > 0 (\gamma \leq \delta)$ and $\mu > 0$ such that, for almost all t and for $\alpha, \alpha' \in (-\gamma, \gamma)$,

$$|(x(t,\alpha), p(t,\alpha)) - (x(t;\alpha'), p(t,\alpha'))| \leq \mu|\alpha - \alpha'|. \tag{3.11}$$

Thus choose γ small enough so that $(\bar{x}, \bar{p})$ is optimal for J_C on $T(\bar{x}, \bar{p}; \mu\gamma)$. Hence, by setting $\mathbb{H}(t;\alpha) := H(t, x(t;\alpha), p(t;\alpha))$ and $\mathbb{J}_C(\alpha) := J_C(x(\cdot;\alpha), p(\cdot;\alpha))$, we get that $\alpha = 0$ is a minimum for $\mathbb{J}_C(\alpha)$ over $(-\gamma, \gamma)$. On the other hand, hypothesis (H_2) and Lemma 2.5 yield that there exists a neighborhood of 0, $(-\bar{\gamma}, \bar{\gamma})$, where $\mathbb{H}, \mathbb{H}_p$, and $\mathbb{H}_x$ are Lipschitz from $\mathbb{R}$ to L^1 and $\mathbb{H}_p(t;\cdot)$ is bounded by an integrable function of t. Set $\gamma_0 := \min(\gamma, \bar{\gamma})$. Thus, for $\alpha \in (-\gamma_0, \gamma_0)$, we have

$$\begin{aligned}
\mathbb{J}_C(\alpha) &:= \ell(x(0;\alpha), x(1;\alpha)) + \int_0^1 \{\mathbb{H}_p(t;\alpha)p(t;\alpha) - \mathbb{H}(t;\alpha)\}dt \\
&= \ell(x(0;\alpha), x(1;\alpha)) + \int_0^1 \{\mathbb{H}_p(t;\alpha)(p(t;\alpha) - \bar{p}(t)) + \dot{x}^T(t;\alpha)\bar{p}(t) - \mathbb{H}(t;\alpha)\}dt \\
&= \ell(x(0;\alpha), x(1;\alpha)) + \bar{p}^T(1)x(1;\alpha) - \bar{p}^T(0)x(0;\alpha) \\
&\quad + \int_0^1 \{\mathbb{H}_p(t;\alpha)(p(t;\alpha) - \bar{p}(t)) - \mathbb{H}(t;\alpha) - \dot{\bar{p}}^T(t)x(t;\alpha)\}dt.
\end{aligned}$$

We shall show that $\mathbb{J}_C(\cdot)$ is twice differentiable at 0. Since on $(-\gamma_0, \gamma_0)$, x and p are Lipschitz from $\mathbb{R}$ to L^∞ (see (3.11)), $\mathbb{H}$ and $\mathbb{H}_p$ are Lipschitz near 0 from $\mathbb{R}$ to L^1, $\mathbb{H}_p(t;\cdot)$ is bounded by an L^1-function of t, and ℓ is Lipschitz, it follows from the dominated convergence theorem and (3.1)(iii) that for $\bar{\alpha} \in (-\gamma_0, \gamma_0)$ and for $x_\alpha = \frac{\partial x}{\partial \alpha}$, $p_\alpha = \frac{\partial p}{\partial \alpha}$, we have

$$\begin{aligned}
\mathbb{J}'_C(\bar{\alpha}) &= (\nabla\ell(x(0;\bar{\alpha}), x(1;\bar{\alpha})) - \nabla\ell(\bar{x}(0), \bar{x}(1)) - \bar{\gamma}^T\bar{M}) \begin{pmatrix} x_\alpha(0;\bar{\alpha}) \\ x_\alpha(1;\bar{\alpha}) \end{pmatrix} \\
&\quad + \int_0^1 \big\{(\mathbb{H}_{px}(t;\bar{\alpha})x_\alpha(t;\bar{\alpha}) + \mathbb{H}_{pp}(t;\bar{\alpha})p_\alpha(t;\bar{\alpha}))^T(p(t;\bar{\alpha}) - \bar{p}(t)) \\
&\quad - (\mathbb{H}_x(t;\bar{\alpha}) + \dot{\bar{p}}^T(t))x_\alpha(t;\bar{\alpha})\big\}dt.
\end{aligned}$$

Using (3.1), system (3.2), and Proposition 3.8, we obtain

$$\mathbb{J}'_C(0) = 0.$$

Let us calculate $\mathbb{J}''_C(0)$. For $h \in (-\gamma_0, \gamma_0)$ and $i \in \{1, \ldots, r\}$ we have

$$\phi^i(x(0;h), x(1;h)) = 0;$$

hence

$$\nabla\phi^i(x(0;h), x(1;h)) \begin{bmatrix} x_\alpha(0;h) \\ x_\alpha(1;h) \end{bmatrix} = 0.$$

Thus, by the mean value theorem applied to each partial derivative $\partial_{x_j}\phi^i$ and $\partial_{x_j}\ell$, where $j = 1, \ldots, 2n$, it results that there exist h_j^i and $\tilde{h}_j$ in between 0 and h

such that, for $\tilde{h}_j$,

$$\bar{M} = \begin{bmatrix} \bar{M}^1 \\ \vdots \\ \bar{M}^r \end{bmatrix} \text{ and } x_\alpha = \begin{pmatrix} x_\alpha^1 \\ \vdots \\ x_\alpha^n \end{pmatrix}$$

we have

$$\begin{aligned} -\frac{\bar{M}^i}{h} \begin{pmatrix} x_\alpha(0;h) \\ x_\alpha(1;h) \end{pmatrix} &= \left[\frac{\nabla\phi^i(x(0;h),x(1;h)) - \bar{M}^i}{h}\right] \begin{pmatrix} x_\alpha(0;h) \\ x_\alpha(1;h) \end{pmatrix} \\ &= \sum_{j=1}^{n} x_\alpha^j(0;h) \nabla\partial_{x_j}\phi^i(x(0,h_j^i),x(1,h_j^i)) \begin{pmatrix} x_\alpha(0;h_j^i) \\ x_\alpha(1;h_j^i) \end{pmatrix} \\ &\quad + \sum_{j=n+1}^{2n} x_\alpha^{j-n}(1;h) \nabla\partial_{x_j}\phi^i(x(0,h_j^i),x(1,h_j^i)) \begin{pmatrix} x_\alpha(0;h_j^i) \\ x_\alpha(1;h_j^i) \end{pmatrix} \end{aligned}$$

and

$$\frac{\partial_{x_j}\ell(x(0;h),x(1;h)) - \partial_{x_j}\ell(\bar{x}(0),\bar{x}(1))}{h} = \nabla\partial_{x_j}\ell(x(0;\tilde{h}_j),x(1;\tilde{h}_j)) \begin{pmatrix} x_\alpha(0;\tilde{h}_j) \\ x_\alpha(1;\tilde{h}_j) \end{pmatrix},$$

whence

$$\begin{aligned} &\lim_{h\to 0} \frac{(\nabla\ell(x(0;h),x(1;h)) - \nabla\ell(\bar{x}(0),\bar{x}(1)) - \bar{\gamma}^T\bar{M})}{h} \begin{pmatrix} x_\alpha(0;h) \\ x_\alpha(1;h) \end{pmatrix} \\ &\qquad = \left\langle \Gamma \begin{pmatrix} \eta(0) \\ \eta(1) \end{pmatrix}, \begin{pmatrix} \eta(0) \\ \eta(1) \end{pmatrix} \right\rangle, \end{aligned}$$

where $\Gamma := \nabla^2(\ell + \phi^T\bar{\gamma})|_{(\bar{x}(0),\bar{x}(1))}$.

On the other hand, from (H_2) and the Lipschitz property of x and p in h obtained in Proposition 3.8, it results that the integrand

$$\begin{aligned} \int_0^1 \{(\mathbb{H}_{px}(t;h)x_\alpha(t;h) + \mathbb{H}_{pp}(t;h)p_\alpha(t;h))^T \frac{(p(t;h) - \bar{p}(t))}{h} \\ - \frac{(\mathbb{H}_x(t;h) - \mathbb{H}_x(t;0))}{h} x_\alpha(t;h)\} dt \end{aligned}$$

is bounded by an integrable function. Therefore, by the dominated convergence theorem we get

$$\begin{aligned} \mathbb{J}_C''(0) := \lim_{h\to 0} \frac{\mathbb{J}_C'(h)}{h} &= \left\langle \Gamma \begin{pmatrix} \eta(0) \\ \eta(1) \end{pmatrix}, \begin{pmatrix} \eta(0) \\ \eta(1) \end{pmatrix} \right\rangle \\ &\quad + \int_0^1 \{v^T(t)\bar{H}_{px}(t)\eta(t) + v^T(t)\bar{H}_{pp}(t)v(t) - \eta^T(t)\bar{H}_{xx}(t)\eta(t) \\ &\qquad - \eta^T(t)\bar{H}_{xp}(t)v(t)\} dt \\ &= 2J_2(\eta, v). \end{aligned}$$

Now, since $\min\{\mathbb{J}_C(\alpha) : \alpha \in (-\gamma_0, \gamma_0)\} = \mathbb{J}_C(0)$ and $\mathbb{J}'_C(0) = 0$,

$$\mathbb{J}''_C(0) \geq 0.$$

Therefore, the proof of the theorem is complete. □

Proof of Theorem 3.3. Let $s \in [1, \infty]$, and let $\bar{x}$ be $W^{1,s}$-weak local minimum for (P) with (H_p^s) satisfied. Consider $\bar{z} = (\bar{x}, \bar{p})$ as in the theorem. Then, by Corollary 3.5, it results that $(\bar{x}, \bar{p})$ provides a classically L^∞-weak local minimum for $(C_\mathcal{H})$. Therefore, by Theorem 3.7 the proof is complete. □

Now, for the rest of this section, assume that the endpoints cost and constraint in (P) are separable; that is,

$$\ell(x, y) := \ell_0(x) + \ell_1(y),$$

$$\phi(x, y) := \begin{pmatrix} \phi_0(x) \\ \phi_1(y) \end{pmatrix},$$

where $\phi_i : \mathbb{R}^n \to \mathbb{R}^{r_i}$ for $i = 0, 1$, and $r_0 + r_1 = r$.

In this case, the matrices $\bar{M}$ and Γ in the accessory problem (AP) take the form

$$\bar{M} = \begin{bmatrix} \bar{M}_0 & 0 \\ 0 & \bar{M}_1 \end{bmatrix} \text{ and } \Gamma = \begin{bmatrix} \Gamma_0 & 0 \\ 0 & \Gamma_1 \end{bmatrix},$$

where

$$\bar{M}_0 := \nabla\phi_0(\bar{x}(0)), \qquad \bar{M}_1 = \nabla\phi_1(\bar{x}(1)),$$

$$\Gamma_0 := \nabla^2(\ell_0 + \phi_0^T\bar{\gamma}_0)\big|_{x=\bar{x}(0)}, \qquad \Gamma_1 = \nabla^2(\ell_1 + \phi_1^T\bar{\gamma}_1)\big|_{x=\bar{x}(1)},$$

and

$$\begin{pmatrix} \gamma_0 \\ \gamma_1 \end{pmatrix} := \gamma.$$

Here $\bar{M}_0 \in \mathbb{R}^{r_0 \times n}$, $\bar{M}_1 \in \mathbb{R}^{r_1 \times n}$, $\gamma_0 \in \mathbb{R}^{r_0}$, and $\gamma_1 \in \mathbb{R}^{r_1}$.

Note that when no final constraint is present, that is, when $\phi_1 : \mathbb{R}^n \to \{0\}$, then $\bar{M}_1$ is automatically of full rank.

Without loss of generality we assume that $\bar{M}_0$ and $\bar{M}_1$ are in fact orthogonal projections in $\mathbb{R}^n$. This can easily be done by taking, instead of $\bar{M}_i$, the orthogonal projections

$$\bar{M}_i^T(\bar{M}_i\bar{M}_i^T)^{-1}\bar{M}_i, \quad i = 0, 1,$$

where the $\bar{M}_i$'s are of full rank, by (H_2)(v).

When system (3.2) is $(\bar{M}_0 : \bar{M}_1)$-normal, the optimality of (η, v) in (AP) yields the existence of an arc q and vectors γ_0 and γ_1 satisfying on $[0, 1]$

$$\begin{cases} \dot{\eta}(t) = A(t)\eta(t) + R(t)q(t) & \text{a.e.}, \\ -\dot{q}(t) = A^T(t)q(t) + P(t)\eta(t) & \text{a.e.}, \end{cases} \tag{3.12}$$

$$\begin{cases} q(0) = \Gamma_0\eta(0) + \bar{M}_0\gamma_0, \\ -q(1) = \Gamma_1\eta(1) + \bar{M}_1\gamma_1, \end{cases}$$

where

$$A(t) := \bar{H}_{px}(t), \quad R(t) := \bar{H}_{pp}(t), \quad \text{and} \quad P(t) := \bar{H}_{xx}(t). \tag{3.13}$$

Consider (U_1, V_1) the solution of the matrix system corresponding to (3.12),

$$\begin{cases} \dot{U}(t) = A(t)U(t) + R(t)V(t) \quad \text{a.e.}, \\ -\dot{V}(t) = A^T(t)V(t) + P(t)U(t) \quad \text{a.e.}, \end{cases} \tag{3.14}$$

with final conditions

$$\begin{cases} U(1) = I - \bar{M}_1, \\ V(1) = -\Gamma_1(I - \bar{M}_1) - \bar{M}_1. \end{cases} \tag{3.15}$$

The following result states that the existence of a conjoined basis on $[0,1]$ for (3.14) and (3.15) with certain initial conditions is *necessary* for the optimality in (P) or in $(C_{\mathcal{H}})$.

THEOREM 3.9. *Let $\bar{z} := (\bar{x}, \bar{p})$ satisfy the conclusions of Theorem* 3.3. *If, in addition, for all $c \in (0,1)$ the system* (3.2) *is $(\bar{M}_0 : I)$-normal on $[0,c]$ and $(I : \bar{M}_1)$-normal on $[c,1]$, then (U_1, V_1) satisfies*

(i) $U^TV = V^TU$ *on* $[0,1]$,

(ii) $\det U(t) \neq 0$ *on* $(0,1)$,

(iii) $U^T(0)(\Gamma_0U(0) - V(0)) \geq 0$ *on* $K := \{\alpha : \bar{M}_0U(0)\alpha = 0\}$,

and

(iv) $V(1) + \Gamma_1(I - \bar{M}_1) + \bar{M}_1 = 0$.

Proof. From the normality hypotheses it follows that the accessory problem (AP) has a minimum value equal to zero. Thus by [17, Theorem 4.1] the result of the theorem follows. □

An immediate consequence of the above theorem is an important result on necessary conditions concerning the Riccati equation.

COROLLARY 3.10. *Under the conditions of Theorem* 3.9, *there exists on* $(0,1)$ *a symmetric absolutely continuous solution W of*

$$\dot{W}(t) + A^T(t)W(t) + W(t)A(t) + W(t)R(t)W(t) + P(t) = 0 \tag{$\mathcal{R}$}$$

satisfying

$$\begin{cases} \lim\limits_{t\to 1} W(t)U_1(t) = -\Gamma_1(I - \bar{M}_1) - \bar{M}_1, \\ U_1^T(0) \lim\limits_{t\to 0^+} (\Gamma_0 - W(t))U_1(t) \geq 0, \\ \qquad\qquad \text{on } K := \{\alpha \mid \bar{M}_0U_1(0)\alpha = 0\}. \end{cases} \tag{3.16}$$

Remark 3.2. When the final state endpoint is free, $\bar{M}_1 = 0$ and (3.15) yield that $U_1(1) = I$. In this case the function W in Corollary 3.10 is defined on $(0,1]$, and (3.16)(i) reduces to

$$W(1) = -\Gamma_1. \tag{3.17}$$

If also the initial condition is fixed, then (3.16)(ii) is trivially satisfied, and thus (3.16) becomes simply (3.17).

By symmetry, we could define (U_0, V_0) as the solution of (3.14), and instead of (3.15),

$$U(0) = I - \bar{M}_0,$$

$$V(0) = \Gamma_0(I - \bar{M}_0) + \bar{M}_0.$$

Then necessary conditions for (P) and ($C_{\mathcal{H}}$) parallel to Theorem 3.9 and Corollary 3.10 could be phrased in terms of (U_0, V_0).

The following example illustrates the utility of the results in this section.

Example 3.1. Define on $[0,1] \times \mathbb{R} \times \mathbb{R}$ the function

$$L(t,x,v) := \begin{cases} x^3 + \dfrac{v^2}{(t-\frac{1}{2})^{2/3}} & \text{if } t \neq \dfrac{1}{2}, \\ x^3 & \text{if } t = \dfrac{1}{2} \text{ and } v = 0, \\ +\infty & \text{if } t = \dfrac{1}{2} \text{ and } v \neq 0. \end{cases}$$

Consider the problem

$$(P_0) \qquad \text{minimize} \qquad -\frac{5}{3}2^{5/3}x^2(1) + \int_0^1 L(t, x(t), \dot{x}(t))dt$$

$$\text{subject to} \qquad x(0) = 0.$$

The Hamiltonian corresponding to (P_0) is

$$H(t,x,p) = \sup_{v \in \mathbb{R}} \{pv - L(t,x,v)\} = -x^3 + \frac{p^2}{4}\left(t - \frac{1}{2}\right)^{2/3}.$$

Take $\bar{x} \equiv 0$ and $\bar{p} \equiv 0$. Then L satisfies ($\mathcal{L}$), and H satisfies hypotheses (H_2) and (H_p^s) for all $s \in [1, \infty]$. Here, $\bar{M}_0 = I$, $\bar{M}_1 = 0$. For $\bar{\gamma} = \binom{0}{0}$, $(\bar{x}, \bar{p})$ with $\bar{\gamma}$ form a normal extremal for (P_0). In this problem, $A(t) \equiv 0$, $R(t) = \frac{1}{2}(t - \frac{1}{2})^{2/3}$, and $P(t) = 0$. Since the system $\dot{y}(t) \equiv 0$ and $(t - \frac{1}{2})^{2/3}y(t) \equiv 0$ admits only $y \equiv 0$ as a solution on any interval $[a,b] \subset [0,1]$, the normality assumptions in Theorem 3.9 are satisfied.

The Riccati equation ($\mathcal{R}$) and the boundary condition (3.17) are equivalent here to

$$\dot{W}(t) + \frac{1}{2}\left(t - \frac{1}{2}\right)^{2/3} W^2(t) = 0 \quad \text{and} \quad W(1) = \frac{10}{3}2^{5/3}. \tag{3.18}$$

The solution is $W(t) = -\frac{10}{3(t-\frac{1}{2})^{5/3}}$, which exists on $(\frac{1}{2}, 1]$ and not on $(0,1]$. Hence the conclusion of Corollary 3.10 fails. Therefore, by Corollary 3.10, $\bar{x} \equiv 0$ is not a $W^{1,s}$-weak local minimum for (P_0), for any $s \in [1, \infty]$.

4. Sufficient conditions for (P). In this section we first derive for the generalized problem of Bolza (P) a sufficiency criterion for strong local optimality. The conditions therein involve the "Weierstrass-type" condition, and the Riccati equation $(\mathcal{R})$ of Corollary 3.10 whose coefficients are given by (3.13), together with the (joint) boundary conditions

$$\Gamma + \begin{bmatrix} -W(0) & 0 \\ 0 & W(1) \end{bmatrix} > 0 \quad \text{on} \quad \left\{(\alpha,\beta) : \bar{M}\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = 0\right\}, \tag{4.1}$$

where $\Gamma = \nabla^2(\ell + \bar{\gamma}^T\phi)|_{(\bar{x}(0),\bar{x}(1))}$.

This result extends the corresponding results in [14] and [16] to the case when both endpoints of x vary (i.e., $\phi(x(0), x(1)) = 0$).

THEOREM 4.1. *Assume that L is $(\mathcal{L}\times\mathcal{B})$-measurable and that, for a vector $\bar{\gamma}\in\mathbb{R}^r$, $\bar{z} := (\bar{x},\bar{p})$ is a normal extremal for* (P) *at which* (H_2) *holds true. Suppose that*

(i) $L(t,\bar{x}(t),\dot{\bar{x}}(t)+v) - L(t,\bar{x}(t),\dot{\bar{x}}(t)) \geq \langle \bar{p}(t), v\rangle$ *for almost all* $t\in[0,1]$ *and for all* $v\in\mathbb{R}^n$;

(ii) *there exists a symmetric absolutely continuous solution on* $[0,1]$ *to the Riccati equation* $(\mathcal{R})$ *with the boundary conditions* (4.1).

Then $J(x)$ is well defined (possibly $+\infty$) near $\bar{x}$, and $\bar{x}$ provides a strict strong local minimum for (P). *Furthermore, there exist $\bar{\varepsilon}>0$ and $\bar{\delta}>0$ such that for all admissible arcs x satisfying $\|x-\bar{x}\|_\infty < \bar{\delta}$ we have*

$$J(x) - J(\bar{x}) \geq \bar{\varepsilon}\|x-\bar{x}\|_2^2.$$

Proof. Define $\bar{L}(t,x,v) := \sup\{\langle p,v\rangle - H(t,x,p) : p\in\mathbb{R}^n\}$. By hypothesis (H_2) and Lemma 2.5, it results that there exist $\bar{\delta}>0$ ($\bar{\delta}<\varepsilon$) and γ such that for $\|z-\bar{z}\|_\infty<\bar{\delta}$, (3.5) is satisfied. Hence, for $x\in W^{1,1}[0,1]$ with $\|x-\bar{x}\|_\infty<\bar{\delta}$,

$$\int_0^1 \bar{L}(t,x(t),\dot{x}(t))dt \geq \int_0^1 \langle \bar{p}(t),\dot{x}(t)\rangle dt - \gamma.$$

Since

$$L(t,x,v) \geq \bar{L}(t,x,v),$$

we conclude that $J(x)$ is well defined (possibly $+\infty$) near $\bar{x}$.

Now, by the embedding theorem of differential equations there exist $\varepsilon_0>0$ and $\bar{W}$ on $[0,1]$ such that

$$\dot{\bar{W}} + A^T\bar{W} + \bar{W}A + \bar{W}R\bar{W} + P = -\varepsilon_0 I \quad \text{for} \quad t\in[0,1] \text{ a.e.}$$

and

$$\Gamma + \begin{pmatrix} -\bar{W}(0) & 0 \\ 0 & \bar{W}(1) \end{pmatrix} > \varepsilon_0 I \quad \text{on} \quad \left\{(\alpha,\beta) : \bar{M}\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = 0\right\}. \tag{4.2}$$

Set

$$p(t,x) := \bar{p}(t) + \bar{W}(t)(x-\bar{x}(t)),$$

and

$$F(x(\cdot)) := \int_0^1 \Big\{ H(t,x(t),p(t,x(t))) - \bar{H}(t) + \langle \dot{\bar{p}}(t), x(t)-\bar{x}(t)\rangle - \langle \dot{\bar{x}}(t), \bar{W}(t)(x(t)-\bar{x}(t))\rangle + \frac{1}{2}\langle x(t)-\bar{x}(t), \dot{\bar{W}}(t)(x(t)-\bar{x}(t))\rangle \Big\} dt, \tag{4.3}$$

where $x(\cdot)$ is continuous. We have that F is twice Gateaux differentiable near $\bar{x}$ with first and second derivatives in the direction $h(\cdot)$:

$$\delta F(x(\cdot);h(\cdot)) = \int_0^1 \langle H_x(t,x(t),p(t,x(t))) + H_p(t,x(t),p(t,x(t)))\bar{W}(t) + \dot{\bar{p}}^T(t) - \dot{\bar{x}}^T(t)\bar{W}(t) + (x(t)-\bar{x}(t))^T\dot{\bar{W}}(t), h(t)\rangle dt,$$

and

$$\delta^2 F(x(\cdot);h(\cdot)) = \int_0^1 \langle \{H_{xx}(t,x(t),p(t,x(t))) + H_{xp}(t,x(t),p(t,x(t)))\bar{W}(t) + \bar{W}(t)H_{px}(t,x(t),p(t,x(t))) + \bar{W}(t)H_{pp}(t,x(t),p(t,x(t)))\bar{W}(t) + \dot{\bar{W}}(t)\}h(t), h(t)\rangle dt.$$

Also, using (3.1) and the definition of $\bar{W}$, it results that, for all continuous $h(\cdot)$,

$$F(\bar{x}(\cdot)) = 0, \quad \delta F(\bar{x}(\cdot);h(\cdot)) = 0, \quad \text{and}$$

$$\delta^2 F(\bar{x}(\cdot);h(\cdot)) = -\varepsilon_0 \int_0^1 |h(t)|^2 dt.$$

Hypothesis (H_2) yields that there exists $\delta_0 > 0$ such that for all continuous $x(\cdot)$ and $h(\cdot)$ satisfying $\|x-\bar{x}\|_\infty < \delta_0$ we have

$$\delta^2 F(x(\cdot);h(\cdot)) \le -\frac{\varepsilon_0}{2}\int_0^1 |h(t)|^2 dt.$$

Therefore, by [18, Theorem 40 A], for all continuous $x(\cdot) : \|x-\bar{x}\|_\infty < \delta_0$, we have

$$F(x(\cdot)) - F(\bar{x}(\cdot)) \le -\frac{\varepsilon_0}{4}\|x-\bar{x}\|_2^2. \tag{4.4}$$

Now define

$$Q(t,x) := \langle \bar{p}(t), x-\bar{x}(t)\rangle + \frac{1}{2}\langle x-\bar{x}(t), \bar{W}(t)(x-\bar{x}(t))\rangle. \tag{4.5}$$

Let $\|x-\bar{x}\|_\infty < \delta_0$ with x admissible for (P). Then, using (4.3), (4.5), (1.2), and condition (i) of the theorem, we obtain

$$\int_0^1 \{(L(t,x(t),\dot{x}(t)) - L(t,\bar{x}(t),\dot{\bar{x}}(t))\}dt \ge F(\bar{x}(\cdot)) - F(x(\cdot)) + \int_0^1 \frac{d}{dt}\{Q(t,x(t))\}dt.$$

Thus, from (4.4) and (4.5) and by integrating the last term, we get

$$\begin{aligned} J(x) - J(\bar{x}) \geq\ & \ell(x(0), x(1)) - \ell(\bar{x}(0), \bar{x}(1)) \\ & + \langle \bar{p}(1), x(1) - \bar{x}(1) \rangle + \frac{1}{2} \langle x(1) - \bar{x}(1), \bar{W}(1)(x(1) - \bar{x}(1)) \rangle \\ & - \langle \bar{p}(0), x(0) - \bar{x}(0) \rangle - \frac{1}{2} \langle x(0) - \bar{x}(0), \bar{W}(0)(x(0) - \bar{x}(0)) \rangle \\ & + \frac{\varepsilon_0}{4} \|x - \bar{x}\|_2^2. \end{aligned}$$

Consider the problem

$$\begin{aligned} \text{minimize } f(x, y) :=\ & \ell(x, y) - \ell(\bar{x}(0), \bar{x}(1)) \\ & + \langle \bar{p}(1), y - \bar{x}(1) \rangle + \frac{1}{2} \langle y - \bar{x}(1), \bar{W}(1)(y - \bar{x}(1)) \rangle \\ & - \langle \bar{p}(0), x - \bar{x}(0) \rangle - \frac{1}{2} \langle x - \bar{x}(0), \bar{W}(0)(x - \bar{x}(0)) \rangle \end{aligned}$$

over $\phi(x, y) = 0$.

Using (3.1)(iii) it follows that

$$\nabla f(\bar{x}(0), \bar{x}(1)) + \bar{\gamma}^T \bar{M} = 0.$$

Furthermore, (4.2) is equivalent to saying that

$$\nabla^2 (f + \bar{g}^T \phi) \Big|_{(\bar{x}(0), \bar{x}(1))} > 0 \quad \text{on} \quad \left\{ (\alpha, \beta) : \bar{M} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = 0 \right\}.$$

Hence, by standard sufficiency criterion for the finite dimensional optimization problem (e.g., [7, p. 307]), we conclude that $(\bar{x}(0), \bar{x}(1))$ is a local minimum for $f(x, y)$ over $\phi(x, y) = 0$. Thus there exists $\bar{\delta} \leq \delta_0$ such that for any admissible x with $\|x - \bar{x}\|_\infty < \bar{\delta}$ we have

$$J(x) - J(\bar{x}) \geq \bar{\varepsilon} \|x - \bar{x}\|_2^2,$$

where $\bar{\varepsilon} := \frac{\varepsilon_0}{4}$. □

Remark 4.1. When the endpoints cost and constraint are separable, we have $\Gamma = \left[\begin{smallmatrix} \Gamma_0 & 0 \\ 0 & \Gamma_1 \end{smallmatrix}\right]$ and $\bar{M} = \left[\begin{smallmatrix} \bar{M}_0 & 0 \\ 0 & \bar{M}_1 \end{smallmatrix}\right]$, and in this case, the boundary conditions (4.1) become

$$\text{(4.6)} \qquad \begin{cases} \Gamma_0 - W(0) > 0 \text{ on } \{\alpha \neq 0 \big| \bar{M}_0 \alpha = 0\}, \\ \\ \Gamma_1 + W(1) > 0 \text{ on } \{\alpha \neq 0 \big| \bar{M}_1 \alpha = 0\}. \end{cases}$$

Note that in this setting the proof of Theorem 4.1 remains valid when (4.6)(ii) is replaced by

$$W(1) = -\Gamma_1 - \beta \bar{M}_1$$

for some β. In fact, by the embedding theorem of differential equations, a solution of $(\mathcal{R})$ satisfying (4.6)(i) and the above condition leads to a solution of $(\mathcal{R})$ satisfying

(4.6). By symmetry, we can replace in Theorem 4.1 condition (4.6)(i) by $W(0) = \Gamma_0 + \beta\bar{M}_0$.

In the remainder of this section we study the connection between the sufficiency criterion presented in Theorem 4.1 and the necessary conditions obtained in section 3 in terms of the accessory problem (i.e., Theorem 3.3), the existence of a conjoined basis (i.e., Theorem 3.9), and a solution to the Riccati equation ($\mathcal{R}$) with boundary conditions (3.16) (i.e., Corollary 3.10). Our study below works for any L^1-matrix functions $A(\cdot), R(\cdot)$, and $P(\cdot)$ from $[0,1]$ to $\mathbb{R}^{n\times n}$ and matrices $\Gamma \in \mathbb{R}^{n\times n}$ and $\bar{M} \in \mathbb{R}^{r\times 2n}$ and not only for the choice given by (3.13) through the data of the problem (P).

Define

$$J_2(\eta, v) := \frac{1}{2}\left\langle \Gamma\begin{pmatrix}\eta(0)\\ \eta(1)\end{pmatrix}, \begin{pmatrix}\eta(0)\\ \eta(1)\end{pmatrix}\right\rangle + \frac{1}{2}\int_0^1 \{\langle R(t)v(t), v(t)\rangle - \langle P(t)\eta(t), \eta(t)\rangle\}dt.$$

In [17, Theorem 5.1] it is shown that the existence of a solution to ($\mathcal{R}$) in $[0,1]$ satisfying condition (4.1), which is condition (ii) of Theorem 4.1, is in fact sufficient for the coercivity in η of the quadratic form J_2; that is, for some $\gamma_0 > 0$,

$$J_2(\eta, v) \geq \frac{\gamma_0}{2}\int_0^1 |\eta(t)|^2 dt \tag{4.7}$$

for all (η, v) satisfying (2.4) and the boundary conditions $\bar{M}\binom{\eta(0)}{\eta(1)} = 0$. Condition (4.7) is a natural strengthening of the necessary condition in Theorem 3.3. However, for the general case of joint boundary conditions on the state, the existence of a solution to $\mathcal{R}$ and (4.1) is sufficient but not necessary for the coercivity of $J_2(\eta, v)$. This latter condition or its equivalent ones are the best to be used in the sufficiency criteria, since the gap between necessary and sufficient conditions would be minimal. This can also be said about the natural strengthenings of the necessary conditions given by Theorem 3.9 and Corollary 3.10.

Now consider the case of separable state endpoints and costs conditions. Hence we have

$$\bar{M} = \begin{bmatrix} \bar{M}_0 & 0 \\ 0 & \bar{M}_1 \end{bmatrix} \quad \text{and} \quad \Gamma = \begin{bmatrix} \Gamma_0 & 0 \\ 0 & \Gamma_1 \end{bmatrix}.$$

It is important to find out whether in this case the existence of a solution to ($\mathcal{R}$) on $[0,1]$ satisfying (4.6) is only sufficient for the coercivity in (4.7) or that the two conditions are rather equivalent. As we shall see below, under a certain controllability assumption, they are indeed equivalent. Observe that the coercivity of J_2 is in η and not in v, due to the lack of the strengthened Legendre–Clebsch hypothesis. However, under the latter hypothesis this result is known. It is worth mentioning that in [17, Theorem 5.2] a result parallel to Theorem 4.2 is obtained for the *positivity* of J_2 in η and *not* for its *coercivity*.

THEOREM 4.2. *Assume that the system* (2.4) *is either* $(I : \bar{M}_1)$*-normal on* $[c, 1]$ *for all* $c \in [0, 1)$ *or* $(\bar{M}_0 : I)$*-normal on* $[0, c]$ *for all* $c \in (0, 1]$. *Then the following are equivalent.*

(1) $J_2(\eta, v)$ *is coercive in* η *for all* (η, v) *solving* (2.4) *and* $\bar{M}_0\eta(0) = \bar{M}_1\eta(1) = 0$.

(2) *There exists a solution* (U, V) *of* (3.14) *satisfying*

(i) $U^T V = V^T U$ *on* $[0, 1]$,

(ii) $\det U(t) \neq 0$ *on* $[0,1]$,
(iii) $U^T(0)[\Gamma_0 U(0) - V(0)] > 0$ *on* $\{\alpha \neq 0 \mid \bar{M}_0 U(0)\alpha = 0\}$, *and*
(iv) $U^T(1)[\Gamma_1 U(1) + V(1)] > 0$ *on* $\{\alpha \neq 0 \mid \bar{M}_1 U(1)\alpha = 0\}$.

(3) *There exists a symmetric absolutely continuous solution* W *on* $[0,1]$ *of* $(\mathcal{R})$ *satisfying* (4.6).

Proof. By taking $W = VU^{-1}$ it follows that condition (2) implies (3). Also from [17, Theorem 5.1], condition (3) implies (1). Thus it remains to show that condition (1) implies condition (2). Assume that (2.4) is $(I : \bar{M}_1)$-normal on $[c,1]$ for all $c \in [0,1)$. By symmetry, a similar argument holds when the $(\bar{M}_0 : I)$-normality occurs. From [17, Theorem 5.5], condition (1) implies that the solution (U_1, V_1) of (3.14) and (3.15) satisfies conditions 2(i), 2(iii),

$$\det U(t) \neq 0 \quad \text{on} \quad [0,1),$$

and

$$U^T(1)[\Gamma_1 U(1) + V(1)] = 0.$$

By [5, Theorem 3.1.2(v)], there exists $\delta_0 > 0$ such that for all $\delta \in (0, \delta_0]$, the matrix

$$U_\delta := (I - \bar{M}_1) + \delta(-\Gamma_1(I - \bar{M}_1) - \bar{M}_1) \tag{4.8}$$

is invertible. Consider the system (3.14) with boundary conditions

$$\begin{cases} U(1) = U_\delta, \\ V(1) = V_1(1) + \delta(\Gamma_1\Gamma_1(I - \bar{M}_1) + I). \end{cases} \tag{4.9}$$

Since for $\delta = 0$ the system (3.14) and (4.9) has (U_1, V_1) as a solution on $[0,1]$ and since from (4.9) we have

$$U^T(1)[V(1) + \Gamma_1 U(1)] = \delta I + \delta^2 \bar{S}$$

for $\bar{S} = [(I - \bar{M}_1)\Gamma_1 + \bar{M}_1][\Gamma_1 \bar{M}_1 - I]$, then by the embedding theorem, there exists $\delta_1 > 0$ ($\delta_1 < \delta_0$) such that for all $\delta \in (0, \delta_1]$ there exists a solution $(U_\delta(\cdot), V_\delta(\cdot))$ of (3.14) and (4.9) satisfying parts (i), (iii), and (iv) of condition (2).

It remains to show that for some $\delta \in (0, \delta_1]$ the solution to (3.14) and (4.9) also satisfies part (ii) of condition (2). Note that since U_δ, given by (4.8), is invertible, then for each $\delta \in (0, \delta_1]$ there exists $t_\delta \in [0,1)$ (minimal) such that $\det U_\delta(t) \neq 0$ on $(t_\delta, 1)$. Now suppose that our claim is not true; then there exists $t_m \in [0,1)$ (minimal) such that

$$\det U_{1/m}(t) \neq 0 \quad \text{on} \quad (t_m, 1] \quad \text{and } \det U_{1/m}(t_m) = 0.$$

Using (4.9) it follows that $(U_{1/m}, V_{1/m}) \underset{m\to\infty}{\longrightarrow} (U_1, V_1)$. However, since U_1 is invertible on $[0,1)$, we conclude that $t_m \to 0$. Let $\alpha_m \in \mathbb{R}^n$ with $|\alpha_m| = 1$ and

$$U_{1/m}(t_m)\alpha_m = 0. \tag{4.10}$$

By passing to a convergent subsequence, say, $\alpha_m \underset{m\to\infty}{\longrightarrow} \alpha$, and upon taking the limit in (4.10), it results that

$$U_1(0)\alpha = 0 \quad \text{and} \quad |\alpha| = 1.$$

This contradicts that $\det U_1(0) \neq 0$. Therefore, the proposition is proved. □

Note that in this setting, condition (ii) of Theorem 4.1 can be replaced by any of the equivalent conditions in Theorem 4.2, as long as the normality assumption of the theorem holds. Another consequence of this theorem is the following connection between condition (ii) of Theorem 4.1, that is, condition (3) of Theorem 4.2, and the necessary conditions in Theorem 3.9 and Corollary 3.10.

COROLLARY 4.3. *Assume that system* (2.4) *is* $(I : \bar{M}_1)$*-normal on* $[c, 1]$ *for all* $c \in [0, 1)$. *Then each of the three equivalent conditions of Theorem* 4.2 *is also equivalent to each of the following conditions.*

(a) *The solution* (U_1, V_1) *of* (3.14) *and* (3.15) *satisfies* (i) $U^T V = V^T U$, (ii) $\det U(t) \neq 0$ *on* $[0, 1)$, (iii) $U^T(0)(\Gamma_0 U(0) - V(0)) > 0$ *on* $\{\alpha \neq 0 \mid \bar{M}_0 U(0)\alpha = 0\}$, *and* $V(1) + \Gamma_1(I - \bar{M}_1) + \bar{M}_1 = 0$.

(b) *There exists a symmetric solution* W *to* $(\mathcal{R})$ *on* $[0, 1)$ *satisfying* (4.6)(i) *and*

$$\lim_{t \to 1^-} W(t)U_1(t) = -\Gamma_1(I - \bar{M}_1) - \bar{M}_1.$$

Proof. From [17, Theorem 5.2] we know that conditions (a) and (b) above are equivalent. Also, in the proof of Theorem 4.2 we showed the following implications:

$$(1) \Rightarrow \text{condition (a)} \Rightarrow (2) \Rightarrow (3) \Rightarrow (1),$$

which proves the corollary. □

Remark 4.2. The above result states that in Theorem 4.1 we can replace condition (ii) by either of conditions (a) or (b) of Corollary 4.3, since they imply the $(I : \bar{M}_1)$-normality required. Hence Corollary 4.3 shows that the sufficiency criterion of Theorem 4.1 is obtained as natural strengthening of the necessary conditions in either Theorem 3.9 or Corollary 3.10.

5. Application to optimal control problems. In this section we consider the optimal control problem (C) defined in the introduction. First, we intend to apply the results of the previous section to the generalized Bolza ($\mathrm{P_C}$) problem defined in section 2 and associated with (C), and hence sufficient conditions for strong local minimality in (C) will be obtained. Next, we plan to derive second-order necessary conditions for optimality in (C). In this regard we shall obtain results parallel to Theorem 3.3, Theorem 3.9, and Corollary 3.10 by either applying Theorem 3.3 to the generalized problem of Bolza ($\mathrm{P_C}$) associated to (C), or by using Theorem 3.7 directly. The former approach yields necessary conditions for $W^{1,s}$-local minimum and requires assuming condition $(\mathcal{B})$, given in section 2, and

$$(\mathcal{L}_C) \qquad \text{for } (t, x) \in T(\bar{x}; \varepsilon), \{(f(t, x, u), g(t, x, u) + r) : u \in U,\ r \geq 0\}$$

is convex and closed.

On the other hand, the second approach produces necessary conditions for $W^{1,s}$- and L^s-weak local minimality, where $s \in [1, \infty]$. For the first type of optimality this approach requires only that the supremum in (1.4) be attained at some $\mathbf{u}(t, x, p)$, while for the second type of optimality, regularity assumptions are needed on $\mathbf{u}$. Of course, all the results of this section apply for the (classical) strong local minimality in (C).

Assume throughout this section that f, g, ℓ, and U satisfy assumption (A), given in section 2; for $t \in [0, 1]$ and $u \in U$, $f(t, \cdot, u)$ and $g(t, \cdot, u)$ are differentiable on $B_n(\bar{x}(t); \varepsilon)$; and ℓ is differentiable on $B_{2n}(\bar{x}(0), \bar{x}(1); \varepsilon)$.

Let $(\bar{x}, \bar{u}, \bar{p}, \bar{\gamma})$ be given. We say that $(\bar{x}, \bar{u}, \bar{p}, \bar{\gamma})$ satisfies the *normal* version of the Pontryagin maximum principle if they satisfy

$$\text{(5.1)} \qquad \begin{cases} -\dot{\bar{p}} = f_x^T(t, \bar{x}(t), \bar{u}(t))\bar{p}(t) - g_x^T(t, \bar{x}(t), \bar{u}(t)), \\ (\bar{p}^T(0), -\bar{p}^T(1)) = \nabla \ell(\bar{x}(0), \bar{x}(1)) + \bar{\gamma}^T \bar{M}, \\ \max\{\bar{p}^T(t) f(t, \bar{x}(t), u) - g(t, \bar{x}(t), u) \mid u \in U\} \\ \text{is attained at } \bar{u}(t), \text{ for almost all } t. \end{cases}$$

Let the Hamiltonian $\mathcal{H}(t, \cdot, \cdot)$ be differentiable at $(\bar{x}, \bar{p})$. If $(\bar{x}, \bar{u})$ is admissible for (C) and (5.1) holds, then $(\bar{x}, \bar{u}, \bar{p}, \bar{\gamma})$ satisfies

$$\begin{aligned} -\dot{\bar{p}}(t) &= \bar{\mathcal{H}}_x(t) \quad \text{a.e.}, \\ \dot{\bar{x}}(t) &= \bar{\mathcal{H}}_p(t) \quad \text{a.e.}, \\ (\bar{p}^T(0), -\bar{p}^T(1)) &= \nabla \ell(\bar{x}(0), \bar{x}(1)) + \bar{\gamma}^T \bar{M}, \\ \mathcal{H}(t, \bar{x}(t), \bar{p}(t)) &= \langle \bar{p}(t), f(t, \bar{x}(t), \bar{u}(t)) \rangle - g(t, \bar{x}(t), \bar{u}(t)) \quad \text{a.e.} \end{aligned}$$

(see, e.g., [2, Proposition 3.2]).

The following result is an application of Theorem 4.1 to the optimal control setting. It extends the corresponding result in [14], [15], and [2] to the case where $\bar{x}$ is not necessarily C^1 and both endpoints of x vary (i.e., $\phi(x(0), x(1)) = 0$). When we specialize to the case considered in [2], that is, a fixed initial and a free final state constraints, our result assumes less regularity assumptions on the Hamiltonian $\mathcal{H}$ defined by (1.4) and on $\mathcal{H}_p$.

THEOREM 5.1. *Let $(\bar{x}, \bar{u})$ be admissible for* (C)*, $\bar{p} \in W^{1,1}[0,1]$ and $\bar{\gamma} \in \mathbb{R}^r$ such that $(\bar{x}, \bar{u}, \bar{p}, \bar{\gamma})$ satisfies the normal version of the Pontryagin maximum principle* (5.1)*. Let $\mathcal{H}$, the Hamiltonian of* (C)*, satisfy* (H_2) *for some $\varepsilon > 0$. Assume that there exists a symmetric, absolutely continuous solution on $[0,1]$ of* (4.1) *and the Riccati equation* ($\mathcal{R}$)*, where H is replaced by $\mathcal{H}$.*

Then there exist $\bar{\varepsilon} > 0$ and $\bar{\delta} > 0$ such that for all admissible pairs (x, u) with $\|x - \bar{x}\|_\infty < \bar{\delta}$ we have

$$J(x, u) \geq J(\bar{x}, \bar{u}) + \bar{\varepsilon} \|x - \bar{x}\|_2^2,$$

and $(\bar{x}, \bar{u})$ is a strong local minimum for (C)*.*

Proof. Define L_C via (2.2), where $T(\bar{x}; \varepsilon)$ is used. In order to prove this theorem, we use Remark 2.2. The maximality condition in (5.1) yields that

$$L_C(t, \bar{x}(t), \dot{\bar{x}}(t)) = g(t, \bar{x}(t), \bar{u}(t)) \quad \text{a.e.}$$

Let us show that $\bar{x}$ is a strong local minimum for (P_C). Again the maximum principle yields that

$$L_C(t, \bar{x}(t), \dot{\bar{x}}(t) + v) - L_C(t, \bar{x}(t), \dot{\bar{x}}(t)) \geq \langle \bar{p}(t), v \rangle$$

for almost all t and for all $v \in \mathbb{R}^n$. Equations (5.1) yield that $(\bar{x}, \bar{p})$ satisfies (3.1), where $H := \mathcal{H}$. Since the Hamiltonian $\mathcal{H}$ of the control problem (C) is the same as the Hamiltonian corresponding to (P_C), all the conditions of Theorem 4.1 are satisfied

for the problem ($\mathrm{P_C}$). It results that there exist $\bar\varepsilon > 0$ and $\bar\delta > 0$ such that for any x admissible for ($\mathrm{P_C}$) with $\|x - \bar x\|_\infty < \bar\delta$ we have

$$J_C(x) - J_C(\bar x) \ge \bar\varepsilon \|x - \bar x\|_2^2.$$

For (x, u) admissible for (C) Remark 2.2 implies that

$$J_C(x) \le J(x, u).$$

We also know that

$$J_C(\bar x) = J(\bar x, \bar u);$$

hence $(\bar x, \bar u)$ is a strong local minimum for (C) and for all admissible pairs (x, u) with $\|x - \bar x\|_\infty < \bar\delta$ we have

$$J(x, u) - J(\bar x, \bar u) \ge \bar\varepsilon \|x - \bar x\|_2^2. \qquad \square$$

Remark 5.1. Consider the case where the state endpoint costs and constraints are separable, that is,

$$\ell(x, y) := \ell_0(x) + \ell_1(y) \quad \text{and} \quad \phi(x, y) := \begin{pmatrix} \phi_0(x) \\ \phi_1(y) \end{pmatrix}. \tag{5.2}$$

Then Theorem 4.2 and Corollary 4.3 yield that the result of Theorem 5.1 remains valid for problem (C) when the assumption regarding the existence of a solution to $(\mathcal{R})$ and (4.1) is replaced by either condition (a) or condition (b) of Corollary 4.3, or by any of the equivalent conditions given in Theorem 4.2, provided that the normality hypothesis of that theorem holds.

Now we intend to derive necessary conditions for (C). This is done either by ensuring that the problem (C), the function $\mathrm{L_C}$, and the Hamiltonian $\mathcal{H}$ satisfy assumptions $(\bar A)$, $(\mathcal{L})$, and (H_2), and hence Theorem 3.3 will apply to ($\mathrm{P_C}$), or by connecting the two optimal control problems (C) and ($\mathrm{C}_\mathcal{H}$), and then ensuring that $H := \mathcal{H}$ satisfies the conditions of Theorem 3.7. Here ($\mathrm{C}_\mathcal{H}$) is exactly the problem ($\mathrm{C}_\mathcal{H}$) introduced in section 3 with $H := \mathcal{H}$. The first method leads to part (1)(i) of the theorem, and the second method is behind the rest of the theorem.

THEOREM 5.2. *Let f, g, and U satisfy* (A). *Suppose that $(\bar x, \bar u, \bar p, \bar\gamma)$ and $\mathcal{H}$ satisfy* (5.1) *and* (H_2). *Then for $H := \mathcal{H}$, the conclusions of Theorem* 3.3 *hold and Theorem* 3.9 *and Corollary* 3.10 *are valid if either of the conditions* (1) *or* (2) *is satisfied.*

(1) *For $s \in [1, \infty]$, $(\bar x, \bar u)$ is a $W^{1,s}$-weak local minimum with* (H_p^s) *satisfied, and either condition* (i) *or condition* (ii) *is in effect.*

(i) *Assumptions* $(\mathcal{L}_C)$ *and* $(\mathcal{B})$ *hold.*

(ii) *The supremum in* (1.4) *is attained for $(t, x, p) \in T(\bar x, \bar p; \varepsilon)$.*

(2) *For $s \in [1, \infty]$, $(\bar x, \bar u)$ is a classically L^s-weak local minimum with* (H_p^s) *satisfied, and the supremum in condition* (ii) *is attained at $\mathbf{u}(t, x, p)$ such that $\mathbf{u}(t, \bar x(t), \bar p(t)) = \bar u(t)$, and $\mathbf{u}$ is continuous at $\bar z := (\bar x, \bar p)$ from L^∞ to L^s.*

Remark 5.2. When U is compact, $(\mathcal{B})$ and condition (ii) of Theorem 5.2 hold automatically, and hence condition (i) is more restrictive than (ii). Thus, in this case, part (1) of the theorem disposes of conditions (i) and (ii). This shows that in this case the second-order necessary conditions obtained for (C) by passing through the associated generalized Bolza problem ($\mathrm{P_C}$) are weaker than those obtained by passing through ($\mathrm{C}_\mathcal{H}$).

The existence of a feedback-type function $\mathbf{u}$ as in part (2) of the theorem has been employed in [13] when deriving sufficient conditions for the strong local minimum.

Proof. Assumptions (A) yield that $\mathcal{H}$ is $(\mathcal{L}\times\mathcal{B})$-measurable. By Remark 2.1, L_C is also $(\mathcal{L}\times\mathcal{B})$-measurable.

For (1), assume that (i) holds. Then, by Remark 2.3, assumptions $(\bar{A})$ are satisfied. Thus, by applying Theorem 2.7, it results that $\bar{x}$ is $W^{1,s}$-weak local for $(\mathrm{P_C})$. We shall check whether the assumptions of Theorem 3.3 are met by $(\mathrm{P_C})$. As mentioned earlier, (5.1) yields that $(\bar{x},\bar{p},\bar{\gamma})$ is a normal extremal for $(\mathrm{P_C})$, whose Hamiltonian is $\mathcal{H}$. Assumption $(\mathcal{L}_{\mathrm{C}})$ implies that $L_C(t,x,\cdot)$ is convex, since for $(t,x)\in T(\bar{x};\varepsilon)$ the set defined in $(\mathcal{L}_{\mathrm{C}})$ is the epigraph of $L_C(t,x,\cdot)$. Thus Assumption $(\mathcal{L})$ is satisfied, whence the conclusions of Theorem 3.3 hold. Assume that ϕ and ℓ are separable, as in (5.2); then, under the normality assumptions therein, the results of Theorem 3.9 and Corollary 3.10 are valid.

To prove parts (1)(ii) and (2) of the theorem we shall use the following result that connects (C) with $(\mathrm{C}_{\mathcal{H}})$.

PROPOSITION 5.3. *Assume that f,g, and U satisfy* (A). *Suppose that condition* (ii) *of Theorem* 5.2 *holds and that $\mathcal{H}(t,\cdot,\cdot)$ is differentiable on $B_{2n}(\bar{x}(t),\bar{p}(t);\varepsilon)$. Let $(\bar{x},\bar{u},\bar{p})$ satisfy* (5.1)(iii). *If for $s\in[1,\infty]$, (H_p^s) holds and*

(a) *$(\bar{x},\bar{u})$ is a $W^{1,s}$-weak local minimum for* (C), *or*

(b) *$(\bar{x},\bar{u})$ is a classically L^s-weak local minimum and the supremum in condition* (ii) *of Theorem* 5.2 *is attained at some $\mathbf{u}(t,x,p)$ such that $\mathbf{u}(t,\bar{x}(t),\bar{p}(t))=\bar{u}(t)$, and $\mathbf{u}$ is continuous at $(\bar{x},\bar{p})$ from L^∞ to L^s,*

then $(\bar{x},\bar{p})$ is classically an L^∞-weak local minimum for $(C_{\mathcal{H}})$.

Proof. For (a), suppose that there exists $0<\varepsilon_0<\varepsilon$ such that $J(x,u)\geq J(\bar{x},\bar{u})$ for all admissible pairs (x,u) with $\|x-\bar{x}\|_{1,s}<\varepsilon_0$. Let (x,p) be admissible for $(\mathrm{C}_{\mathcal{H}})$ with $(x,p)\in T(\bar{x},\bar{p};\varepsilon)$ and $\|x-\bar{x}\|_{1,s}<\varepsilon_0$. Condition (ii) of Theorem 5.2 and the differentiability of $\mathcal{H}(t,\cdot,\cdot)$ yield that

$$\mathcal{H}_p(t,x(t),p(t))=f(t,x(t),A(t,x(t),p(t))),$$

where

$$A(t,x,p):=\{u\in U:\langle p,f(t,x,u)\rangle-g(t,x,u)=\mathcal{H}(t,x,p)\},$$

and $f(t,x,A(t,x,p))$ is a singleton. Hence, by the measurable selection theorem, there exists a measurable function u such that $u(t)\in A(t,x(t),p(t))$ for almost all t. Thus

$$\dot{x}(t)=f(t,x(t),u(t))=\mathcal{H}_p(t,x(t),p(t))\quad\text{a.e.,}\tag{5.3}$$

and

$$\langle p(t),f(t,x(t),u(t))\rangle-g(t,x(t),u(t))=\mathcal{H}(t,x(t),p(t))\quad\text{a.e.}\tag{5.4}$$

It follows that (x,u) is admissible for (C), and then

$$J(x,u)\geq J(\bar{x},\bar{u}).$$

However, using (5.3) and (5.4), we get

$$J_C(x,p)=J(x,u)\quad\text{and}\quad J_C(\bar{x},\bar{p})=J(\bar{x},\bar{u}).$$

Therefore, $(\bar{x},\bar{p})$ is a $W^{1,s}$-weak local minimum for $(\mathrm{C}_{\mathcal{H}})$. Now applying Lemma 2.6 to the problem $(\mathrm{C}_{\mathcal{H}})$, we conclude the result.

For (b), suppose for $0 < \varepsilon_0 < \varepsilon$ we have $J(x,u) \geq J(\bar{x},\bar{u})$ for $\|x-\bar{x}\|_\infty + \|u-\bar{u}\|_s < \varepsilon_0$. By continuity of $\mathbf{u}$, there exists $0 < \delta_0 < \frac{\varepsilon_0}{2}$ such that for $\|x-\bar{x}\|_\infty + \|p-\bar{p}\|_\infty < \delta_0$, we have

$$\|\mathbf{u}(\cdot, x(\cdot), p(\cdot)) - \mathbf{u}(\cdot, \bar{x}(\cdot), \bar{p}(\cdot))\|_s < \varepsilon_0/2.$$

Now let (x,p) be such that $\|x-\bar{x}\|_\infty + \|p-\bar{p}\|_\infty < \delta_0$ and (x,p) is admissible for $(\mathrm{C}_\mathcal{H})$. In this case, $\mathbf{u}(t,x(t),p(t))$ belongs to $A(t,x(t),p(t))$ defined above, and hence, for $u(\cdot) := \mathbf{u}(\cdot, x(\cdot), p(\cdot))$, (x,u) is admissible for (C). Since

$$\|x-\bar{x}\|_\infty + \|u-\bar{u}\|_s < \delta_0 + \frac{\varepsilon_0}{2} < \varepsilon_0,$$

it results that $J(x,u) \geq J(\bar{x},\bar{u})$. Arguments similar to those used for (a) yield that $J_\mathrm{C}(x,p) \geq J_\mathrm{C}(\bar{x},\bar{p})$. That is, $(\bar{x},\bar{p})$ is an L^∞-weak local minimum for $(\mathrm{C}_\mathcal{H})$. □

Return to the proof of part (1) of Theorem 5.2. Assume condition (ii) of the theorem holds. Then Proposition 5.3(a) and Theorem 3.7 yield that the conclusions of Theorem 3.3 hold, and thus each of Theorem 3.9 and Corollary 3.10 is valid. Part (2) of the theorem follows from Proposition 5.3(b) and Theorem 3.7. □

Remark 5.3. When the initial state value is fixed and the final state value is free, the problem (C) reduces to the one studied in [2]. In this case, ℓ and ϕ in (5.2) take the form $\ell(x,y) = \ell_1(y)$ and $\phi(x,y) = \phi_0(x) = x - x_0$. In this special setting necessary conditions in terms of the Riccati equation $(\mathcal{R})$ were derived in [2, Theorem 5.1] for the $W^{1,1}$-weak local minimum. On the other hand, specialized to this setting, part (1) of Theorem 5.2 here yields that, under hypotheses (A), (5.1), (H_2), and condition (ii) of Theorem 5.2, Corollary 3.10 holds true. Hence, by Remark 3.2, these assumptions yield necessary conditions in terms of a solution on $(0,1]$ to $(\mathcal{R})$ and (3.17). However, these assumptions are considerably weaker than those imposed in [2, Theorem 5.1]. This is so since we do *not* assume (1.5), $\nabla_z^2 \bar{\mathcal{H}}(\cdot)$ to be L^∞, $\nabla_z^2 \mathcal{H}$ to be Lipschitz near $\bar{z} = (\bar{x},\bar{p})$ from L^∞ to L^1, nor the strengthened Legendre–Clebsch condition $(S\mathcal{L})$ to hold. Also, our necessary conditions apply for the $W^{1,s}$-, or classically L^s-weak local minimum, where $s \in [1,\infty]$, and not only for the $W^{1,1}$-weak local minimum.

As we shall see below, the assumption $(S\mathcal{L})$ in [2] is very strong. In fact, it renders the corresponding function L_C to be locally C^1 in (x,v), and hence, the control problem (C) resembles a calculus of variations problem.

LEMMA 5.4. *Let $(\bar{x},\bar{p}) \in W^{1,\infty}[0,1] \times L^\infty[0,1]$ satisfy $\dot{\bar{x}}(t) = \bar{\mathcal{H}}_p(t)$ a.e., and let $\mathcal{H}$ satisfy (H_2), (H_p^∞), and $(S\mathcal{L})$. Assume that $\mathcal{H}_{pp}$ is continuous at $(\bar{x},\bar{p})$, from L^∞ to L^∞, and $\bar{\mathcal{H}}_p(\cdot)$ and $\mathcal{H}_{pp}(\cdot)$ are in L^∞.*

If L_C, defined by (2.2), satisfies $(\mathcal{L})$, then L_C is $(\mathcal{L}\times\mathcal{B})$-measurable and lower semicontinuous in (x,v), and there exists $\tilde{\varepsilon} > 0$ $(\tilde{\varepsilon} \leq \varepsilon)$ and $p(t,x,v)$ such that $p(t,\cdot,\cdot)$ is C^1 on $T(\bar{x},\dot{\bar{x}};\tilde{\varepsilon})$ and ∇p is continuous at $(\bar{x},\dot{\bar{x}})$, from L^∞ to L^∞, such that

$$L_C(t,x,v) = p(t,x,v)\cdot v - H(t,x,p(t,x,v)).$$

Proof. Since $\mathcal{H}$ is also the Hamiltonian corresponding to L_C, which is convex in v,

$$L_C(t,x,v) = \sup\{\langle p,v\rangle - \mathcal{H}(t,x,p) \mid p \in \mathbb{R}^n\}$$

for $(t,x) \in T(\bar{x};\varepsilon)$. Hence by [9], L_C is $(\mathcal{L}\times\mathcal{B})$-measurable and lower semicontinuous in (x,v). Consider the equation

$$v = \mathcal{H}_p(t,x,p). \tag{5.5}$$

We have that (5.5) is satisfied for almost all t, for $v = \dot{\bar{x}}(t)$, for $x = \bar{x}(t)$, and for $p = \bar{p}(t)$. Set

$$N(z(\cdot);\varepsilon) := \{x(\cdot) \in L^\infty[0,1] \mid \|z - x\|_\infty < \varepsilon\},$$

and define

$$\mathcal{F} : N(\bar{x};\varepsilon) \times N(\bar{p};\varepsilon) \to L^\infty[0,1],$$

$$(x(\cdot), p(\cdot)) \to v(\cdot) = \mathcal{F}(x(\cdot), p(\cdot)),$$

where

$$\mathcal{F}(x(\cdot), p(\cdot))(t) = \mathcal{H}_p(t, x(t), p(t)).$$

Consider the equation

$$v(\cdot) = \mathcal{F}(x(\cdot), p(\cdot)).$$

(i) For $x(\cdot) \in N(\bar{p};\varepsilon)$, $\mathcal{F}(x(\cdot),\cdot)$ is differentiable on $N(\bar{p};\varepsilon)$.

For, let $p(\cdot) \in N(\bar{p};\varepsilon)$ and $\bar{\varepsilon} > 0$ given. The continuity of $\mathcal{H}_{pp}$ from L^∞ to L^∞ yields the existence of $\delta > 0$ such that for $\|q - p\|_\infty < \delta$ we have

$$|\mathcal{H}_{pp}(t, x(t), q(t)) - \mathcal{H}_{pp}(t, x(t), p(t))| < \frac{\bar{\varepsilon}}{2} \text{ a.e.}$$

Let $\bar{q}(\cdot) \in N(p(\cdot);\delta)$. We shall show

$$\|\mathcal{F}(x(\cdot), \bar{q}(\cdot)) - \mathcal{F}(x(\cdot), p(\cdot)) - \mathcal{F}_p(x(\cdot), p(\cdot))(\bar{q}(\cdot) - p(\cdot))\|_\infty \le \bar{\varepsilon}\|\bar{q} - p\|_\infty.$$

By the mean value theorem for vector-valued functions applied to $\mathcal{H}_p(t, x(t), \cdot) - \mathcal{H}_{pp}(t, x(t), p(t))(\cdot)$ on the line segment joining $p(t)$ and $\bar{q}(t)$, there exists $\tilde{q}$ with $|\tilde{q}(t) - p(t)| < \delta$ such that

$$\begin{aligned}
&|\mathcal{H}_p(t, x(t), \bar{q}(t)) - \mathcal{H}_p(t, x(t), p(t)) - \mathcal{H}_{pp}(t, x(t), p(t))(\bar{q}(t) - p(t))| \\
&\quad \le \big|\big(\mathcal{H}_{pp}(t, x(t), \tilde{q}(t)) - \mathcal{H}_{pp}(t, x(t), p(t))\big)(\bar{q}(t) - p(t))\big| \\
&\quad \le \frac{\bar{\varepsilon}}{2}\|\bar{q} - p\|_\infty.
\end{aligned}$$

Set

$$\mathcal{F}_p(x(\cdot), p(\cdot))(t) := \mathcal{H}_{pp}(t, x(t), p(t)).$$

The result follows.

(ii) Let $\mathcal{B}(L^\infty, L^\infty)$ be the normed space of all bounded linear operators from L^∞ into L^∞. Then

$$\mathcal{F}_p : N(\bar{x};\varepsilon) \times N(\bar{p};\varepsilon) \to \mathcal{B}(L^\infty, L^\infty)$$

is continuous, since $\mathcal{H}_{pp}(t,\cdot,\cdot)$ is continuous on $N(\bar{x};\varepsilon) \times N(\bar{p};\varepsilon)$, L^∞-uniformly in t.

(iii) $\mathcal{F}_p(\bar{x},\bar{p})$ is a homeomorphism of L^∞ onto L^∞. The strengthened Legendre condition $(S\mathcal{L})$ yields that $\mathcal{F}_p(\bar{x},\bar{p})$ is a bijection and that $\bar{\mathcal{H}}_{pp}^{-1}(t)$ is essentially bounded. Since $\bar{\mathcal{H}}_{pp}(\cdot)\in L^\infty[0,1]$, we get that $\mathcal{F}_p(\bar{x},\bar{p})$ is a homeomorphism.

(i)–(iii) allow us to apply the implicit function theorem [11, II.3.8] to deduce the existence of $\varepsilon_0,\varepsilon_1>0$ and a unique C^1-function $\mathcal{P}$ such that

$$\mathcal{P}: N(\bar{x},\varepsilon_0)\times N(\dot{\bar{x}},\varepsilon_0)\to N(\bar{p},\varepsilon_1),$$

$$v(\cdot)=\mathcal{F}(x(\cdot),\mathcal{P}(x(\cdot),v(\cdot))),$$

and

$$\mathcal{P}(\bar{x},\dot{\bar{x}})=\bar{p}.$$

Let us construct $p(t,x,v)$ on $T(\bar{x},\dot{\bar{x}};\frac{\varepsilon_0}{2})$.

For given $(t,x,v)\in T(\bar{x},\dot{\bar{x}};\frac{\varepsilon_0}{2})$, set

$$x(s):=\bar{x}(s)+(\dot{\bar{x}}(t)-v)(t-s)+x-\bar{x}(t).$$

We have $x(t)=x$, and $\dot{x}(t)=v$. Define

$$p(t,x,v):=\mathcal{P}(x(\cdot),\dot{x}(\cdot))(t).$$

Then, on $T(\bar{x},\dot{\bar{x}};\frac{\varepsilon_0}{2})$, $p(\cdot,\cdot,\cdot)$ is well defined with values in $N(\bar{p},\varepsilon_1)$,

$$v=\mathcal{H}_p(t,x,p(t,x,v)), \tag{5.6}$$

and

$$p(t,\bar{x}(t),\dot{\bar{x}}(t))=\bar{p}(t).$$

Condition $(S\mathcal{L})$ yields that $p(\cdot,\cdot,\cdot)$ is unique. The C^1-property of $\mathcal{P}$ on $N(\bar{x};\varepsilon_0)\times N(\dot{\bar{x}},\varepsilon_0)$ implies that p is continuously differentiable from L^∞ to L^∞. Thus, by the convexity of $\mathcal{H}(t,x,\cdot)$, for (t,x,v) in $T(\bar{x},\dot{\bar{x}};\frac{\varepsilon_0}{2})$,

$$L_C(t,x,v)=p(t,x,v)\cdot v-\mathcal{H}(t,x,p(t,x,v)),$$

whence the result follows. □

The following example shows the power of the results of this section, as it produces information where previous work offered none.

Example 5.1. Consider the optimal control problem

$$(\mathrm{C}_0)\qquad \text{minimize } J(x,u):=-\frac{5}{3}2^{5/3}x^2(1)+\int_0^1 (x^3(t)+u^2(t))dt$$

subject to

$$\dot{x}(t)=\left(t-\frac{1}{2}\right)^{1/3}u(t),$$

$$x(0)=0,$$

$$u(t)\in\mathbb{R}\quad \text{a.e.}$$

The Hamiltonian corresponding to (C_0) is

$$\mathcal{H}(t,x,p) := \sup_{u\in\mathbb{R}} \left\{ p\left(t-\frac{1}{2}\right)^{1/3} u - x^3 - u^2 \right\}$$
$$= -x^3 + \frac{p^2}{4}\left(t-\frac{1}{2}\right)^{2/3},$$

and hence the supremum in $\mathcal{H}$ is attained for all $(t,x,p)\in[0,1]\times\mathbb{R}\times\mathbb{R}$ at

$$\mathbf{u}(t,x,p) = \frac{p}{2}\left(t-\frac{1}{2}\right)^{1/3}$$

with $\mathbf{u}$ continuous everywhere from $L^\infty \times L^\infty$ to L^∞. Moreover, $\mathcal{H}$ is the same Hamiltonian H of the problem (P_0) in Example 3.1.

Let $\bar{x}\equiv 0, \bar{u}\equiv 0, \bar{p}\equiv 0$, and $\bar{\gamma} = \binom{0}{0}$. Then $(\bar{x},\bar{p},\bar{u},\bar{\gamma})$ satisfy the normal version of the Pontryagin maximum principle (5.1). Furthermore, assumptions (A) and (H_2) hold, and, as seen in Example 3.1, the normality requirements of Theorem 3.9 are satisfied.

The Riccati equation $(\mathcal{R})$ for this problem with the boundary condition (3.17) is exactly given by (3.18). The solution $W(t) = -\frac{10}{3(t-\frac{1}{2})^{5/3}}$ exists on $(\frac{1}{2},1]$ and cannot be extended (by continuity) to $[\frac{1}{2},1]$. Even though (1.4) holds, the strengthened Legendre–Clebsch condition $(S\mathcal{L})$ required in the necessity theorem [2, Theorem 5.1] fails to hold near $t_c = \frac{1}{2}$, since

$$\bar{H}_{pp}(t) = \frac{1}{2}\left(t-\frac{1}{2}\right)^{2/3} \Rightarrow \inf_{\alpha\le t<t_c} \bar{H}_{pp}(t) = 0 \qquad \forall\ \ \alpha\in[0,t_c).$$

Therefore, the result in [2, Theorem 5.1] does not apply to this problem, and no information can be produced from [2] about the optimality of $(\bar{x},\bar{u})$ for (C_0).

On the other hand, as seen in Example 3.1, the conclusion of Corollary 3.10 does not hold since the solution W does not exist on $(0,1]$. Given that condition (ii) of part (1) of Theorem 5.2 is satisfied, it results from applying Theorem 5.2(1) that $(\bar{x},\bar{u})$ is not a $W^{1,s}$-weak local minimum for (C_0), for any $s\in[1,\infty]$. Since the assumptions for part (2) of Theorem 5.2 hold, we also conclude that $(\bar{x},\bar{u})$ is not a L^s-weak local minimum for (C_0), for any $s\in[1,\infty]$.

Acknowledgment. The author wishes to thank a referee for the thorough reading of the paper and for valuable remarks and comments that helped improve the presentation of the results.

REFERENCES

[1] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser Boston, Boston, 1990.
[2] N. Caroff and H. Frankowska, *Conjugate points and shocks in nonlinear optimal control*, Trans. Amer. Math. Soc., 348 (1996), pp. 3133–3153.
[3] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Canad. Math. Soc. Series Monographs Adv. Texts, John Wiley, New York, 1983.
[4] F. H. Clarke, *Hamiltonian analysis of the generalized problem of Bolza*, Trans. Amer. Math. Soc., 301 (1987), pp. 385–400.
[5] W. Kratz, *Quadratic Functionals in Variational Analysis and Control Theory*, Akademie-Verlag, Berlin, 1995.

[6] P. D. Loewen and R. T. Rockafellar, *New necessary conditions for the generalized problem of Bolza*, SIAM J. Control. Optim., 34 (1996), pp. 1496–1511.
[7] D. G. Luenberger, *Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, 1984.
[8] Z. Páles and V. Zeidan, *First and second order necessary conditions for control problems with constraints*, Trans. Amer. Math. Soc., 346 (1994), pp. 421–453.
[9] R. T. Rockafellar, *Existence theorems for general control problems of Bolza and Lagrange*, Adv. Math., 15 (1975), pp. 312–333.
[10] R. T. Rockafellar, *Equivalent subgradient versions of Hamiltonian and Euler–Lagrange equations in variational analysis*, SIAM J. Control Optim., 34 (1996), pp. 1300–1314.
[11] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
[12] V. Zeidan and P. Zezza, *The conjugate point condition for smooth control sets*, J. Math. Anal. Appl., 132 (1988), pp. 572–589.
[13] V. Zeidan, *Sufficient conditions with minimal regularity assumptions*, Appl. Math. Optim., 20 (1989), pp. 19–31.
[14] V. Zeidan, *Sufficient conditions for the generalized problem of Bolza*, Trans. Amer. Math. Soc., 275 (1983), pp. 561–585.
[15] V. Zeidan, *Extended Jacobi sufficiency criterion for optimal control*, SIAM J. Control. Optim., 22 (1984), pp. 294–301.
[16] V. Zeidan, *A modified Hamilton-Jacobi approach in the generalized problem of Bolza*, Appl. Math. Optim., 11 (1984), pp. 97–109.
[17] V. Zeidan, *Nonnegativity and positivity of a quadratic functional*, Dynam. Systems Appl., 8 (1999), pp. 571–588.
[18] E. Zeidler, *Nonlinear Functional Analysis and Its Applications* III. *Variational Methods and Optimization*, Springer-Verlag, New York, 1985.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 2, pp. 610–627

A ROBUST CONTROL FRAMEWORK FOR LINEAR, TIME-INVARIANT, SPATIALLY DISTRIBUTED SYSTEMS*

J. REINSCHKE†, M. W. CANTONI‡, AND M. C. SMITH§

Abstract. A robust control framework for linear, time-invariant (LTI), spatially distributed systems is outlined in this paper. We adopt an input-output approach which takes account of the spatially distributed nature of the input and output signals for such systems. The approach is a generalization of H^∞ control in the sense that the 2-norm (in both time and space) is used to quantify the size of signals. It is shown that a frequency-domain representation, in the form of a graph symbol, exists for every LTI, spatially distributed system under very mild assumptions. The graph symbol gives rise to left and right coprime representations if the system is also stabilizable. We investigate fundamental issues of feedback control such as feedback stability and robust stability to plant and/or controller uncertainty quantified in the gap-metric. This includes a generalization of the Sefton–Ober gap formula to the infinite-dimensional operator case. A design example in which an electrostatically destabilized membrane is feedback-stabilized concludes the paper.

Key words. spatially distributed systems, distributed-parameter systems, robust stability, gap-metric, H^∞-control

AMS subject classifications. 93C20, 93C30, 93D09, 93D15, 93D25

PII. S0363012999357057

1. Introduction.

1.1. Motivation. Robust control theory attempts to account for the fact that there is always some mismatch between any model of a physical plant and the actual physical plant itself. A fundamental problem in the theory is to find a suitable measure of the distance between two systems. A solution to this problem (initially only for finite-dimensional, lumped-parameter, linear, time-invariant (LTI) systems) was proposed in terms of the so-called *gap-metric,* which was introduced into the control literature by Zames and El-Sakkary [ZE] in 1980. In the 1980's and early 1990's a system- and control-theoretic framework for *finite-dimensional,* lumped-parameter, LTI systems emerged which, in particular, complemented the use of the gap-metric as a distance measure to quantify uncertainty; see, e.g., [DLMS], [VSF], [V1], [E], [VK], [G1], [GM], [GS1], [OS], [QD], [SO], [V3] as well as the monographs [V2], [F], [ZDG], and [V4]. Substantial parts of this robust control theory were subsequently generalized to certain classes of infinite-dimensional, lumped-parameter, LTI systems (see [C2] and [GS2], for example), to the case of linear, time-varying systems (see [DS], [FGS], [CG1], and [CG2]), and to the case of nonlinear systems (see [DGS], [G2], and [GS5], for example). In this paper we systematically generalize much of this theory to the class of LTI, spatially distributed systems. The results and proofs are in parts based on similar results available in the literature for finite-dimensional,

*Received by the editors May 31, 1999; accepted for publication (in revised form) February 16, 2000; published electronically DATE. Part of this reserach was performed while the first author received financial support from the German Academic Exchange Service (DAAD) through their programme HSP III. The paper was also supported by EPSRC.

http://www.siam.org/journals/sicon/40-2/35705.html

†Siemens AG, I & S MP TC, P.O. Box 3240, D-91050 Erlangen, Germany (Johannes.Reinschke @SIEMENS.com).

‡Deparment of Electrical and Electronic Engineering, The University of Melbourne, Parkville, VIC 3010, Australia (m.cantoni@ee.mu.oz.au).

§Department of Engineering, The University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, UK (mcs@eng.cam.ac.uk).

lumped-parameter, LTI systems; we refer, in particular, to [GS1], [GS3], and [SO]. As in [GS1] and [SO], we will use techniques that are mostly function-theoretic, relying on specific representations for the graph of an LTI, dynamical system. A more general, geometric framework based on operator-theoretic methods, in particular, the use of parallel projection operators, was presented in [FGS]. However, the results of [FGS] do not consider the shift-invariance or causality of systems, and hence the necessity parts of the robust stability results as stated here are not known to hold in the geometric framework.

We take an input-output view of spatially distributed systems in which the system's input and output signals are spatially distributed, i.e., the signals may depend on spatial variables as well as time. This view is motivated by the observation that there exist numerous examples of feedback control problems in which inputs, outputs, disturbances, etc. are naturally considered as spatially distributed signals. A common first step in analyzing spatially distributed systems is to reduce them to "lumped-parameter" form. By contrast, we seek to retain the spatial element of signals in the definition of input-output performance measures. In particular, we will make use of the 2-norm for both the space- and the time-dependence of signals and thereby define the induced norm of a system. As we are concerned with LTI systems, we can think interchangeably of signals and systems in the time and frequency domain. In this way our approach can be considered as a generalization of H^∞ control to spatially distributed systems. We will not be concerned with distributed systems in state-space form (as in [CZ]) but concentrate on input-output properties and issues such as the existence of graph symbols and coprime representations, feedback stability, robustness, etc.

The paper is organized as follows. In subsection 1.2 we discuss a simple example of a spatially distributed system to illustrate the input-output view taken. A number of mathematical results of complex analysis and operator theory needed in the remainder of the paper are gathered in section 1.3. In section 2 we present an abstract approach to spatially distributed systems and show that a frequency-domain representation exists for every spatially distributed LTI system under very mild assumptions. The representation takes the form of a graph symbol from which so-called *(normalized) right and left coprime representations* can be obtained if the system is stabilizable. In section 3 we generalize the Sefton–Ober gap formula [SO] to the infinite-dimensional operator case. In section 4 we present (generalized) robust stability results for feedback loops of shift-invariant systems in which the plant, or the controller, or both, may be uncertain with the uncertainty being measured in the gap-metric. Finally, in section 5 the use of the previously obtained results is illustrated by a controller design example.

1.2. Example. In this subsection we discuss a simple example of a spatially distributed system to illustrate the distributed input-output view taken, and the choice of signal spaces. Consider an elastic string stretched between $x=0$ and $x=1$ and clamped at both ends. Denote the string's deflection from the equilibrium position by $y(x,t)$, and assume the string is set in motion under the action of a distributed load $u(x,t)$ ($u(x,t)\equiv 0$ for $t<0$). The dynamics of the string are governed by the PDE

$$\frac{\partial^2 y(x,t)}{\partial t^2}+\delta\,\frac{\partial y(x,t)}{\partial t}-\tau\,\frac{\partial^2 y(x,t)}{\partial x^2}=u(x,t), \tag{1}$$

$x\in(0,1)$, $t\geq 0$, together with the boundary conditions $y(0,t)=y(1,t)=0$. In (1), $\delta>0$ is a frictional coefficient, and $\tau>0$ represents the tension per unit mass of the string.

We select spaces of input and output signals as follows. At any given time instant we assume that $u(x,t)$ is square-integrable in x over the spatial domain $\mathcal{D}^i=(0,1)$. That is, for fixed t, $u(.,t)\in L^2(\mathcal{D}^i)=:\mathscr{U}$, where $L^2(\mathcal{D}^i)$ is the standard Lebesgue space. Bringing in the time-dependence, we consider $u(.,t)=:u(t)$ to belong to the Lebesgue space of $\mathscr{U}$-valued, square-integrable functions, $L^2_{\mathscr{U}}[0,\infty)$, which is defined as

$$L^2_{\mathscr{U}}[0,\infty):=\left\{u:[0,\infty)\to\mathscr{U}\;\Big|\;\int_0^\infty\langle u(t),\,u(t)\rangle_{\mathscr{U}}\,\mathrm{d}t<\infty\right\},$$

where $\langle .,.\rangle_{\mathscr{U}}$ denotes the scalar product in $\mathscr{U}$. Similarly, a suitable space for the output signals is $L^2_{\mathscr{Y}}[0,\infty)$, where $\mathscr{Y}:=L^2(\mathcal{D}^o)$, and in this case the spatial domain is $\mathcal{D}^o=(0,1)$. Thus we consider the system (1) as defining an operator from $L^2_{\mathscr{U}}[0,\infty)$ to $L^2_{\mathscr{Y}}[0,\infty)$. In this example, $\mathcal{D}^i$ and $\mathcal{D}^o$ are both equal to $(0,1)$. In general, however, we allow for $\mathcal{D}^i$ and $\mathcal{D}^o$ to be different, which permits the description of plants with boundary controls or boundary observations, for example. Also, in the case of spatially multidimensional systems, $\mathcal{D}^i$ and $\mathcal{D}^o$ will be subsets of $\mathbb{R}^{n_u}$ and $\mathbb{R}^{n_y}$ with n_u, $n_y\in\mathbb{Z}_+$.

Taking Laplace transforms of (1) with zero initial conditions gives

$$\left((s^2+\delta\,s)-\tau\frac{\mathrm{d}^2}{\mathrm{d}x^2}\right)\hat{y}(x;s)=\hat{u}(x;s),$$

$x\in(0,1)$, $s\in\mathbb{C}$, plus the boundary conditions $\hat{y}(0;s)=\hat{y}(1;s)=0$. Then $\hat{u}(.;s)=:\hat{u}(s)$ may be regarded as the system's Laplace-transformed spatially distributed input variable and $\hat{y}(.;s)=:\hat{y}(s)$ as its Laplace-transformed spatially distributed output variable. After taking Laplace transforms, the input and output signal spaces become Hardy 2-spaces of $\mathscr{U}$- (respectively, $\mathscr{Y}$-) valued functions; see Definition 1.2 below. In the frequency-domain, the input-output relationship of a spatially distributed LTI plant can often be represented as

$$\hat{y}(x;s)=\int_{\mathcal{D}^i}\mathrm{d}\xi\;\kappa_{\hat{\mathbf{P}}}(x,\xi;s)\,\hat{u}(\xi;s),\quad x\in\mathcal{D}^o,$$

where $\kappa_{\hat{\mathbf{P}}}(x,\xi;s)$ denotes the kernel of an s-dependent, infinite-dimensional, integral operator. For the vibrating string, the integral kernel is given by $\kappa_{\hat{\mathbf{P}}}(x,\xi;s)=\sum_{k=1}^{\infty}2\sin(k\pi x)\big(s^2+\delta\,s+\tau\,(k\,\pi)^2\big)^{-1}\sin(k\pi\xi)$, which is the string's Laplace-transformed Green's function.

1.3. Mathematical preliminaries. In this subsection we introduce the notation and the mathematical background that will be used repeatedly in this paper. Except for Proposition 1.12, the material is taken from the literature (see especially [RR], [FF], and [N]).

Let $\mathbb{C}_+:=\big\{s\in\mathbb{C}\,|\,\mathrm{Re}(s)>0\big\}$. Let $\mathfrak{L}$ represent the Laplace-transform operator. $\mathbf{I}_{\mathscr{V}}$, $\|.\|_{\mathscr{V}}$, and $\langle .,.\rangle_{\mathscr{V}}$ denote the identity operator, the norm, and the scalar product in the Hilbert space $\mathscr{V}$. Throughout, Hilbert spaces will be assumed to be separable. Let $\mathscr{V}_1$ be a subspace of $\mathscr{V}$. The projection operator onto $\mathscr{V}_1$ is denoted by $\mathcal{P}_{\mathscr{V}_1}$. For $\mathscr{U}$, $\mathscr{Y}$ Banach spaces, let the space of all bounded, linear operators from $\mathscr{U}$ to $\mathscr{Y}$, equipped with the induced norm $\|.\|_{\mathrm{ind}}$, be denoted by $\mathscr{B}(\mathscr{U},\mathscr{Y})$. The domain, the null-space, and the range-space of an operator $T:\mathscr{U}\to\mathscr{Y}$ are denoted by $\mathrm{dom}(T):=\{u\in\mathscr{U}\,|\,T\,u\in\mathscr{Y}\}$, $\mathrm{null}(T):=\{u\in\mathscr{U}\,|\,T\,u=0\}$, and $\mathrm{ran}(T):=\{y\in\mathscr{Y}\,|\,y=T\,u,\;u\in\mathrm{dom}(T)\}$.

THEOREM 1.1 (see [H, Problem 52]). *Let $\mathscr{U}$, $\mathscr{Y}$ be Hilbert spaces. If $T\in\mathscr{B}(\mathscr{U},\mathscr{Y})$ is bounded from below (i.e., $\|Tu\|_{\mathscr{Y}}\geq\delta\|u\|_{\mathscr{U}}$ for all $u\in\mathscr{U}$ and some $\delta>0$) and has dense range, then T is boundedly invertible.*

DEFINITION 1.2 (Lebesgue 2-space, Hardy 2-space [RR]). *Let $\mathcal{V}$ be a Hilbert space. The* Lebesgue 2-space $L^2_{\mathcal{V}}(\jmath\mathbb{R})$ *is the Banach space of $\mathcal{V}$-valued functions $\hat{v}$ of a complex variable that are bounded in the norm* $\|\hat{v}\|_{L^2} := \left(\frac{1}{2\pi}\int_{-\infty}^{\infty} \left\|\hat{v}(\jmath\omega)\right\|^2_{\mathcal{V}} \mathrm{d}\omega\right)^{1/2}$. *By the* Hardy 2-space $H^2_{\mathcal{V}}(\mathbb{C}_+)$ *we mean the Banach space of $\mathcal{V}$-valued functions $\hat{v}$ of a complex variable that are analytic on $\mathbb{C}_+$ and bounded in the norm* $\|\hat{v}\|_{H^2} := \left(\sup_{\sigma>0} \frac{1}{2\pi}\int_{-\infty}^{\infty} \left\|\hat{v}(\sigma+\jmath\omega)\right\|^2_{\mathcal{V}} \mathrm{d}\omega\right)^{1/2}$. *If $\hat{v} \in H^2_{\mathcal{V}}(\mathbb{C}_+)$, then $\hat{v}$ can be extended to a boundary function on $\jmath\mathbb{R}$ and $\|\hat{v}\|_{L^2} = \|\hat{v}\|_{H^2}$. The Lebesgue 2-space $L^2_{\mathcal{V}}(\jmath\mathbb{R})$ and the Hardy 2-space $H^2_{\mathcal{V}}(\mathbb{C}_+)$ are Hilbert spaces with inner product* $\langle f, g\rangle_2 := \frac{1}{2\pi}\int_{-\infty}^{\infty} \langle f(\jmath\omega), g(\jmath\omega)\rangle_{\mathcal{V}} \mathrm{d}\omega$, *and $H^2_{\mathcal{V}}(\mathbb{C}_+) \subset L^2_{\mathcal{V}}(\jmath\mathbb{R})$.*

DEFINITION 1.3 (truncation and shift operators). *Let $\mathcal{V}$ be a Hilbert space. The (time-domain) truncation and shift operators, $\mathbf{T}^\tau$ and $\mathbf{S}^\tau$, on the signal space $L^2_{\mathcal{V}}[0,\infty)$ are defined by*

$$\mathbf{T}^\tau v(t) = \begin{cases} v(t) & \text{if } t < \tau, \\ 0 & \text{otherwise,} \end{cases} \qquad \text{and} \qquad \mathbf{S}^\tau v(t) = \begin{cases} v(t-\tau) & \text{if } \tau \le t, \\ 0 & \text{if } 0 \le t < \tau, \end{cases}$$

where $v \in L^2_{\mathcal{V}}[0,\infty)$. The (frequency-domain) shift operator $\hat{\mathbf{S}}^\tau$ on the signal space $H^2_{\mathcal{V}}(\mathbb{C}_+)$ is defined as $\hat{\mathbf{S}}^\tau := \mathcal{L}\mathbf{S}^\tau\mathcal{L}^{-1} = \mathrm{e}^{-\tau s}\,\mathbf{I}_{\mathcal{V}}$. On $L^2_{\mathcal{V}}(\jmath\mathbb{R})$, $\mathrm{e}^{-\jmath\omega\tau}\,\mathbf{I}_{\mathcal{V}}$ corresponds to the bilateral shift by τ.

DEFINITION 1.4 (Lebesgue ∞-space, Hardy ∞-space [RR]). *Let $\mathcal{U}$, $\mathcal{Y}$ be Hilbert spaces. The* Lebesgue ∞-space $L^\infty_{\mathcal{B}(\mathcal{U},\mathcal{Y})}(\jmath\mathbb{R})$ *is the Banach space of $\mathcal{B}(\mathcal{U},\mathcal{Y})$-valued functions T of a complex variable that are bounded in the norm* $\|T\|_{L^\infty} := \operatorname{ess\,sup}_{\omega\in\mathbb{R}} \{\|T(\jmath\omega)\|_{\mathrm{ind}}\}$. *The* Hardy ∞-space $H^\infty_{\mathcal{B}(\mathcal{U},\mathcal{Y})}(\mathbb{C}_+)$ *is the Banach space of $\mathcal{B}(\mathcal{U},\mathcal{Y})$-valued functions T of a complex variable that are analytic on $\mathbb{C}_+$ and bounded in the norm* $\|T\|_{H^\infty} := \sup_{s\in\mathbb{C}_+} \{\|T(s)\|_{\mathrm{ind}}\}$. *If $T \in H^\infty_{\mathcal{B}(\mathcal{U},\mathcal{Y})}(\mathbb{C}_+)$, then $\|T\|_{H^\infty} = \operatorname{ess\,sup}_{\omega\in\mathbb{R}} \|T(\jmath\omega)\|_{\mathrm{ind}}$, and hence we can identify the norms $\|.\|_{L^\infty}$ and $\|.\|_{H^\infty}$ with each other and write simply $\|.\|_\infty$, calling it the* ∞-norm.

DEFINITION 1.5 (multiplication operator [RR]). *Let $T \in H^\infty_{\mathcal{B}(\mathcal{U},\mathcal{Y})}(\mathbb{C}_+)$. The operator $\mathcal{M}_T : H^2_{\mathcal{U}}(\mathbb{C}_+) \to H^2_{\mathcal{Y}}(\mathbb{C}_+)$ defined by $\mathcal{M}_T : f \mapsto Tf$ for all $f \in H^2_{\mathcal{U}}(\mathbb{C}_+)$ is called the* multiplication operator with symbol T.

DEFINITION 1.6 (Laurent operator and para-hermitian conjugate). *Each $G \in L^\infty_{\mathcal{B}(\mathcal{U},\mathcal{Y})}(\jmath\mathbb{R})$ induces an operator $\mathcal{L}_G : L^2_{\mathcal{U}}(\jmath\mathbb{R}) \to L^2_{\mathcal{Y}}(\jmath\mathbb{R})$, called the* Laurent operator with symbol G, *and defined by $\mathcal{L}_G : g \mapsto Gg$ for all $g \in L^2_{\mathcal{U}}(\jmath\mathbb{R})$. The adjoint $\mathcal{L}_G^*$ is equal to the Laurent operator with symbol $G^\sim$, where $G^\sim$ is defined via $\langle f, Gg\rangle_2 = \langle G^\sim f, g\rangle_2$ for all $g \in L^2_{\mathcal{U}}(\jmath\mathbb{R})$, $f \in L^2_{\mathcal{Y}}(\jmath\mathbb{R})$. $G^\sim \in L^\infty_{\mathcal{B}(\mathcal{Y},\mathcal{U})}(\jmath\mathbb{R})$ is called the* para-hermitian conjugate *of G, and it is given by $G^\sim(s) = \left(G(-\overline{s})\right)^*$, where $(.)^*$ denotes the $\mathcal{B}(\mathcal{U},\mathcal{Y})$-adjoint and $\overline{s}$ is the complex-conjugate of s. If $G \in H^\infty_{\mathcal{B}(\mathcal{U},\mathcal{Y})}(\mathbb{C}_+)$, then $\|G\|_{H^\infty} = \sqrt{\|G^\sim G\|_{L^\infty}} = \sqrt{\|G\,G^\sim\|_{L^\infty}}$.*

THEOREM 1.7 ([FF, p. 235]). *Let $\mathcal{U}$, $\mathcal{Y}$ be Hilbert spaces. A bounded linear operator $M : H^2_{\mathcal{U}}(\mathbb{C}_+) \to H^2_{\mathcal{Y}}(\mathbb{C}_+)$ is shift-invariant iff $M = \mathcal{M}_G$ for some $G \in H^\infty_{\mathcal{B}(\mathcal{U},\mathcal{Y})}(\mathbb{C}_+)$. In this case, $\|M\|_{\mathrm{ind}} = \|G\|_\infty$. Similarly, a bounded linear operator $L : L^2_{\mathcal{U}}(\jmath\mathbb{R}) \to L^2_{\mathcal{Y}}(\jmath\mathbb{R})$ commutes with the bilateral shift iff $L = \mathcal{L}_K$ for some $K \in L^\infty_{\mathcal{B}(\mathcal{U},\mathcal{Y})}(\jmath\mathbb{R})$. In this case, $\|L\|_{\mathrm{ind}} = \|K\|_\infty$.*

DEFINITION 1.8 (invertible in H^∞). *An (operator-valued) function $T \in H^\infty_{\mathcal{B}(\mathcal{U},\mathcal{U})}(\mathbb{C}_+)$ is said to be* invertible in H^∞ *if there exists a function $T^{-1} \in H^\infty_{\mathcal{B}(\mathcal{U},\mathcal{U})}(\mathbb{C}_+)$ such that $T^{-1}T = TT^{-1} = \mathbf{I}_{\mathcal{U}}$. The terms* right invertible in H^∞ *and* left invertible in H^∞ *are defined analogously.*

DEFINITION 1.9 (inner and coinner functions [FF, p. 234]). *Let $\mathscr{U}$, $\mathscr{Y}$ be Hilbert spaces. A function $G \in H^\infty_{\mathscr{B}(\mathscr{U},\mathscr{Y})}(\mathbb{C}_+)$ is called* inner *if $G^\sim(\jmath\omega)\,G(\jmath\omega) = \mathbf{I}_\mathscr{U}$ for (almost) all $\omega \in \mathbb{R}$, and* coinner *if $G(\jmath\omega)\,G^\sim(\jmath\omega) = \mathbf{I}_\mathscr{Y}$ for (almost) all $\omega \in \mathbb{R}$.*

DEFINITION 1.10 (outer function [FF, p. 240]). *Let $\mathscr{U}$, $\mathscr{Y}$ be Hilbert spaces. $G_o \in H^\infty_{\mathscr{B}(\mathscr{U},\mathscr{Y})}(\mathbb{C}_+)$ is called an* outer function *(or simply* outer*) if $\operatorname{cl}\big(G_o\,H^2_\mathscr{U}(\mathbb{C}_+)\big) = H^2_\mathscr{Y}(\mathbb{C}_+)$, where* cl(.) *denotes "closure."*

THEOREM 1.11 (Szegö [RR, section 6.14]). *Let $\mathscr{Y}$ be a Hilbert space, and F a (weakly measurable) nonnegative $\mathscr{B}(\mathscr{Y},\mathscr{Y})$-valued function of $\omega \in \mathbb{R}$ that has invertible values for (almost) all $\omega \in \mathbb{R}$. Assume that*

$$\int_{-\infty}^{\infty} \frac{\log^+\|F(\jmath\omega)\|_{\mathrm{ind}}}{1+\omega^2}\,\mathrm{d}\omega < \infty \quad \textit{and} \quad \int_{-\infty}^{\infty} \frac{\log^+\|F^{-1}(\jmath\omega)\|_{\mathrm{ind}}}{1+\omega^2}\,\mathrm{d}\omega < \infty, \tag{2}$$

where $\log^+ t := \max(\log t, 0)$ for $t > 0$ and $\log^+ 0 := -\infty$. Then $F(\jmath\omega) = V^\sim(\jmath\omega)\,V(\jmath\omega)$ for (almost) all $\omega \in \mathbb{R}$, where $V(s)$, $s \in \mathbb{C}$, is a $\mathscr{B}(\mathscr{Y},\mathscr{Y}_1)$-valued outer function with $\mathscr{Y}_1$ being a closed subspace of $\mathscr{Y}$.

The following Proposition 1.12 on the spectral factorization of operator-valued L^∞-functions is a consequence of Szegö's theorem. A proof of the proposition is included as the authors were unable to find the result, as stated below, in the literature.

PROPOSITION 1.12 (spectral factorization). *Let $\mathscr{U}$, $\mathscr{Y}$ be Hilbert spaces.*

(i) *For every function $G \in L^\infty_{\mathscr{B}(\mathscr{U},\mathscr{Y})}(\jmath\mathbb{R})$ that has a left inverse in L^∞, there exists a function $X \in H^\infty_{\mathscr{B}(\mathscr{U},\mathscr{U})}(\mathbb{C}_+)$, which is invertible in H^∞, such that $G^\sim(\jmath\omega)\,G(\jmath\omega) = X^\sim(\jmath\omega)\,X(\jmath\omega)$ for (almost) all $\omega \in \mathbb{R}$.*

(ii) *For every function $\tilde{G} \in L^\infty_{\mathscr{B}(\mathscr{Y},\mathscr{U})}(\jmath\mathbb{R})$ that has a right inverse in L^∞, there exists a function $Y \in H^\infty_{\mathscr{B}(\mathscr{U},\mathscr{U})}(\mathbb{C}_+)$, which is invertible in H^∞, such that $\tilde{G}(\jmath\omega)\,\tilde{G}^\sim(\jmath\omega) = Y(\jmath\omega)\,Y^\sim(\jmath\omega)$ for (almost) all $\omega \in \mathbb{R}$.*

Proof. Given a function $K \in L^\infty_{\mathscr{B}(\mathscr{U},\mathscr{Y})}(\jmath\mathbb{R})$ that has a left inverse $L \in L^\infty_{\mathscr{B}(\mathscr{Y},\mathscr{U})}(\jmath\mathbb{R})$, set $F := K^\sim K$. Since $L\,K = \mathbf{I}_\mathscr{U}$, it follows that $\|K\,\hat{u}\|_2 \geq \|L\|_\infty^{-1}\,\|\hat{u}\|_2$, and hence, since $\|L\|_\infty^{-2}\langle\hat{u},\hat{u}\rangle_2 \leq \langle K\hat{u}, K\hat{u}\rangle_2 = \langle\hat{u}, F\hat{u}\rangle_2 \leq \|\hat{u}\|_2\,\|F\hat{u}\|_2$, we have $\|F\,\hat{u}\|_2 \geq \|L\|_\infty^{-2}\,\|\hat{u}\|_2$ for all $\hat{u} \in L^2_\mathscr{U}(\jmath\mathbb{R})$. Thus $\mathcal{L}_F\colon L^2_\mathscr{U}(\jmath\mathbb{R}) \to L^2_\mathscr{U}(\jmath\mathbb{R})$ is bounded from below. Since K is left invertible, $K^\sim$ is right invertible which means that $\mathcal{L}_{K^\sim}$ has full range. Since K is bounded below, $\operatorname{ran}(\mathcal{L}_K) = \operatorname{cl}\big(\operatorname{ran}(\mathcal{L}_K)\big) = \operatorname{null}(\mathcal{L}_{K^\sim})^\perp$. Hence $\operatorname{ran}(\mathcal{L}_F) = L^2_\mathscr{U}[0,\infty)$. This means that $\mathcal{L}_F$ is boundedly invertible by Theorem 1.1. Since $\mathcal{L}_F$ commutes with the bilateral shift, so does $\mathcal{L}_F^{-1}$, implying the existence of a function $F^{-1} \in L^\infty_{\mathscr{B}(\mathscr{U},\mathscr{U})}(\jmath\mathbb{R})$ such that $\mathcal{L}_F^{-1} = \mathcal{L}_{F^{-1}}$ by Theorem 1.7. Note that both F and F^{-1} satisfy the conditions (2) in Theorem 1.11. Consequently, there exist outer functions $V_1 \in H^\infty_{\mathscr{B}(\mathscr{U},\mathscr{U}_1)}(\mathbb{C}_+)$ and $V_2 \in H^\infty_{\mathscr{B}(\mathscr{U},\mathscr{U}_2)}(\mathbb{C}_+)$, with $\mathscr{U}_1$ and $\mathscr{U}_2$ being closed subspaces of $\mathscr{U}$, such that $F(\jmath\omega) = V_1^\sim(\jmath\omega)\,V_1(\jmath\omega)$ and $F^{-1}(\jmath\omega) = V_2^\sim(\jmath\omega)\,V_2(\jmath\omega)$ for (almost) all $\omega \in \mathbb{R}$. The multiplication operators $\mathcal{M}_{V_1}\colon H^2_\mathscr{U}(\mathbb{C}_+) \to H^2_{\mathscr{U}_1}(\mathbb{C}_+)$ and $\mathcal{M}_{V_2}\colon H^2_\mathscr{U}(\mathbb{C}_+) \to H^2_{\mathscr{U}_2}(\mathbb{C}_+)$ have dense range (since V_1 and V_2 are outer) and are bounded from below (since $\|F\hat{u}\|_2 = \|V_1^\sim\,V_1\,\hat{u}\|_2 \geq \|L\|_\infty^{-2}\,\|\hat{u}\|_2$ implies $\|V_1\,\hat{u}\|_2 \geq \|V_1\|_\infty^{-1}\,\|L\|_\infty^{-2}\,\|\hat{u}\|_2$, and similarly for V_2). Combining Theorems 1.1 and 1.7, it follows that both V_1 and V_2 are invertible in H^∞. Furthermore, $\dim\mathscr{U}_1 = \dim\mathscr{U}_2 = \dim\mathscr{U}$, i.e., both $\mathscr{U}_1$ and $\mathscr{U}_2$ are isometrically isomorphic to $\mathscr{U}$ [K2, p. 173], and therefore there exist (constant) isometric isomorphisms $T_1\colon \mathscr{U}_1 \to \mathscr{U}$ and $T_2\colon \mathscr{U}_2 \to \mathscr{U}$.

To prove part (i), set $K := G$ and $X := T_1\,V_1$. Now note that $X, X^{-1} \in H^\infty_{\mathscr{B}(\mathscr{U},\mathscr{U})}(\mathbb{C}_+)$ and that $G^\sim(\jmath\omega)\,G(\jmath\omega) = F(\jmath\omega) = V_1^\sim(\jmath\omega)\,V_1(\jmath\omega) = X^\sim(\jmath\omega)\,X(\jmath\omega)$ for (almost) all $\omega \in \mathbb{R}$.

To prove part (ii), set $K := \tilde{G}^{\sim}$ and $Y := \begin{pmatrix} T_2 \\ V_2 \end{pmatrix}^{-1}$. Now note that $Y, Y^{-1} \in H^{\infty}_{\mathscr{B}(\mathscr{U},\mathscr{U})}(\mathbb{C}_+)$ and that $\big(\tilde{G}(\jmath\omega)\,\tilde{G}^{\sim}(\jmath\omega)\big)^{-1} = F^{-1}(\jmath\omega) = V_2^{\sim}(\jmath\omega)\,V_2(\jmath\omega) = \big(Y(\jmath\omega)\,Y^{\sim}(\jmath\omega)\big)^{-1}$, i.e., $\tilde{G}(\jmath\omega)\,\tilde{G}^{\sim}(\jmath\omega) = Y(\jmath\omega)\,Y^{\sim}(\jmath\omega)$ for (almost) all $\omega \in \mathbb{R}$. □

2. System representations. Mathematical models for spatially distributed systems may arise in the form of PDEs, integral operators, or as a result of direct modelling techniques such as the finite element method. In this section we take an abstract approach to show that frequency domain representations exist for LTI, spatially distributed systems under very mild assumptions. The representation takes the form of a graph symbol. We will furthermore show that the graph symbol gives rise to so-called *(normalized) right and left coprime representations* if the system is (feedback-) stabilizable.

In the time-domain, a spatially distributed system is considered to be an operator

$$\mathbf{P} : \mathrm{dom}(\mathbf{P}) \subseteq L^2_{\mathscr{U}}[0,\infty) \to L^2_{\mathscr{Y}}[0,\infty),$$

where $\mathscr{U}$, $\mathscr{Y}$ are (infinite-dimensional) Hilbert spaces. The *graph* of $\mathbf{P}$ is defined by $\mathcal{G}_{\mathbf{P}} := \left[\begin{smallmatrix} \mathbf{I}_{\mathscr{U}} \\ \mathbf{P} \end{smallmatrix}\right] \mathrm{dom}(\mathbf{P}) \subseteq L^2_{\left[\begin{smallmatrix}\mathscr{U}\\ \mathscr{Y}\end{smallmatrix}\right]}[0,\infty)$. A system $\mathbf{P}$ is said to be *linear* if $\mathcal{G}_{\mathbf{P}}$ is a linear subspace of $L^2_{\left[\begin{smallmatrix}\mathscr{U}\\ \mathscr{Y}\end{smallmatrix}\right]}[0,\infty)$. A linear system $\mathbf{P}$ is said to be *stable* if $\mathrm{dom}(\mathbf{P}) = L^2_{\mathscr{U}}[0,\infty)$ and $\|\mathbf{P}\|_{\mathrm{ind}} < \infty$. A system $\mathbf{P}$ is said to be *shift-invariant* (or *time-invariant*) if, for all $\tau \in (0,\infty)$ and all $u \in \mathrm{dom}(\mathbf{P})$, $\mathbf{S}^{\tau}\,\mathbf{P}\,u = \mathbf{P}\,\mathbf{S}^{\tau}\,u$. This is equivalent to the graph $\mathcal{G}_{\mathbf{P}}$ being a shift-invariant subspace of $L^2_{\left[\begin{smallmatrix}\mathscr{U}\\ \mathscr{Y}\end{smallmatrix}\right]}[0,\infty)$, i.e., $\mathbf{S}^{\tau}\,\mathcal{G}_{\mathbf{P}} \subseteq \mathcal{G}_{\mathbf{P}}$ for all $\tau \in (0,\infty)$. A linear system $\mathbf{P}$ is said to be *causal* if, for all $\tau \in (0,\infty)$ and all $u_1, u_2 \in \mathrm{dom}(\mathbf{P})$, $\mathbf{T}^{\tau}\,u_1 = \mathbf{T}^{\tau}\,u_2$ implies $\mathbf{T}^{\tau}\,\mathbf{P}\,u_1 = \mathbf{T}^{\tau}\,\mathbf{P}\,u_2$. This is equivalent to the graph of $\mathbf{P}$ satisfying the following: $\left(\begin{smallmatrix} 0 \\ \hat{v} \end{smallmatrix}\right) \in \mathbf{T}^{\tau}\,\mathcal{G}_{\mathbf{P}}$ implies $\hat{v} = 0$ for all $\tau \in (0,\infty)$. A stable, linear system $\mathbf{P}$ is causal iff $\mathbf{T}^{\tau}\,\mathbf{P}\,u = \mathbf{T}^{\tau}\,\mathbf{P}\,\mathbf{T}^{\tau}\,u$ for all $u \in L^2_{\mathscr{U}}[0,\infty)$. Furthermore, it is a standard fact that every linear, stable, and shift-invariant system, which is defined on the singly infinite time axis $[0,\infty)$, is causal. A causal, linear system $\mathbf{P}$ is called *causally extendible* if $\mathbf{T}^{\tau}\,\mathrm{dom}(\mathbf{P}) = L^2_{\mathscr{U}}[0,\tau) := \mathbf{T}^{\tau} L^2_{\mathscr{U}}[0,\infty)$ for all $\tau \in (0,\infty)$. Causal extendibility means that, given any $\tau > 0$, we can choose the input to $\mathbf{P}$ arbitrarily over the interval $[0,\tau)$ and yet be able to continue the input to an element of $\mathrm{dom}(\mathbf{P})$.

Consider the LTI system $\mathbf{P} : \mathrm{dom}(\mathbf{P}) \subseteq L^2_{\mathscr{U}}[0,\infty) \to L^2_{\mathscr{Y}}[0,\infty)$ with input-output relationship given by $y = \mathbf{P}\,u$. Denote the Laplace-transforms of u and y by $\hat{u} := \mathcal{L}u$ and $\hat{y} := \mathcal{L}y$, respectively. In the "frequency domain," the system is represented by the operator $\hat{\mathbf{P}} : \mathrm{dom}(\hat{\mathbf{P}}) \subseteq H^2_{\mathscr{U}}(\mathbb{C}_+) \to H^2_{\mathscr{Y}}(\mathbb{C}_+)$, defined via $\hat{\mathbf{P}} := \mathcal{L}\mathbf{P}\mathcal{L}^{-1}$. The Laplace-transformed system graph is denoted by $\mathcal{G}_{\hat{\mathbf{P}}} := \mathcal{L}\mathcal{G}_{\mathbf{P}}$. It is well known that the Laplace-transform (of signals) is an isometric isomorphism between Hilbert spaces. Thus we can think interchangeably of operators defined in the "time domain" or the "frequency domain." The equivalence between "time domain" and "frequency domain" can be depicted in diagrammatic form:

$$\begin{array}{ccccc} L^2_{\mathscr{U}}[0,\infty) & \supseteq & \mathrm{dom}(\mathbf{P}) & \xrightarrow{\ \mathbf{P}\ } & L^2_{\mathscr{Y}}[0,\infty) \\ \mathcal{L} \Updownarrow \mathcal{L}^{-1} & & \mathcal{L} \Updownarrow \mathcal{L}^{-1} & & \mathcal{L} \Updownarrow \mathcal{L}^{-1} \\ H^2_{\mathscr{U}}[\mathbb{C}_+) & \supseteq & \mathrm{dom}(\hat{\mathbf{P}}) & \xrightarrow{\ \hat{\mathbf{P}}\ } & H^2_{\mathscr{Y}}[\mathbb{C}_+). \end{array}$$

The following result, which is a generalization of a fundamental result for lumped-parameter systems, states mild conditions under which the graph of a spatially distributed LTI system can be represented as the range of an operator-valued H^{∞}-function.

THEOREM 2.1. *Let $\mathscr{U}$, $\mathscr{Y}$ be (separable) Hilbert spaces. Consider a closed, linear, shift-invariant, causally extendible system* $\mathbf{P}\colon \operatorname{dom}(\mathbf{P}) \subseteq L^2_{\mathscr{U}}[0,\infty) \to L^2_{\mathscr{Y}}[0,\infty)$. *Then there exist (operator-valued) functions* $M \in H^\infty_{\mathscr{B}(\mathscr{U},\,\mathscr{U})}(\mathbb{C}_+)$ *and* $N \in H^\infty_{\mathscr{B}(\mathscr{U},\,\mathscr{Y})}(\mathbb{C}_+)$ *such that* $G := \left[\begin{smallmatrix} M \\ N \end{smallmatrix}\right]$ *is inner and satisfies* $\mathcal{G}_{\hat{\mathbf{P}}} = G\, H^2_{\mathscr{U}}(\mathbb{C}_+)$.

Proof (the proof is based on the proofs of Propositions 1 and 11 in [GS3]). Call $\mathscr{V} := \left[\begin{smallmatrix} \mathscr{U} \\ \mathscr{Y} \end{smallmatrix}\right]$. By assumption $\mathcal{G}_{\hat{\mathbf{P}}}$ is a closed, shift-invariant subspace of $H^2_{\mathscr{V}}(\mathbb{C}_+)$, and hence by the Beurling–Lax theorem [FF, p. 239] there exists an at most countably infinite-dimensional Hilbert space $\mathscr{Z}$ and an inner function $G' \in H^\infty_{\mathscr{B}(\mathscr{Z},\,\mathscr{V})}(\mathbb{C}_+)$ such that $\mathcal{G}_{\hat{\mathbf{P}}} = G'\, H^2_{\mathscr{Z}}(\mathbb{C}_+)$. We will show next that the assumption of causal extendibility of $\mathbf{P}$ implies that $\mathscr{Z}$ cannot be finite-dimensional if $\mathscr{U}$ is infinite-dimensional. Set $G' =: \left[\begin{smallmatrix} M' \\ N' \end{smallmatrix}\right]$, where $M' \in H^\infty_{\mathscr{B}(\mathscr{Z},\,\mathscr{U})}(\mathbb{C}_+)$ and $N' \in H^\infty_{\mathscr{B}(\mathscr{Z},\,\mathscr{Y})}(\mathbb{C}_+)$. $\mathbf{P}$ is causally extendible, meaning that for all $\tau > 0$ and all $\hat{u} \in H^2_{\mathscr{U}}(\mathbb{C}_+) \ominus \mathrm{e}^{-\tau s} H^2_{\mathscr{U}}(\mathbb{C}_+)$ there exists a $\hat{z} \in H^2_{\mathscr{Z}}(\mathbb{C}_+)$ such that $\hat{u} = \mathcal{P}_{H^2_{\mathscr{U}}(\mathbb{C}_+) \ominus \mathrm{e}^{-\tau s} H^2_{\mathscr{U}}(\mathbb{C}_+)} M'\, \hat{z}$. Equivalently, for every $\hat{u} \in H^2_{\mathscr{U}}(\mathbb{C}_+)$ and $\tau > 0$, $\hat{u} = M'\, \hat{w} - \mathrm{e}^{-\tau s}\, \hat{v} =: M'' \left(\begin{smallmatrix} \hat{w} \\ \hat{v} \end{smallmatrix}\right)$ has a solution $\left(\begin{smallmatrix} \hat{w} \\ \hat{v} \end{smallmatrix}\right) \in H^2_{\left[\begin{smallmatrix} \mathscr{Z} \\ \mathscr{U} \end{smallmatrix}\right]}(\mathbb{C}_+)$. This condition is satisfied iff $M'' H^2_{\left[\begin{smallmatrix} \mathscr{Z} \\ \mathscr{U} \end{smallmatrix}\right]}(\mathbb{C}_+) = H^2_{\mathscr{U}}(\mathbb{C}_+)$, i.e., the multiplication operator $\mathcal{M}_{M''} : H^2_{\left[\begin{smallmatrix} \mathscr{Z} \\ \mathscr{U} \end{smallmatrix}\right]}(\mathbb{C}_+) \to H^2_{\mathscr{U}}(\mathbb{C}_+)$ is onto. This implies that $\mathcal{M}_{M''}$ has a right inverse, which by [FF, Corollary VI.6.2, p. 218] can be expressed as a multiplication operator with symbol in H^∞. Hence $\big(M''(s)\big)^*$ is left invertible for fixed $s \in \mathbb{C}_+$. Moreover, $\big(M''(s)\big)^*$ is uniformly (in $\mathbb{C}_+$) bounded from below. Thus $\inf_{s \in \mathbb{C}_+} \left\langle \big(M''(s)\big)^* u,\, \big(M''(s)\big)^* u \right\rangle_{\left[\begin{smallmatrix} \mathscr{Z} \\ \mathscr{U} \end{smallmatrix}\right]} > 0$ for all $0 \neq u \in \mathscr{U}$, and hence

$$\inf_{s \in \mathbb{C}_+} \left(\left\langle \big(M'(s)\big)^* u,\, \big(M'(s)\big)^* u \right\rangle_{\mathscr{Z}} + \mathrm{e}^{-2\tau \mathrm{Re}(s)} \langle u,\, u \rangle_{\mathscr{U}} \right) > 0 \qquad \text{for all} \quad 0 \neq u \in \mathscr{U}. \tag{3}$$

Thus, for all $s \in \mathbb{C}_+$, the bounded linear operator $\big(M'(s)\big)^* : \mathscr{U} \to \mathscr{Z}$ has zero null-space, and hence $\dim \mathscr{U} = \dim \operatorname{ran} \big(M'(s)\big)^* \leq \dim \mathscr{Z}$. Therefore, if $\mathscr{U}$ is countably infinite-dimensional, so must be $\mathscr{Z}$. Since the Hilbert spaces $\mathscr{Z}$ and $\mathscr{U}$ have the same dimension, they are isometrically isomorphic ([K2, p. 173]), i.e., there exists a (constant) isometric isomorphism $U : \mathscr{U} \to \mathscr{Z}$. Finally, $G := G'\, U$ satisfies $\mathcal{G}_{\hat{\mathbf{P}}} = G\, H^2_{\mathscr{U}}(\mathbb{C}_+)$, where G so defined is inner since $G \in H^\infty_{\mathscr{B}(\mathscr{U},\,\left[\begin{smallmatrix} \mathscr{U} \\ \mathscr{Y} \end{smallmatrix}\right])}(\mathbb{C}_+)$ and $\big(G(\jmath\omega)\big)^{\sim} G(\jmath\omega) = U^* \big(G'(\jmath\omega)\big)^{\sim} G'(\jmath\omega)\, U = U^* \mathbf{I}_{\mathscr{Z}}\, U = \mathbf{I}_{\mathscr{U}}$ for (almost) all $\omega \in \mathbb{R}$. ∎

DEFINITION 2.2 (graph symbol). *Given a closed, linear, and shift-invariant system* $\mathbf{P}\colon\ L^2_{\mathscr{U}}[0,\infty) \to L^2_{\mathscr{Y}}[0,\infty)$, *a function* $G \in H^\infty_{\mathscr{B}(\mathscr{U},\,\left[\begin{smallmatrix} \mathscr{U} \\ \mathscr{Y} \end{smallmatrix}\right])}(\mathbb{C}_+)$ *satisfying* $\mathcal{G}_{\hat{\mathbf{P}}} = G\, H^2_{\mathscr{U}}(\mathbb{C}_+)$ *is called a* graph symbol *of* $\mathbf{P}$.

In the physical sciences and engineering it is often the case that the input-output behavior of a system can be represented in terms of an integral operator. Though we are not able to provide an existence theorem for such representations, based on fundamental system properties as in Theorem 2.1, we now briefly discuss the possible relationship between such representations and the graph symbol arising from Theorem 2.1. Suppose that $\mathbf{P}$ is a causal, linear, shift-invariant system with integral operator representation of the form

$$y(x,t) = \int_0^t \int_{\mathcal{D}^i} \kappa_{\mathbf{P}}(x,\xi,t-\tau)\, u(\xi,\tau)\, \mathrm{d}\xi\, \mathrm{d}\tau, \qquad x \in \mathcal{D}^o, \tag{4}$$

where $\mathcal{D}^i$ and $\mathcal{D}^o$ denote the input and output spatial domains. Taking Laplace-

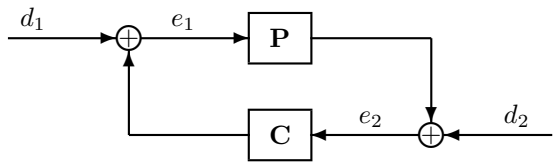


FIG. 1. *Standard feedback configuration* $[\mathbf{P}, \mathbf{C}]$ *of spatially distributed systems in the "time domain."*

transforms of (4) gives

$$\hat{y}(x;s) = \int_{\mathcal{D}^i} \kappa_{\hat{\mathbf{P}}}(x,\xi;s)\,\hat{u}(\xi;s)\,\mathrm{d}\xi, \qquad x \in \mathcal{D}^o,$$

where $\kappa_{\hat{\mathbf{P}}}(x,\xi;s)$ is the Laplace-transform of the system's Green's function $\kappa_{\mathbf{P}}(x,\xi;t)$. This leads to a "frequency domain" interpretation of $\mathbf{P}$ as a parameter-dependent (i.e., s-dependent) integral operator $P(s)$ with kernel $\kappa_{\hat{\mathbf{P}}}(x,\xi;s)$. Now suppose that $\mathbf{P}$ is closed and causally extendible, so that by Theorem 2.1 there exists an $M \in H^\infty_{\mathscr{B}(\mathscr{U},\mathscr{U})}(\mathbb{C}_+)$ and an $N \in H^\infty_{\mathscr{B}(\mathscr{U},\mathscr{Y})}(\mathbb{C}_+)$ such that $\mathcal{G}_{\hat{\mathbf{P}}} = G\,H^2_{\mathscr{U}}(\mathbb{C}_+)$, where $G =: \left[\begin{smallmatrix} M \\ N \end{smallmatrix}\right]$, $\mathscr{U} := L^2(\mathcal{D}^i)$, and $\mathscr{Y} := L^2(\mathcal{D}^o)$. If $\mathbf{P}$ is also bounded on the whole of $L^2_{\mathscr{U}}[0,\infty)$, then it can be shown that the corresponding M is invertible in $H^\infty_{\mathscr{B}(\mathscr{U},\mathscr{U})}(\mathbb{C}_+)$ and that $P(s) = N(s)M^{-1}(s)$. In general, however, although it can be shown that $\mathcal{M}_M$ is always injective, whether $(\mathcal{M}_M)^{-1}$ corresponds to multiplication by some operator-valued function (frequency-domain symbol) appears to be an open question. In such cases, the relationship between the integral operator representation and graph symbol needs further clarification.

Consider now the spatially distributed feedback system in Figure 1, which we denote by $[\mathbf{P}, \mathbf{C}]$, where $\mathbf{P} : \mathrm{dom}(\mathbf{P}) \subseteq L^2_{\mathscr{U}}[0,\infty) \to L^2_{\mathscr{Y}}[0,\infty)$ and $\mathbf{C} : \mathrm{dom}(\mathbf{C}) \subseteq L^2_{\mathscr{Y}}[0,\infty) \to L^2_{\mathscr{U}}[0,\infty)$ are linear operators.

DEFINITION 2.3 (feedback stability). *Consider the systems* $\mathbf{P} : \mathrm{dom}(\mathbf{P}) \subseteq L^2_{\mathscr{U}}[0,\infty) \to L^2_{\mathscr{Y}}[0,\infty)$ *and* $\mathbf{C} : \mathrm{dom}(\mathbf{C}) \subseteq L^2_{\mathscr{Y}}[0,\infty) \to L^2_{\mathscr{U}}[0,\infty)$*, where* $\mathscr{U}$ *and* $\mathscr{Y}$ *are Hilbert spaces. The standard feedback configuration* $[\mathbf{P}, \mathbf{C}]$ *in Figure* 1 *is said to be* stable *if*

$$\mathbf{F}(\mathbf{P},\mathbf{C}) := \begin{pmatrix} \mathbf{I}_{\mathscr{U}} & -\mathbf{C} \\ -\mathbf{P} & \mathbf{I}_{\mathscr{Y}} \end{pmatrix} : \ \mathrm{dom}(\mathbf{P}) \times \mathrm{dom}(\mathbf{C}) \subseteq L^2_{\left[\begin{smallmatrix}\mathscr{U}\\ \mathscr{Y}\end{smallmatrix}\right]}[0,\infty) \to L^2_{\left[\begin{smallmatrix}\mathscr{U}\\ \mathscr{Y}\end{smallmatrix}\right]}[0,\infty) : \ \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \mapsto \begin{pmatrix} d_1 \\ d_2 \end{pmatrix}$$

has a bounded inverse $\big(\mathbf{F}(\mathbf{P},\mathbf{C})\big)^{-1} =: \mathbf{H}(\mathbf{P},\mathbf{C})$ *on* $L^2_{\left[\begin{smallmatrix}\mathscr{U}\\ \mathscr{Y}\end{smallmatrix}\right]}[0,\infty)$.

The *inverse graph* of $\mathbf{C}$ is defined as $\mathcal{G}^{\mathrm{i}}_{\mathbf{C}} := \left[\begin{smallmatrix}\mathbf{C}\\ \mathbf{I}_{\mathscr{Y}}\end{smallmatrix}\right] \mathrm{dom}(\mathbf{C})$, and its frequency-domain equivalent is denoted by $\mathcal{G}^{\mathrm{i}}_{\hat{\mathbf{C}}} := \mathcal{L}\mathcal{G}^{\mathrm{i}}_{\mathbf{C}}$. The following result characterizes feedback stability in terms of the graphs of $\mathbf{P}$ and $\mathbf{C}$; see [FGS1, OS, GS3, FGS], for example.

PROPOSITION 2.4. *Consider the systems* $\mathbf{P} : \mathrm{dom}(\mathbf{P}) \subseteq L^2_{\mathscr{U}}[0,\infty) \to L^2_{\mathscr{Y}}[0,\infty)$ *and* $\mathbf{C} : \mathrm{dom}(\mathbf{C}) \subseteq L^2_{\mathscr{Y}}[0,\infty) \to L^2_{\mathscr{U}}[0,\infty)$*, where* $\mathscr{U}$ *and* $\mathscr{Y}$ *are Hilbert spaces. The*

feedback configuration $[\mathbf{P}, \mathbf{C}]$ *is stable iff* $\mathcal{G}_{\mathbf{P}}$, $\mathcal{G}^{\mathrm{i}}_{\mathbf{C}}$ *are closed subspaces of* $L^2_{\left[\begin{smallmatrix}\mathscr{U}\\ \mathscr{Y}\end{smallmatrix}\right]}[0,\infty)$,

$$\mathcal{G}_{\mathbf{P}} \cap \mathcal{G}^{\mathrm{i}}_{\mathbf{C}} = \{0\}, \tag{5}$$

$$\mathcal{G}_{\mathbf{P}} + \mathcal{G}^{\mathrm{i}}_{\mathbf{C}} = L^2_{\left[\begin{smallmatrix}\mathscr{U}\\ \mathscr{Y}\end{smallmatrix}\right]}[0,\infty). \tag{6}$$

On the basis of Theorem 2.1 and Definition 2.3, we now define the class of spatially distributed systems considered in this paper.

DEFINITION 2.5 (class of shift-invariant, spatially distributed systems). *Consider the Hilbert spaces* $\mathscr{U} := L^2(\mathcal{D}^i)$ *and* $\mathscr{Y} := L^2(\mathcal{D}^o)$, *where* $\mathcal{D}^i$ *and* $\mathcal{D}^o$ *are spatial domains. We define* $\mathcal{S}\big(\mathscr{U}, \mathscr{Y}; [0,\infty)\big)$ *to be the set of all closed, linear, shift-invariant, causally extendible operators* $\mathbf{P} : \mathrm{dom}(\mathbf{P}) \subseteq L^2_{\mathscr{U}}[0,\infty) \to L^2_{\mathscr{Y}}[0,\infty)$. *By* $\mathcal{SS}\big(\mathscr{U}, \mathscr{Y}; [0,\infty)\big)$ *we denote the subclass of* stabilizable *systems in* $\mathcal{S}\big(\mathscr{U}, \mathscr{Y}; [0,\infty)\big)$.

The requirement that the elements of $\mathcal{S}\big(\mathscr{U}, \mathscr{Y}; [0,\infty)\big)$ be closed operators (i.e., operators with closed graphs) is not a major restriction since, by Proposition 2.4, closedness of the system graph is necessary for the system to be stabilizable. We next define *coprime representations,* which are shown to exist for every system in $\mathcal{SS}\big(\mathscr{U}, \mathscr{Y}; [0,\infty)\big)$. Coprime representations have proved to be useful notions in feedback control, particularly in dealing with (open-loop) unstable systems, and as shown below they lead to a neat parameterization of stabilizing controllers. The following definition is adapted from [DS].

DEFINITION 2.6 (coprime representations). *Given a system* $\mathbf{P} \in \mathcal{S}\big(\mathscr{U}, \mathscr{Y}; [0,\infty)\big)$, $\left[\begin{smallmatrix}M\\ N\end{smallmatrix}\right]$ *is said to be a* right coprime representation *of* $\mathbf{P}$ *if* $M \in H^\infty_{\mathscr{B}(\mathscr{U},\mathscr{U})}(\mathbb{C}_+)$ *and* $N \in H^\infty_{\mathscr{B}(\mathscr{U},\mathscr{Y})}(\mathbb{C}_+)$ *satisfy* $\mathcal{G}_{\hat{\mathbf{P}}} = \left[\begin{smallmatrix}M\\ N\end{smallmatrix}\right] H^2_{\mathscr{U}}(\mathbb{C}_+)$, *and* $\left[\begin{smallmatrix}M\\ N\end{smallmatrix}\right]$ *is left invertible in* H^∞. *If, in addition,* $\left[\begin{smallmatrix}M\\ N\end{smallmatrix}\right]$ *is inner, then the right coprime representation is called* normalized. *Given* $\mathbf{P} \in \mathcal{S}\big(\mathscr{U}, \mathscr{Y}; [0,\infty)\big)$, $\begin{bmatrix}-\tilde{N} & \tilde{M}\end{bmatrix}$ *is said to be a* left coprime representation *of* $\mathbf{P}$ *if* $\tilde{M} \in H^\infty_{\mathscr{B}(\mathscr{Y},\mathscr{Y})}(\mathbb{C}_+)$ *and* $\tilde{N} \in H^\infty_{\mathscr{B}(\mathscr{U},\mathscr{Y})}(\mathbb{C}_+)$ *satisfy* $\mathcal{G}_{\hat{\mathbf{P}}} = \big\{\hat{v} \in H^2_{\left[\begin{smallmatrix}\mathscr{U}\\ \mathscr{Y}\end{smallmatrix}\right]}(\mathbb{C}_+) \;\big|\; \begin{bmatrix}-\tilde{N} & \tilde{M}\end{bmatrix} \hat{v} = 0\big\}$, *and* $\begin{bmatrix}-\tilde{N} & \tilde{M}\end{bmatrix}$ *is right invertible in* H^∞. *If, in addition,* $\begin{bmatrix}-\tilde{N} & \tilde{M}\end{bmatrix}$ *is coinner, then the left coprime representation is called* normalized.

THEOREM 2.7. *Every* $\mathbf{P} \in \mathcal{SS}\big(\mathscr{U}, \mathscr{Y}; [0,\infty)\big)$ *has normalized right and normalized left coprime representations* $\left[\begin{smallmatrix}M\\ N\end{smallmatrix}\right] \in H^\infty_{\mathscr{B}(\mathscr{U},\left[\begin{smallmatrix}\mathscr{U}\\ \mathscr{Y}\end{smallmatrix}\right])}(\mathbb{C}_+)$ *and* $\begin{bmatrix}-\tilde{N} & \tilde{M}\end{bmatrix} \in H^\infty_{\mathscr{B}(\left[\begin{smallmatrix}\mathscr{U}\\ \mathscr{Y}\end{smallmatrix}\right],\mathscr{Y})}(\mathbb{C}_+)$, *respectively.*

Proof. Suppose that $\mathbf{C} \in \mathcal{SS}\big(\mathscr{Y}, \mathscr{U}; [0,\infty)\big)$ stabilizes $\mathbf{P}$. By Theorem 2.1 there exist (operator-valued) inner functions $G \in H^\infty_{\mathscr{B}(\mathscr{U},\left[\begin{smallmatrix}\mathscr{U}\\ \mathscr{Y}\end{smallmatrix}\right])}(\mathbb{C}_+)$ and $K \in H^\infty_{\mathscr{B}(\mathscr{Y},\left[\begin{smallmatrix}\mathscr{U}\\ \mathscr{Y}\end{smallmatrix}\right])}(\mathbb{C}_+)$ such that $\mathcal{G}_{\hat{\mathbf{P}}} = GH^2_{\mathscr{U}}(\mathbb{C}_+)$ and $\mathcal{G}^{\mathrm{i}}_{\hat{\mathbf{C}}} = KH^2_{\mathscr{Y}}(\mathbb{C}_+)$. Let $T := (G, K)$. Then since $[\mathbf{P}, \mathbf{C}]$ is stable, it follows by Proposition 2.4 that

$$T\, H^2_{\left[\begin{smallmatrix}\mathscr{U}\\ \mathscr{Y}\end{smallmatrix}\right]}(\mathbb{C}_+) = GH^2_{\mathscr{U}}(\mathbb{C}_+) + KH^2_{\mathscr{Y}}(\mathbb{C}_+) = \mathcal{G}_{\hat{\mathbf{P}}} + \mathcal{G}^{\mathrm{i}}_{\hat{\mathbf{C}}} = H^2_{\left[\begin{smallmatrix}\mathscr{U}\\ \mathscr{Y}\end{smallmatrix}\right]}(\mathbb{C}_+).$$

Moreover, since $\mathrm{null}\big(\mathcal{M}_G\big) = \{0\}$ and $\mathrm{null}\big(\mathcal{M}_K\big) = \{0\}$, it follows from (5) that $\mathcal{M}_T : H^2_{\left[\begin{smallmatrix}\mathscr{U}\\ \mathscr{Y}\end{smallmatrix}\right]}(\mathbb{C}_+) \to H^2_{\left[\begin{smallmatrix}\mathscr{U}\\ \mathscr{Y}\end{smallmatrix}\right]}(\mathbb{C}_+)$ is also injective, and hence boundedly invertible. As the inverse is also shift-invariant, it follows by Theorem 1.7 that T is invertible in H^∞. Partition $T =: \left(\begin{smallmatrix}M & Y\\ N & X\end{smallmatrix}\right) \in H^\infty_{\mathscr{B}(\left[\begin{smallmatrix}\mathscr{U}\\ \mathscr{Y}\end{smallmatrix}\right],\left[\begin{smallmatrix}\mathscr{U}\\ \mathscr{Y}\end{smallmatrix}\right])}(\mathbb{C}_+)$ and $T^{-1} =: \left(\begin{smallmatrix}\tilde{X} & -\tilde{Y}\\ -\tilde{N} & \tilde{M}\end{smallmatrix}\right) \in H^\infty_{\mathscr{B}(\left[\begin{smallmatrix}\mathscr{U}\\ \mathscr{Y}\end{smallmatrix}\right],\left[\begin{smallmatrix}\mathscr{U}\\ \mathscr{Y}\end{smallmatrix}\right])}(\mathbb{C}_+)$. Clearly, $\begin{bmatrix}\tilde{X} & -\tilde{Y}\end{bmatrix}$ is a left inverse of $G = \left[\begin{smallmatrix}M\\ N\end{smallmatrix}\right]$, and correspondingly G is a right coprime representation of $\mathbf{P}$. Furthermore, it is normalized since G is inner.

We now show that $\tilde{G} := \begin{bmatrix}-\tilde{N} & \tilde{M}\end{bmatrix}$ is a left coprime representation of $\mathbf{P}$. First note that $\left[\begin{smallmatrix}Y\\ X\end{smallmatrix}\right]$ is a right inverse of $\tilde{G}$. Let $\hat{v}_1, \hat{v}_2 \in H^2_{\left[\begin{smallmatrix}\mathscr{U}\\ \mathscr{Y}\end{smallmatrix}\right]}(\mathbb{C}_+)$ satisfy $0 \neq \hat{v}_1 \in \mathcal{G}_{\hat{\mathbf{P}}}$ and

$0 \neq \hat{v}_2 \notin \mathcal{G}_{\hat{\mathbf{P}}}$, noting that there exist a $0 \neq \hat{u}_1 \in H^2_{\mathscr{U}}(\mathbb{C}_+)$ such that $\hat{v}_1 = \left[\begin{smallmatrix} M \\ N \end{smallmatrix}\right] \hat{u}_1$ and a $\hat{u}_2 \in H^2_{\mathscr{U}}(\mathbb{C}_+)$ and $0 \neq \hat{y} \in H^2_{\mathscr{Y}}(\mathbb{C}_+)$ such that $\hat{v}_2 = \left[\begin{smallmatrix} M \\ N \end{smallmatrix}\right] \hat{u}_2 + \left[\begin{smallmatrix} Y \\ X \end{smallmatrix}\right] \hat{y}$. It follows that $\tilde{G}\,\hat{v}_1 = \begin{bmatrix} -\tilde{N} & \tilde{M} \end{bmatrix} \left[\begin{smallmatrix} M \\ N \end{smallmatrix}\right] \hat{u}_1 = 0$ and $\tilde{G}\,\hat{v}_2 = \begin{bmatrix} -\tilde{N} & \tilde{M} \end{bmatrix} \left[\begin{smallmatrix} Y \\ X \end{smallmatrix}\right] \hat{y} = \hat{y} \neq 0$, which implies that $\mathcal{G}_{\hat{\mathbf{P}}} = \left\{ \hat{v} \in H^2_{\left[\begin{smallmatrix} \mathscr{U} \\ \mathscr{Y} \end{smallmatrix}\right]}(\mathbb{C}_+) \;\middle|\; \tilde{G}\,\hat{v} = 0 \right\}$. It remains to show that $\tilde{G}$ can be normalized. Since $\tilde{G}$ is right invertible, it follows by Proposition 1.12(ii) that there exists a function $Y \in H^\infty_{\mathscr{B}(\mathscr{Y},\,\mathscr{Y})}(\mathbb{C}_+)$, invertible in H^∞, such that $\tilde{G}(\jmath\omega)\,\tilde{G}^\sim(\jmath\omega) = Y(\jmath\omega)\,Y^\sim(\jmath\omega)$ for (almost) all $\omega \in \mathbb{R}$. It is now straightforward to verify that $(Y^{-1}\,\tilde{G})$ is a normalized left coprime representation of $\mathbf{P}$. □

PROPOSITION 2.8 (Youla-parametrization of all stabilizing controllers). *Let* $\mathbf{P} \in \mathcal{SS}\big(\mathscr{U}, \mathscr{Y}; [0,\infty)\big)$. *Consider any (operator-valued) functions* M, $\tilde{X} \in H^\infty_{\mathscr{B}(\mathscr{U},\,\mathscr{U})}(\mathbb{C}_+)$; Y, $\tilde{Y} \in H^\infty_{\mathscr{B}(\mathscr{Y},\,\mathscr{U})}(\mathbb{C}_+)$; N, $\tilde{N} \in H^\infty_{\mathscr{B}(\mathscr{U},\,\mathscr{Y})}(\mathbb{C}_+)$; *and* X, $\tilde{M} \in H^\infty_{\mathscr{B}(\mathscr{Y},\,\mathscr{Y})}(\mathbb{C}_+)$ *with* $\left[\begin{smallmatrix} M \\ N \end{smallmatrix}\right]$ *a right coprime representation of* $\mathbf{P}$ *and* $\begin{bmatrix} -\tilde{N} & \tilde{M} \end{bmatrix}$ *a left coprime representation of* $\mathbf{P}$ *such that the following double Bezout identity holds:*

$$\begin{pmatrix} M & Y \\ N & X \end{pmatrix} \begin{pmatrix} \tilde{X} & -\tilde{Y} \\ -\tilde{N} & \tilde{M} \end{pmatrix} = \begin{pmatrix} \tilde{X} & -\tilde{Y} \\ -\tilde{N} & \tilde{M} \end{pmatrix} \begin{pmatrix} M & Y \\ N & X \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{\mathscr{U}} & 0 \\ 0 & \mathbf{I}_{\mathscr{Y}} \end{pmatrix}.$$

Then $\mathbf{C} \in \mathcal{SS}\big(\mathscr{Y}, \mathscr{U}; [0,\infty)\big)$ *stabilizes* $\mathbf{P}$ *iff* $\mathbf{C}$ *has a right coprime representation* $\left[\begin{smallmatrix} Y - M\,Q \\ X - N\,Q \end{smallmatrix}\right]$ *and a left coprime representation* $\begin{bmatrix} (-\tilde{Y} + Q\,\tilde{M}) & (\tilde{X} - Q\,\tilde{N}) \end{bmatrix}$ *for some* $Q \in H^\infty_{\mathscr{B}(\mathscr{Y},\,\mathscr{U})}(\mathbb{C}_+)$.

Proof. The result follows largely by arguments in [F], with extra care needed to ensure invertibility of certain frequency-domain symbols in H^∞, as in the first part of the proof of Theorem 2.7 above. □

3. The gap-metric. Consider two systems $\mathbf{P}_1, \mathbf{P}_2 \in \mathcal{SS}\big(\mathscr{U}, \mathscr{Y}; [0,\infty)\big)$. Let their (Laplace-transformed) graphs be denoted by $\mathcal{G}_1 := \mathcal{G}_{\hat{\mathbf{P}}_1}$ and $\mathcal{G}_2 := \mathcal{G}_{\hat{\mathbf{P}}_2}$, and let the inner functions G_1 and G_2 denote the respective graph symbols. The *gap* $\delta_g(.,.)$ between the two systems $\mathbf{P}_1$ and $\mathbf{P}_2$ is defined to be the aperture between their graphs (cf. [K1, p. 197ff] and [ZE]), i.e., $\delta_g(\mathbf{P}_1, \mathbf{P}_2) := \|\mathcal{P}_{\mathcal{G}_1} - \mathcal{P}_{\mathcal{G}_2}\|_{\mathrm{ind}}$, where $\mathcal{P}_{\mathcal{G}_1}$ and $\mathcal{P}_{\mathcal{G}_2}$ denote the projection operators from $H^2_{\left[\begin{smallmatrix} \mathscr{U} \\ \mathscr{Y} \end{smallmatrix}\right]}(\mathbb{C}_+)$ onto the (closed) graphs $\mathcal{G}_1$ and $\mathcal{G}_2$. It is easy to verify that the gap $\delta_g(.,.)$ so defined is a metric. Furthermore, $0 \leq \delta_g(\mathbf{P}_1, \mathbf{P}_2) \leq 1$. In [KVZRS, p. 205f], it was shown that $\delta_g(\mathbf{P}_1, \mathbf{P}_2) = \max\big\{\vec{\delta}_g(\mathbf{P}_1, \mathbf{P}_2), \vec{\delta}_g(\mathbf{P}_2, \mathbf{P}_1)\big\}$, where $\vec{\delta}_g(\mathbf{P}_1, \mathbf{P}_2) := \big\|\mathcal{P}_{\mathcal{G}_2^\perp}\,\mathcal{P}_{\mathcal{G}_1}\big\|_{\mathrm{ind}}$, which is termed the *directed gap*. In addition,

$$\delta_g(\mathbf{P}_1, \mathbf{P}_2) = \vec{\delta}_g(\mathbf{P}_1, \mathbf{P}_2) = \vec{\delta}_g(\mathbf{P}_2, \mathbf{P}_1) \qquad \text{if} \quad \delta_g(\mathbf{P}_1, \mathbf{P}_2) < 1. \tag{7}$$

Generalizing the result of [G1] to the infinite-dimensional, linear operator case, it was shown in [CG2, Theorem 3.3] that

$$\vec{\delta}_g(\mathbf{P}_1, \mathbf{P}_2) = \Big\|\mathcal{P}_{\mathcal{G}_2^\perp}\,\mathcal{P}_{\mathcal{G}_1}\Big\|_{\mathrm{ind}} = \inf_{Q \in H^\infty_{\mathscr{B}(\mathscr{U},\,\mathscr{U})}(\mathbb{C}_+)} \|G_1 - G_2\,Q\|_\infty. \tag{8}$$

Building on the result of [G1], Sefton and Ober [SO] derived a useful formula for the gap (as opposed to the directed gap) between lumped-parameter LTI systems, which is similar to (8) except that Q is restricted to being invertible in H^∞. In what follows the Sefton–Ober gap formula is generalized to the case of spatially distributed LTI systems.

THEOREM 3.1. *Let the inner functions* $\left[\begin{smallmatrix}M_1\\N_1\end{smallmatrix}\right], \left[\begin{smallmatrix}M_2\\N_2\end{smallmatrix}\right] \in H^\infty_{\mathscr{B}(\mathscr{U},\left[\begin{smallmatrix}\mathscr{Y}\\\mathscr{U}\end{smallmatrix}\right])}(\mathbb{C}_+)$ *be graph symbols of the systems* $\mathbf{P}_1, \mathbf{P}_2 \in \mathcal{S}\big(\mathscr{U}, \mathscr{Y}; [0,\infty)\big)$.

(i) *If there exists a* $\tilde{Q} \in H^\infty_{\mathscr{B}(\mathscr{U},\mathscr{U})}(\mathbb{C}_+)$ *such that* $\| \left[\begin{smallmatrix}M_1\\N_1\end{smallmatrix}\right] - \left[\begin{smallmatrix}M_2\\N_2\end{smallmatrix}\right] \tilde{Q}\|_\infty < 1$ *and* $\tilde{Q}$ *is not invertible in* H^∞, *then* $\delta_g(\mathbf{P}_1, \mathbf{P}_2) = 1$.

(ii) *If there exists a* $\tilde{Q} \in H^\infty_{\mathscr{B}(\mathscr{U},\mathscr{U})}(\mathbb{C}_+)$ *such that* $\|G_1 - G_2\,\tilde{Q}\|_\infty < 1$ *and* $\tilde{Q}^{-1} \in H^\infty_{\mathscr{B}(\mathscr{U},\mathscr{U})}(\mathbb{C}_+)$, *then* $\delta_g(\mathbf{P}_1, \mathbf{P}_2) < 1$.

Proof. (i) This part follows as shown in the proof of Proposition 4.6 in [SO], provided it can be shown that when $\tilde{Q} \in H^\infty_{\mathscr{B}(\mathscr{U},\mathscr{U})}(\mathbb{C}_+)$ satisfies

$$\left\| \begin{bmatrix} M_1 \\ N_1 \end{bmatrix} - \begin{bmatrix} M_2 \\ N_2 \end{bmatrix} \tilde{Q} \right\|_\infty < 1, \tag{9}$$

then taking an inner-outer factorization (see [FF, p. 241]), the outer part is invertible in H^∞. To see that this is the case, first note that for any $\tilde{Q} \in H^\infty_{\mathscr{B}(\mathscr{U},\mathscr{U})}(\mathbb{C}_+)$ satisfying (9), $\mathcal{M}_{\tilde{Q}} : H^2_{\mathscr{U}}(\mathbb{C}_+) \to H^2_{\mathscr{U}}(\mathbb{C}_+)$ is bounded from below. This follows by assuming the contrary and taking a sequence $\hat{v}_n \in H^2_{\mathscr{U}}(\mathbb{C}_+)$ such that $\|\hat{v}_n\|_2 = 1$ for $n \in \mathbb{Z}_+$ and $\lim_{n\to\infty} \|\tilde{Q}\hat{v}_n\|_2 = 0$. Then $\lim_{n\to\infty} \| \left[\begin{smallmatrix}M_1\\N_1\end{smallmatrix}\right] \hat{v}_n - \left[\begin{smallmatrix}M_2\\N_2\end{smallmatrix}\right] \tilde{Q}\hat{v}_n\|_2 = 1$, which contradicts (9). Now taking an inner-outer factorization $\tilde{Q} = \tilde{Q}_i\,\tilde{Q}_o$, where $\tilde{Q}_i \in H^\infty_{\mathscr{B}(\mathscr{V},\mathscr{U})}(\mathbb{C}_+)$ is inner, $\tilde{Q}_o \in H^\infty_{\mathscr{B}(\mathscr{U},\mathscr{V})}(\mathbb{C}_+)$ is outer, and $\mathscr{V}$ is a Hilbert space of suitable dimension, it follows that $\mathcal{M}_{\tilde{Q}_o} : H^2_{\mathscr{U}}(\mathbb{C}_+) \to H^2_{\mathscr{V}}(\mathbb{C}_+)$ is bounded from below, since $\mathcal{M}_{\tilde{Q}_i} : H^2_{\mathscr{V}}(\mathbb{C}_+) \to H^2_{\mathscr{U}}(\mathbb{C}_+)$ is isometric. But $\mathcal{M}_{\tilde{Q}_o}$ has dense range, since $\tilde{Q}_o$ is outer, which by Theorem 1.1 implies that $\mathcal{M}_{\tilde{Q}_o}$ is boundedly invertible. Hence, $\tilde{Q}_o$ is invertible in H^∞ by Theorem 1.7, as claimed.

(ii) This part of the result follows by the same arguments used to prove Theorem 4.7 in [SO]. ∎

The Sefton–Ober gap formula is a direct consequence of Theorem 3.1 as stated below.

COROLLARY 3.2. *Let the inner functions* $G_1, G_2 \in H^\infty_{\mathscr{B}(\mathscr{U},\left[\begin{smallmatrix}\mathscr{Y}\\\mathscr{U}\end{smallmatrix}\right])}(\mathbb{C}_+)$ *be graph symbols of the systems* $\mathbf{P}_1, \mathbf{P}_2 \in \mathcal{S}\big(\mathscr{U}, \mathscr{Y}; [0,\infty)\big)$. *Then* $\delta_g(\mathbf{P}_1, \mathbf{P}_2) = \inf_{Q,\, Q^{-1} \in H^\infty_{\mathscr{B}(\mathscr{U},\mathscr{U})}(\mathbb{C}_+)} \|G_1 - G_2\,Q\|_\infty$.

4. Robust feedback stability. In this section we establish a quantitative robust stability result for feedback loops comprised of spatially distributed LTI systems subject to gap-metric perturbations. First, we present a qualitative result which demonstrates that the gap-metric is a measure of the difference between open-loop systems with respect to variation in closed-loop performance.

PROPOSITION 4.1. *With* $\mathbf{P}, \mathbf{P}_i \in \mathcal{SS}\big(\mathscr{U}, \mathscr{Y}; [0,\infty)\big)$ *for* $i \in \mathbb{Z}_+$, *the following statements are equivalent:*

(a) $\delta_g(\mathbf{P}_i, \mathbf{P}) \to 0$ *as* $i \to \infty$;

(b) $\|\mathbf{H}(\mathbf{P}_i,\mathbf{C}) - \mathbf{H}(\mathbf{P},\mathbf{C})\|_{\text{ind}} \to 0$ *as* $i \to \infty$ *for any* $\mathbf{C}$ *stabilizing* $\mathbf{P}$;

(c) *there exist right coprime representations* $\left[\begin{smallmatrix}M_i\\N_i\end{smallmatrix}\right], \left[\begin{smallmatrix}M\\N\end{smallmatrix}\right]$ *of* $\mathbf{P}_i, \mathbf{P}$ *such that* $\left\| \left[\begin{smallmatrix}M_i\\N_i\end{smallmatrix}\right] - \left[\begin{smallmatrix}M\\N\end{smallmatrix}\right] \right\|_\infty \to 0$ *as* $i \to \infty$.

Proof. (a) ⇔ (b) is a direct consequence of Theorem 4.1 in [C1] and sufficiency of Theorem 3 in [FGS]. The equivalence (c) ⇔ (a) follows directly from Corollary 3.2. ∎

Suppose that $[\mathbf{P}, \mathbf{C}]$ is stable and that $\mathbf{P}$ has a right coprime representation $\left[\begin{smallmatrix}M\\N\end{smallmatrix}\right]$. By Proposition 4.1, every plant $\mathbf{P}'$, with a (perturbed) right coprime representation

$\left[\begin{smallmatrix} M' \\ N' \end{smallmatrix}\right] = \left[\begin{smallmatrix} M+\Delta_M \\ N+\Delta_N \end{smallmatrix}\right]$, will be stabilized by $\mathbf{C}$ provided that the perturbation $\left[\begin{smallmatrix} \Delta_M \\ \Delta_N \end{smallmatrix}\right]$, commonly referred to as the *(right) coprime factor uncertainty,* has a sufficiently small ∞-norm. The following result shows that a *normalized* coprime factor ball of systems is equivalent to a gap-metric ball of systems.

PROPOSITION 4.2. *Given* $\mathbf{P} \in \mathcal{SS}\big(\mathscr{U}, \mathscr{Y}; [0,\infty)\big)$, *let* $\left[\begin{smallmatrix} M \\ N \end{smallmatrix}\right]$ *be a normalized right coprime representation of* $\mathbf{P}$. *For all* $b:\ 0 < b \le 1$ *we have*

$$\begin{aligned} &\big\{\mathbf{P}' \in \mathcal{SS}\big(\mathscr{U}, \mathscr{Y}; [0,\infty)\big) \,\big|\, \delta_g(\mathbf{P}, \mathbf{P}') < b\big\} \\ &\quad = \big\{\mathbf{P}' \in \mathcal{SS}\big(\mathscr{U}, \mathscr{Y}; [0,\infty)\big) \,\big|\, \left[\begin{smallmatrix} M+\Delta_M \\ N+\Delta_N \end{smallmatrix}\right] \\ &\qquad \text{is a right coprime representation of } \mathbf{P}' \text{ with } \left\| \begin{smallmatrix} \Delta_M \\ \Delta_N \end{smallmatrix} \right\|_\infty < b\big\}. \end{aligned}$$

Proof. The proof is the same as that of Theorem 5.2 in [SO] for the lumped-parameter LTI case. □

Given $\mathbf{P} \in \mathcal{SS}\big(\mathscr{U}, \mathscr{Y}; [0,\infty)\big)$ and $\mathbf{C} \in \mathcal{SS}\big(\mathscr{Y}, \mathscr{U}; [0,\infty)\big)$ such that the standard feedback configuration $[\mathbf{P}, \mathbf{C}]$ in Figure 1 is stable, let

$$b_{\mathbf{P},\mathbf{C}} := \left\| \begin{bmatrix} \mathbf{I}_\mathscr{U} \\ \mathbf{P} \end{bmatrix} \begin{bmatrix} \big(\mathbf{I}_\mathscr{U} - \mathbf{C}\,\mathbf{P}\big)^{-1} & -\mathbf{C}\big(\mathbf{I}_\mathscr{Y} - \mathbf{P}\,\mathbf{C}\big)^{-1} \end{bmatrix} \right\|_{\mathrm{ind}}^{-1},$$

and define

$$b_{\mathrm{opt}}(\mathbf{P}) := \sup_{\mathbf{C} \in \mathcal{SS}(\mathscr{U}, \mathscr{Y}, [0,\infty))} b_{\mathbf{P},\mathbf{C}}. \tag{10}$$

With $G = \left[\begin{smallmatrix} M \\ N \end{smallmatrix}\right]$ as a normalized right coprime representation of $\mathbf{P}$ and $\begin{bmatrix} -\tilde{U} & \tilde{V} \end{bmatrix}$ as a left coprime representation of $\mathbf{C}$, setting $\tilde{K} := \begin{bmatrix} \tilde{V} & -\tilde{U} \end{bmatrix}$ it follows that $(\tilde{K}G)^{-1} \in H^\infty_{\mathscr{B}(\mathscr{U},\mathscr{U})}(\mathbb{C}_+) \Leftrightarrow [\mathbf{P}, \mathbf{C}]$ is stable, and $b_{\mathbf{P},\mathbf{C}} = \big\|\big(\tilde{K}G\big)^{-1}\tilde{K}\big\|_\infty^{-1}$. Furthermore, $b_{\mathbf{P},\mathbf{C}}$ satisfies (see, e.g., [FGS]) $0 \le b_{\mathbf{P},\mathbf{C}} \le 1$ and $b_{\mathbf{P},\mathbf{C}} = b_{\mathbf{C},\mathbf{P}}$. The number $b_{\mathbf{P},\mathbf{C}}$ is often called the *stability margin.* The reason for this terminology is evident from the following result.

THEOREM 4.3 (robustness under plant gap uncertainty). *Assume that* $\mathbf{P} \in \mathcal{SS}\big(\mathscr{U}, \mathscr{Y}; [0,\infty)\big)$ *and* $\mathbf{C} \in \mathcal{SS}\big(\mathscr{Y}, \mathscr{U}; [0,\infty)\big)$ *are such that* $[\mathbf{P}, \mathbf{C}]$ *is stable and* $0 < b_{\mathbf{P},\mathbf{C}} < b_{\mathrm{opt}}(\mathbf{P})$. *Then* $[\mathbf{P}', \mathbf{C}]$ *is stable for all* $\mathbf{P}' \in \mathcal{SS}\big(\mathscr{U}, \mathscr{Y}; [0,\infty)\big)$ *such that* $\delta_g(\mathbf{P}, \mathbf{P}') < b$ *iff* $b \le b_{\mathbf{P},\mathbf{C}}$.

Proof. (i) "if": Using Proposition 4.2, this can be proved as for lumped-parameter LTI systems (see Theorem 9.6 in [ZDG], for example). Alternatively, it follows from [FGS, Theorem 3].

(ii) "only if": Let $G := \left[\begin{smallmatrix} M \\ N \end{smallmatrix}\right]$, where $M \in H^\infty_{\mathscr{B}(\mathscr{U},\mathscr{U})}(\mathbb{C}_+)$ and $N \in H^\infty_{\mathscr{B}(\mathscr{U},\mathscr{Y})}(\mathbb{C}_+)$, be a normalized right coprime representation of $\mathbf{P}$. Also let $\begin{bmatrix} -\tilde{U} & \tilde{V} \end{bmatrix}$ be a left coprime representation of $\mathbf{C}$, and set $\tilde{K} := \begin{bmatrix} \tilde{V} & -\tilde{U} \end{bmatrix}$. Assuming that the claim does not hold, we now show that it is then possible to construct a right coprime representation of a system $\mathbf{P}' \in \mathcal{SS}\big(\mathscr{U}, \mathscr{Y}; [0,\infty)\big)$ such that $\delta_g(\mathbf{P}, \mathbf{P}') < b$ with $b_{\mathbf{P},\mathbf{C}} < b \le b_{\mathrm{opt}}(\mathbf{P})$ and $[\mathbf{P}', \mathbf{C}]$ unstable. The right coprime representation of the required $\mathbf{P}'$ takes the form $(G + \Delta) \in H^\infty_{\mathscr{B}(\mathscr{U},\left[\begin{smallmatrix} \mathscr{U} \\ \mathscr{Y} \end{smallmatrix}\right])}(\mathbb{C}_+)$, where $\|\Delta\|_\infty < b$ (cf. Proposition 4.2.)

Before going on to construct an appropriate Δ, we make the following observations. First, note that provided $\|\Delta\|_\infty < b_{\mathrm{opt}}(\mathbf{P})$, then $(G+\Delta) \in H^\infty_{\mathscr{B}(\mathscr{U},\left[\begin{smallmatrix} \mathscr{U} \\ \mathscr{Y} \end{smallmatrix}\right])}(\mathbb{C}_+)$ is left invertible in H^∞, which can be shown as follows. By the definition of $b_{\mathrm{opt}}(\mathbf{P})$, there exists a $\mathbf{C}_0$ with left coprime representation $\begin{bmatrix} -\tilde{U}_0 & \tilde{V}_0 \end{bmatrix}$ such that $\tilde{K}_0 := \begin{bmatrix} \tilde{V}_0 & -\tilde{U}_0 \end{bmatrix}$

satisfies $b_{\mathrm{opt}}^{-1}(\mathbf{P}) \le b_{\mathbf{P},\mathbf{C}_0}^{-1} = \left\|\left(\tilde{K}_0 G\right)^{-1}\tilde{K}_0\right\|_\infty < \|\Delta\|_\infty^{-1}$. Setting $X := \left(\tilde{K}_0 G\right)^{-1}\tilde{K}_0$, and noting that $X(G+\Delta) = \mathbf{I}_{\mathscr{U}} + X\Delta$, where $\|X\Delta\|_\infty \le \|X\|_\infty \|\Delta\|_\infty < 1$, it follows that $X(G+\Delta)$ is invertible in H^∞. Then $X_0 := \big(X(G+\Delta)\big)^{-1}X \in H^\infty_{\mathscr{B}(\left[\begin{smallmatrix}\mathscr{U}\\ \mathscr{Y}\end{smallmatrix}\right],\,\mathscr{U})}(\mathbb{C}_+)$ is a left inverse of $(G+\Delta)$. Second, note that if $(G+\Delta) \in H^\infty_{\mathscr{B}(\mathscr{U},\,\left[\begin{smallmatrix}\mathscr{U}\\ \mathscr{Y}\end{smallmatrix}\right])}(\mathbb{C}_+)$ is a right coprime representation of a system $\mathbf{P}' \in \mathcal{SS}\big(\mathscr{U}, \mathscr{Y}; [0,\infty)\big)$, then $[\mathbf{P}', \mathbf{C}]$ is stable iff $\big(\mathbf{I}_{\mathscr{U}} + (\tilde{K}G)^{-1}\tilde{K}\Delta\big)^{-1} \in H^\infty_{\mathscr{B}(\mathscr{U},\,\mathscr{U})}(\mathbb{C}_+)$. This follows from the fact that $\left(\tilde{K}\,G\right)^{-1}\tilde{K}$ is a left coprime representation of $\mathbf{C}$.

Now, since $\left\|\left(\tilde{K}G\right)^{-1}\tilde{K}\right\|_\infty = b_{\mathbf{P},\mathbf{C}}^{-1} > b^{-1}$, it follows that for arbitrary small $\sigma > 0$, there exists a frequency $\omega_0 \in (0,\infty)$ and a unit vector $\hat{v} \in \left[\begin{smallmatrix}\mathscr{U}\\ \mathscr{Y}\end{smallmatrix}\right]$ such that $\hat{u} := \big(\tilde{K}(\sigma+\jmath\omega_0)\,G(\sigma+\jmath\omega_0)\big)^{-1}\tilde{K}(\sigma+\jmath\omega_0)\,\hat{v}$ satisfies $b_{\mathbf{P},\mathbf{C}}^{-1} \ge \|\hat{u}\|_{\mathscr{U}} > b^{-1}$. Define $p(s) := \frac{s}{s^2+2\delta s+\omega_0^2}$ and $\Delta(s) := \frac{-p(s)}{p(\sigma+\jmath\omega_0)}\,\frac{\mathrm{e}^{-\alpha s}}{\mathrm{e}^{-\alpha(\sigma+\jmath\omega_0)}}\,\frac{\hat{v}}{\|\hat{u}\|_{\mathscr{U}}^2}\,\langle\hat{u},\,.\rangle_{\mathscr{U}}$, where $\delta > 0$ and $\alpha > 0$. It follows that $\|\Delta\|_\infty = \frac{|p(\jmath\omega_0)|}{|p(\sigma+\jmath\omega_0)|}\,\mathrm{e}^{\alpha\sigma}\,\|\hat{u}\|_{\mathscr{U}}^{-1}$. Since $\|\hat{u}\|_{\mathscr{U}}^{-1} < b$ and $\frac{|p(\jmath\omega_0)|}{|p(\sigma+\jmath\omega_0)|}\,\mathrm{e}^{\alpha\sigma} = \frac{\sqrt{(\sigma^2+2\delta\sigma)^2+4\omega_0^2(\sigma+\delta)^2}}{2\delta\sqrt{\sigma^2+\omega_0^2}}\,\mathrm{e}^{\alpha\sigma} \to 1$ as $\delta \to \infty$ and $\alpha \to 0$, we can choose $\delta > 0$ sufficiently large and $\alpha > 0$ sufficiently small so that $\|\Delta\|_\infty < b$. Furthermore, by construction

$$\Big(\mathbf{I}_{\mathscr{U}} + \big(\tilde{K}(\sigma+\jmath\omega_0)\,G(\sigma+\jmath\omega_0)\big)^{-1}\tilde{K}(\sigma+\jmath\omega_0)\,\Delta(\sigma+\jmath\omega_0)\Big)\,\hat{u} = 0,$$

and hence $(\mathbf{I}_{\mathscr{U}} + \left(\tilde{K}\,G\right)^{-1}\tilde{K}\,\Delta)$ cannot be invertible in H^∞. Therefore, if $\big(G+\Delta\big)$ corresponds to a right coprime representation of a system $\mathbf{P}' \in \mathcal{SS}\big(\mathscr{U}, \mathscr{Y}; [0,\infty)\big)$, then the feedback loop $[\mathbf{P}', \mathbf{C}]$ is unstable as claimed. The remainder of the proof is devoted to showing that indeed $\mathbf{P}' \in \mathcal{SS}\big(\mathscr{U}, \mathscr{Y}; [0,\infty)\big)$.

Since the perturbation constructed satisfies $\|\Delta\|_\infty < b_{\mathrm{opt}}(\mathbf{P}) \le 1$, it follows that $(G+\Delta)$ is a right coprime representation of a closed, linear, shift-invariant system $\mathbf{P}'$. Furthermore, using the result of part (i) and the fact that $\delta_g(\mathbf{P}, \mathbf{P}') < b_{\mathrm{opt}}(\mathbf{P})$ (cf. Proposition 4.2), we see that $\mathbf{P}'$ is also stabilizable. For $\mathbf{P}'$ to be in $\mathcal{SS}\big(\mathscr{U}, \mathscr{Y}; [0,\infty)\big)$ it remains to show that $\mathbf{P}'$ is causally extendible. The following proof that $\mathbf{P}'$ is causally extendible uses arguments similar to those in [CG2]. Partition $(G+\Delta) =: \left[\begin{smallmatrix}M'\\ N'\end{smallmatrix}\right]$ such that $M' \in H^\infty_{\mathscr{B}(\mathscr{U},\,\mathscr{U})}(\mathbb{C}_+)$ and $N' \in H^\infty_{\mathscr{B}(\mathscr{U},\,\mathscr{Y})}(\mathbb{C}_+)$. Define $\mathbf{M} := \mathcal{L}^{-1}\,\mathcal{M}_M\,\mathcal{L}$, $\mathbf{N} := \mathcal{L}^{-1}\,\mathcal{M}_N\,\mathcal{L}$, $\mathbf{M}' := \mathcal{L}^{-1}\,\mathcal{M}_{M'}\,\mathcal{L}$, and $\mathbf{N}' := \mathcal{L}^{-1}\,\mathcal{M}_{N'}\,\mathcal{L}$, and recall that for $\mathbf{P}'$ to be causally extendible, we must show that $\mathbf{P}'$ is causal and that $\mathbf{T}^\tau \mathrm{dom}(\mathbf{P}') = L^2_{\mathscr{U}}[0,\tau)$ for all $\tau > 0$. We will first show that $\mathbf{P}'$ is causal. Pick a $\tau > 0$ and a $v \in L^2_{\mathscr{U}}[0,\infty)$ such that $\mathbf{T}^\tau\,\mathbf{M}'\,v = 0$. Setting $\widetilde{\mathscr{U}} := L^2_{\mathscr{U}}[0,\alpha)$, define the time-lifting isomorphism $\mathcal{W}\colon L^2_{\mathscr{U}}[0,\infty) \to \ell^2_{\widetilde{\mathscr{U}}}[0,\infty)$ via

$$\widetilde{f}_k(\theta) := \big(\mathcal{W}f\big)_k(\theta) := f(k\,\alpha+\theta), \qquad f \in L^2_{\mathscr{U}}[0,\infty), \quad \theta \in [0,\alpha), \quad k = 0,1,2,\ldots\,.$$

Representing signals in $\ell^2_{\widetilde{\mathscr{U}}}[0,\infty)$ by (infinite-dimensional) column vectors with the kth entry corresponding to the value of the signal at the kth time instant, which

corresponds to the continuous-time interval $\big[k\,\alpha, (k+1)\,\alpha\big)$, we obtain

(11)

$$0 = \mathbf{T}^{\tau}\,\mathbf{M}'\,v = \mathbf{T}^{\tau}\,\mathcal{W}^{-1}\,[\mathcal{W}\,\mathbf{M}'\,\mathcal{W}^{-1}]\,[\mathcal{W}\,v] = \mathbf{T}^{\tau}\,\mathcal{W}^{-1}\begin{bmatrix} \overbrace{\begin{bmatrix} \widetilde{\mathbf{M}}'_0 & 0 & \cdots & 0 \\ \widetilde{\mathbf{M}}'_1 & \widetilde{\mathbf{M}}'_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \widetilde{\mathbf{M}}'_N & \cdots & \widetilde{\mathbf{M}}'_1 & \widetilde{\mathbf{M}}'_0 \end{bmatrix}}^{=:\,\left[\widetilde{\mathbf{M}}'\right]_N} \begin{bmatrix} \widetilde{v}_0 \\ \widetilde{v}_1 \\ \vdots \\ \widetilde{v}_N \end{bmatrix} \\ \begin{matrix} 0 \\ \vdots \end{matrix} \end{bmatrix},$$

where $N :=$ the integer part of τ/α, $\widetilde{v}_k := (\mathcal{W}\,v)_k$, and $\widetilde{\mathbf{M}}'_k \colon \widetilde{\mathscr{U}} \to \widetilde{\mathscr{U}}$ is the kth entry of the sequence uniquely identifiable with the block-Toeplitz representation of the causal, discrete-time operator $\mathcal{W}\,\mathbf{M}'\,\mathcal{W}^{-1} \colon \ell^2_{\widetilde{\mathscr{U}}}[0,\infty) \to \ell^2_{\widetilde{\mathscr{U}}}[0,\infty)$. Now, noting that $\mathbf{M}' = \mathbf{M} + \mathbf{S}^{\alpha}\boldsymbol{\Delta}'$, where $\boldsymbol{\Delta}'$ is a causal system, it follows that $\widetilde{\mathbf{M}}'_0 = \widetilde{\mathbf{M}}_0$, where $\widetilde{\mathbf{M}}_0$ is the 0th entry of the block-Toeplitz representation of $\mathcal{W}\,\mathbf{M}\,\mathcal{W}^{-1}$. As such, $[\widetilde{\mathbf{M}}']_N$ defined in (11) is boundedly invertible because of its block lower triangular structure and since $\widetilde{\mathbf{M}}_0$ is boundedly invertible as shown next. First note that since $\mathbf{P}$ is causal, $\mathbf{T}^{\alpha}\,\mathbf{M}\,u = 0$ implies $\mathbf{T}^{\alpha}\,\mathbf{N}\,u = 0$ for all $u \in L^2_{\mathscr{U}}[0,\infty)$. Next, since $\left[\begin{smallmatrix} M \\ N \end{smallmatrix}\right]$ is a right *coprime* representation of $\mathbf{P}$, $\left[\begin{smallmatrix} \mathbf{M} \\ \mathbf{N} \end{smallmatrix}\right]$ has a bounded, shift-invariant (and hence causal) left inverse, $\mathbf{K}$. Therefore, $\mathbf{T}^{\alpha}\,u = \mathbf{T}^{\alpha}\,\mathbf{K}\left[\begin{smallmatrix} \mathbf{M} \\ \mathbf{N} \end{smallmatrix}\right] u = \mathbf{T}^{\alpha}\,\mathbf{K}\,\mathbf{T}^{\alpha}\left[\begin{smallmatrix} \mathbf{M} \\ \mathbf{N} \end{smallmatrix}\right] u = 0$ if $\mathbf{T}^{\alpha}\,\mathbf{M}\,u = \mathbf{T}^{\alpha}\,\mathbf{M}\,\mathbf{T}^{\alpha}\,u = 0$, which implies that $\widetilde{\mathbf{M}}_0 \colon \widetilde{\mathscr{U}} \to \widetilde{\mathscr{U}}$ is injective. Moreover, since $\mathbf{P}$ is causally extendible, we have that $\mathbf{T}^{\alpha}\,\mathbf{M}\,L^2_{\mathscr{U}}[0,\infty) = \mathbf{T}^{\alpha}\,\mathrm{dom}(\mathbf{P}) = \mathbf{T}^{\alpha}L^2_{\mathscr{U}}[0,\infty) = \widetilde{\mathscr{U}}$, implying that $\widetilde{\mathbf{M}}_0 \colon \widetilde{\mathscr{U}} \to \widetilde{\mathscr{U}}$ is surjective. Consequently, $\widetilde{\mathbf{M}}_0$, and hence $[\widetilde{\mathbf{M}}']_N$, are boundedly invertible. In view of this, (11) implies that $\mathbf{T}^{\tau}\,v = 0$, by which it follows that $\mathbf{T}^{\tau}\,\mathbf{N}'\,v = \mathbf{T}^{\tau}\,\mathbf{N}'\,\mathbf{T}^{\tau}\,v = 0$, where the first equality is a consequence of the fact that $\mathbf{N}'$ is a causal operator, since it is bounded and shift-invariant. Correspondingly, we have shown that, for all $\tau > 0$ and $v \in L^2_{\mathscr{U}}[0,\infty)$, $\mathbf{T}^{\tau}\left[\begin{smallmatrix} \mathbf{M}' \\ \mathbf{N}' \end{smallmatrix}\right] v = \left[\begin{smallmatrix} 0 \\ y \end{smallmatrix}\right]$ implies $y = 0$. That is, $\mathbf{P}'$ is causal. Causal extendibility of $\mathbf{P}'$ now follows by noting that, since $\left[\widetilde{\mathbf{M}}'\right]_N$ has full range for all $N = 0, 1, 2, \ldots$, for every $\tau > 0$ (taking $N > \tau/\alpha$)

$$\mathbf{T}^{\tau}\,\mathrm{dom}(\mathbf{P}') = \mathbf{T}^{\tau}\,\mathbf{M}'\,L^2_{\mathscr{U}}[0,\infty) = \mathbf{T}^{\tau}\,\mathcal{W}^{-1}\begin{bmatrix} \left[\widetilde{\mathbf{M}}'\right]_N \overbrace{\left(\widetilde{\mathscr{U}} \times \cdots \times \widetilde{\mathscr{U}}\right)}^{N \text{ times}} \\ 0 \\ \vdots \end{bmatrix} = \mathbf{T}^{\tau}\,L^2_{\mathscr{U}}[0, N\alpha)$$

$$= L^2_{\mathscr{U}}[0,\tau).$$

This completes the proof. □

COROLLARY 4.4 (robustness under controller gap uncertainty). *Let* $\mathbf{P} \in \mathcal{SS}\big(\mathscr{U}, \mathscr{Y}; [0,\infty)\big)$ *and* $\mathbf{C} \in \mathcal{SS}\big(\mathscr{Y}, \mathscr{U}; [0,\infty)\big)$ *be such that* $[\mathbf{P}, \mathbf{C}]$ *is stable and* $0 < b_{\mathbf{P},\mathbf{C}} < b_{\mathrm{opt}}(\mathbf{P})$. *Then* $b \le b_{\mathbf{P},\mathbf{C}}$ *iff* $[\mathbf{P}, \mathbf{C}']$ *is stable for all* $\mathbf{C}' \in \mathcal{SS}\big(\mathscr{U}, \mathscr{Y}; [0,\infty)\big)$ *with* $\delta_g(\mathbf{C}, \mathbf{C}') < b$.

Proof. Thinking of the plant and the controller interchanged and recalling that $b_{\mathbf{P},\mathbf{C}} = b_{\mathbf{C},\mathbf{P}}$, the result follows immediately by Theorem 4.3. □

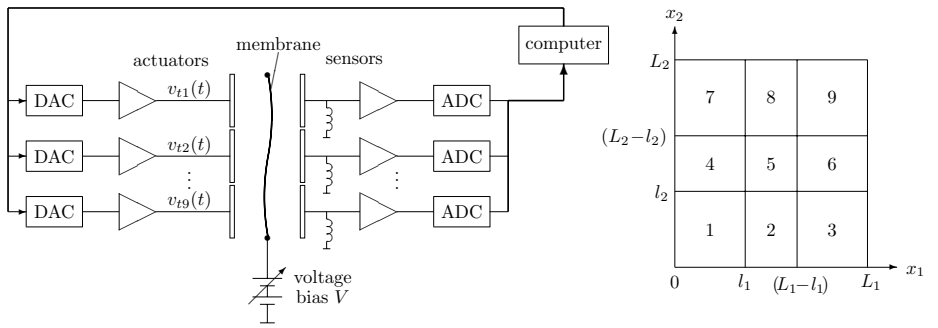


FIG. 2. *Left: Functional diagram of the electrostatically destabilized membrane with feedback arrangement for stabilization. Right: Geometry of actuator and sensor plates (assumed equal).*

For completeness, we quote a result on the stability of feedback loops with *simultaneous* plant and controller uncertainty, which was originally formulated for finite-dimensional, lumped-parameter, LTI systems in [QD], and which was later generalized to infinite-dimensional, time-varying systems in [FGS].

THEOREM 4.5 (robustness to plant and controller gap uncertainty [FGS]). *Given a plant* $\mathbf{P} \in \mathcal{SS}\big(\mathscr{U}, \mathscr{Y}; [0,\infty)\big)$ *and controller a* $\mathbf{C} \in \mathcal{SS}\big(\mathscr{Y}, \mathscr{U}; [0,\infty)\big)$*, fix nonnegative numbers* b_1 *and* b_2 *such that* $b_1^2 + b_2^2 < 1$*. Suppose that* $[\mathbf{P}, \mathbf{C}]$ *is stable and that* $b_1\sqrt{1-b_2^2} + b_2\sqrt{1-b_1^2} < b_{\mathbf{P},\mathbf{C}}$*. Then* $[\mathbf{P}', \mathbf{C}']$ *is stable for all* $\mathbf{P}' \in \mathcal{S}\big(\mathscr{U}, \mathscr{Y}; [0,\infty)\big)$ *and all* $\mathbf{C}' \in \mathcal{S}\big(\mathscr{Y}, \mathscr{U}; [0,\infty)\big)$ *which satisfy* $\delta_g(\mathbf{P}, \mathbf{P}') \le b_1$ *and* $\delta_g(\mathbf{C}, \mathbf{C}') \le b_2$*.*

5. Controller design example. Based on the control-theoretic framework described above, a controller design method was proposed in [R]. The design method, termed *coprime factor synthesis*, can be regarded as an extension or adaptation of "normalized coprime factor robust stabilization" [GM] and "H^∞ loop-shaping" [MG1, MG2] to the case of spatially distributed LTI systems. In the following we will briefly outline the method by means of a numerical example treating the feedback stabilization of an electrostatically destabilized, electrically conducting membrane. For details see [R].

The experimental setup is schematically depicted in Figure 2 and can be described as follows. A rectangular, flexible, electrically conducting membrane is suspended vertically, clamped at its boundaries, and biased by a high-voltage source V. The two spatial coordinates within the membrane plane shall be denoted by x_1 and x_2. Assuming the membrane deflection $\hat{y}(x_1, x_2; s)$ from the level-flat equilibrium position to be small, this (Laplace-transformed) quantity is governed by the PDE

$$\left(s^2\,\mu + s\,\delta - \frac{2\epsilon_0 V^2}{H^3}\right)\hat{y}(x_1,x_2;s) - \tau_1\,\frac{\partial^2\hat{y}(x_1,x_2;s)}{\partial x_1^2} - \tau_2\,\frac{\partial^2\hat{y}(x_1,x_2;s)}{\partial x_2^2} = -\frac{\epsilon_0\,V}{H^2}\,\hat{u}(x_1,x_2;s), \tag{12}$$

$\{x_1, x_2\} \in [0, L_1] \times [0, L_2]$, together with the boundary conditions $\hat{y}(x_1, x_2; s) = 0$ if $x_1 = 0$, or $x_1 = L_1$, or $x_2 = 0$, or $x_2 = L_1$. The distributed variable $\hat{u}(x_1, x_2; s)$ on the right-hand side of (12) stands for a distributed control voltage to be realized by the actuators. (The fact that the form of $\hat{u}(x_1, x_2; s)$ is restricted due to the discreteness of the actuators shall be disregarded for the moment.)

The parameters appearing in (12) have the following meaning and numerical values: membrane mass density $\mu = 0.033\,\mathrm{kg\,m^{-2}}$, viscous damping coefficient $\delta =$

$0.48\,\mathrm{kg\,m^{-2}\,s^{-1}}$, membrane tension in x_1-direction $\tau_1 = 8.4\,\mathrm{N/m}$, membrane tension in x_2-direction $\tau_2 = 7.8\,\mathrm{N/m}$, distance between membrane and actuator plates $H = 9.2\cdot 10^{-3}\,\mathrm{m}$, free space permittivity $\epsilon_0 = 8.85\cdot 10^{-12}\,\mathrm{As/V/m}$. The size of the actuator and sensor surfaces is assumed to be $L_1 \times L_2 = 1.04\,\mathrm{m} \times 1.12\,\mathrm{m}$.

The explicit input-output relationship of the system described by (12) is given by $\hat{y}(x_1, x_2; s) = \int_0^{L_1} \mathrm{d}\xi_1 \int_0^{L_2} \mathrm{d}\xi_2\, \kappa_{\hat{\mathbf{P}}^\infty}(x_1, x_2, \xi_1, \xi_2; s)\, \hat{u}(\xi_1, \xi_2; s)$, where

(13)
$$\kappa_{\hat{\mathbf{P}}^\infty}(x_1, x_2, \xi_1, \xi_2; s) = \sum_{j=1}^{\infty} \alpha_j(x_1, x_2) \left(-\frac{\epsilon_0 V}{H^2}\right) \left(s^2\,\mu + s\,\delta + \omega_j^2 - \Omega\right)^{-1} \beta_j(\xi_1, \xi_2)$$

with $\alpha_j(x_1, x_2) = \beta_j(x_1, x_2) := \frac{2}{\sqrt{L_1 L_2}} \sin(\frac{i\pi x_1}{L_1}) \sin(\frac{l\pi x_2}{L_2})$, $\omega_j := \sqrt{\tau_1 \left(\frac{i\pi}{L_1}\right)^2 + \tau_2 \left(\frac{l\pi}{L_2}\right)^2}$, and $\Omega := \frac{2\epsilon_0 V^2}{H^3}$. The mapping $(i, l) \mapsto j$ is defined such that $\{\omega_j\}_{j\in\mathbb{Z}_+}$ is an increasing sequence. From (13) it is clear that the first mode becomes unstable if V exceeds $\pi\sqrt{\frac{H^3}{2\epsilon_0}\left(\frac{\tau_1}{L_1^2} + \frac{\tau_2}{L_2^2}\right)} = 2460\,\mathrm{V}$; the second mode goes unstable at $V = 3760\,\mathrm{V}$, etc. We take $V = 2500\,\mathrm{V}$, whence precisely the first mode of the open-loop plant is unstable.

Let $\hat{\mathbf{P}}^\infty$ denote the integral operator whose kernel is given by (13); its finite-dimensional approximation, $\hat{\mathbf{P}}^{(10)}$, is obtained by truncating the infinite sum in (13) after the first $m = 10$ terms. We select the scalar, stable, stably invertible, weighting function $\hat{W}(s) = \frac{s+w_1}{s+w_2}$ with $w_1 = 3.1158 \cdot 10^6$ and $w_2 = 70$, and we define $\hat{\mathbf{P}}_W^\infty := \hat{\mathbf{P}}^\infty \cdot \hat{W}$ as well as $\hat{\mathbf{P}}_W^{(10)} := \hat{\mathbf{P}}^{(10)} \cdot \hat{W}$. Using the generalized Sefton–Ober gap formula (Corollary 3.2), the gap distance between the weighted infinite-dimensional plant model and the weighted finite-dimensional plant model can be evaluated as $\delta_g(\hat{\mathbf{P}}_W^\infty, \hat{\mathbf{P}}_W^{(10)}) = 0.0458$. By means of (10) and standard H^∞-techniques for finite-dimensional, lumped-parameter systems, the optimal stability margin of $\hat{\mathbf{P}}_W^{(10)}$ can be computed to be $b_{\mathrm{opt}}(\hat{\mathbf{P}}_W^{(10)}) = 0.3732$, where the optimally stabilizing controller, $\hat{\mathbf{C}}^{\mathrm{opt}}$, achieving this stability margin has a kernel of the form

$$\kappa_{\hat{\mathbf{C}}^{\mathrm{opt}}}(\xi_1, \xi_2, x_1, x_2; s) = E^\beta(\xi_1, \xi_2)\, \hat{C}^{\mathrm{opt}}(s) \left(E^\alpha(x_1, x_2)\right)^T$$

with $E^\alpha(x_1, x_2) := \big(\alpha_1(x_1, x_2), \ldots, \alpha_{10}(x_1, x_2)\big)$, $E^\beta(\xi_1, \xi_2) := \big(\beta_1(\xi_1, \xi_2), \ldots, \beta_{10}(\xi_1, \xi_2)\big)$, and $\hat{C}^{\mathrm{opt}}$ being a 10×10, finite-dimensional, lumped-parameter, transfer matrix.

The controller $\hat{\mathbf{C}}^{\mathrm{opt}}$ cannot be realized by the discrete sensors and actuators, as depicted in Figure 2. Therefore, we wish to find an implementable controller, $\hat{\mathbf{C}}^{\mathrm{imp}}$, whose kernel is of the form

$$\kappa_{\hat{\mathbf{C}}^{\mathrm{imp}}}(x_1, x_2, \xi_1, \xi_2, s) = E^{\alpha^c}(x_1, x_2)\, \hat{C}^{\mathrm{imp}}(s) \left(E^{\beta^c}(\xi_1, \xi_2)\right)^T,$$

where $E^{\alpha^c}(x_1, x_2) = E^{\beta^c}(x_1, x_2) = \big(\alpha_1^c(x_1, x_2), \ldots, \alpha_9^c(x_1, x_2)\big)$ with

$$\alpha_1^c(x_1, x_2) := \begin{cases} (l_1 \cdot l_2)^{-1/2} & \text{if } 0 < x_1 < l_1 \text{ and } 0 < x_2 < l_2, \\ 0 & \text{otherwise,} \end{cases}$$

etc. The parameters l_1 and l_2 as well as the 9×9, lumped-parameter, transfer matrix $\hat{C}^{\mathrm{imp}}(s)$ are to be determined such that the gap distance $\delta_g\big(\hat{\mathbf{C}}^{\mathrm{opt}}, \hat{\mathbf{C}}^{\mathrm{imp}}\big)$ is sufficiently small for the implementable controller to stabilize the weighted, infinite-dimensional, plant model $\hat{\mathbf{P}}_W^\infty$ according to Theorem 4.5. By means of a direct, global search in l_1

and l_2, in conjunction with standard H^∞-techniques for computing $\hat{C}^{\mathrm{imp}}(s)$, we found $l_1 = 0.20$, $l_2 = 0.22$, and $C^{\mathrm{imp}}(s)$ with (minimal) state-space realization $[A, B, C, D]$, where

$$A = \begin{pmatrix} -905.0798 & -193.9911 & -109.3168 \\ 193.9911 & -1.2967 & -5.4102 \\ -109.3168 & 5.4102 & -93.9223 \end{pmatrix},$$

$$B = \begin{pmatrix} -1.667 & -8.664 & -1.667 & -8.385 & -43.583 & -8.385 & -1.667 & -8.664 & -1.667 \\ 0.053 & 0.275 & 0.053 & 0.266 & 1.384 & 0.266 & 0.053 & 0.275 & 0.053 \\ -0.104 & -0.543 & -0.104 & -0.526 & -2.732 & -0.526 & -0.104 & -0.543 & -0.104 \end{pmatrix},$$

$$C = \begin{pmatrix} -1.6668 & -0.0530 & -0.1045 \\ -8.6642 & -0.2752 & -0.5432 \\ -1.6668 & -0.0530 & -0.1045 \\ -8.3847 & -0.2663 & -0.5256 \\ -43.5833 & -1.3838 & -2.7316 \\ -8.3847 & -0.2663 & -0.5256 \\ -1.6668 & -0.0530 & -0.1045 \\ -8.6642 & -0.2752 & -0.5432 \\ -1.6668 & -0.0530 & -0.1045 \end{pmatrix}, \qquad D = 0_{9\times 9},$$

to be nearly optimal. For the gap distance between optimal and implementable controllers we obtained $\delta_g(\hat{\mathbf{C}}^{\mathrm{opt}}, \hat{\mathbf{C}}^{\mathrm{imp}}) = 0.3195$, thus satisfying the inequality in Theorem 4.5. Since the weight $\hat{W}(s)$ is scalar, stable, and stably invertible, $\hat{\mathbf{C}}^{\mathrm{imp}}$ stabilizing $\hat{\mathbf{P}}_W^\infty$ implies that the implementable controller $\hat{\mathbf{C}}^{\mathrm{imp}} \cdot \hat{W}$ stabilizes the infinite-dimensional, open-loop unstable, spatially distributed plant $\hat{\mathbf{P}}^\infty$.

REFERENCES

[C1] M. W. Cantoni, *Gap–metric performance bounds for linear feedback systems*, in Proceedings of the 38th IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, 1999, pp. 4505–4510.

[CG1] M. W. Cantoni and K. Glover, *Existence of right and left representations of the graph for linear periodically time-varying systems*, SIAM J. Control Optim., 38 (2000), pp. 786–802.

[CG2] M. W. Cantoni and K. Glover, *Gap-metric robustness analysis of linear periodically time-varying feedback systems*, SIAM J. Control Optim., 38 (2000), pp. 803–822.

[C2] R. F. Curtain, *Robust stabilization of normalized coprime factors: The infinite–dimensional case*, Internat. J. Control, 51 (1990), pp. 1173–1190.

[CZ] R. F. Curtain and H. J. Zwart, *An Introduction to Infinite–Dimensional Linear Systems Theory*, Texts Appl. Math., 21 Springer-Verlag, Berlin, 1995.

[DS] W. Dale and M. C. Smith, *Stabilizability and existence of system representations for discrete-time, time-varying systems*, SIAM J. Control Optim., 31 (1993), pp. 1538–1557.

[DLMS] C. A. Desoer, R. W. Liu, J. Murray, and R. Saeks, *Feedback system design: The fractional representation approach*, IEEE Trans. Automat. Control, 25 (1980), pp. 399–412.

[DGS] J. C. Doyle, T. T. Georgiou, and M. C. Smith, *The parallel projection operators of a nonlinear feedback system*, Systems Control Lett., 20 (1993), pp. 79–85.

[E] A. K. El-Sakkary, *The gap metric: Robustness of stabilization of feedback systems*, IEEE Trans. Automat. Control, 30 (1985), pp. 240–247.

[FF] C. Foias and A. Frazho, *The Commutant Lifting Approach to Interpolation Problems*, Oper. Theory Adv. Appl. 44, Birkhäuser, Berlin, 1990.

[FGS1] C. Foias, T. T. Georgiou, and M. C. Smith, *Geometric techniques for robust stabilization of linear time-varying systems*, in Proceedings of the 29th IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, 1990, pp. 2868–2873.

[FGS] C. Foias, T. T. Georgiou, and M. C. Smith, *Robust Stability of feedback systems: A geometric approach using the gap metric*, SIAM J. Control Optim., 31 (1993), pp. 1518–1537.

[F] B. A. Francis, *A Course in H^∞ Control Theory*, Lecture Notes in Control and Inform. Sci. 88, Springer-Verlag, Berlin, 1987.

[G1] T. T. Georgiou, *On the computation of the gap metric*, Systems and Control Lett., 11 (1988), pp. 253–257.

[G2] T. T. Georgiou, *Differential stability and robust control of nonlinear systems*, Math. Control Signals Systems, 6 (1993), pp. 289–306.

[GS1] T. T. Georgiou and M. C. Smith, *Optimal robustness in the gap metric*, IEEE Trans. Automat. Control, 35 (1990), pp. 673–686.

[GS2] T. T. Georgiou and M. C. Smith, *Robust stabilization in the gap metric: Controller design for distributed plants*, IEEE Trans. Automat. Control, 37 (1992), pp. 1133–1143.

[GS3] T.T. Georgiou and M. C. Smith, *Graphs, causality and stabilizability: Linear, shift–invariant systems*, Math. Control Signals Systems, 6 (1993), pp. 195–223.

[GS5] T. T. Georgiou and M. C. Smith, *Robustness analysis of nonlinear feedback systems: An input–output approach*, IEEE Trans. Automat. Control, 42 (1997), pp. 1200–1221.

[GM] K. Glover and D. McFarlane, *Robust stabilization of normalized coprime factor plant descriptions with H^∞–bounded uncertainty*, IEEE Trans. Automat. Control, 34 (1989), pp. 821–830.

[H] P. R. Halmos, *A Hilbert Space Problem Book*, 2nd ed., Springer-Verlag, New York, Heidelberg, Berlin, 1982.

[K1] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1980.

[KVZRS] M. A. Krasnosel'skiĭ, G. M. Vainikko, P. P. Zabreiko, Ya. B. Rutitskiĭ, and V. Ya. Stetsenko, *Approximate Solution of Operator Equations*, Wolters–Noordhoff, Groningen, The Netherlands, 1972.

[K2] E. Kreyszig, *Introductory Functional Analysis with Applications*, John Wiley and Sons, New York, 1978.

[MG1] D. C. McFarlane and K. Glover, *Robust Controller Design Using Normalized Coprime Factor Plant Descriptions*, Lecture Notes in Control and Inform. Sci. 138, Springer-Verlag, New York, 1990.

[MG2] D. C. McFarlane and K. Glover, *A loop shaping design procedure using H^∞ synthesis*, IEEE Trans. Automat. Control, 37 (1992), pp. 759–769.

[N] N. K. Nikol'skiĭ, *Treatise on the Shift Operator*, Springer-Verlag, Berlin, 1986.

[OS] R. J. Ober and J. A. Sefton, *Stability of control systems and graphs of linear systems*, Systems and Control Lett., 17 (1991), pp. 265–280.

[QD] L. Qiu and E. J. Davison, *Feedback stability under simultaneous gap metric uncertainties in plant and controller*, Systems Control Lett., 18 (1992), pp. 9–22.

[R] J. Reinschke, *H^∞–Control of Spatially Distributed Systems*, Ph.D. thesis, Department of Engineering, University of Cambridge, Cambridge, UK, 1999.

[RR] M. Rosenblum and J. Rovnyak, *Hardy Classes and Operator Theory*, Oxford University Press, Oxford, UK, 1985.

[SO] J. A. Sefton and R. J. Ober, *On the gap metric and coprime factor perturbations*, Automatica J. IFAC, 29 (1993), pp. 723–734.

[V1] M. Vidyasagar, *The graph metric for unstable plants and robustness estimates for feedback stability*, IEEE Trans. Automat. Control, 29 (1984), pp. 203–209.

[V2] M. Vidyasagar, *Control System Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA, 1985.

[VK] M. Vidyasagar and H. Kimura, *Robust controllers for uncertain linear multivariable systems*, Automatica J. IFAC, 22 (1986), pp. 240–247.

[VSF] M. Vidyasagar, H. Schneider, and B. A. Francis, *Algebraic and topological aspects of feedback stabilization*, IEEE Trans. Automat. Control, 27 (1982), pp. 880–894.

[V3] G. Vinnicombe, *Frequency domain uncertainty and the graph topology*, IEEE Trans. Automat. Control, 38 (1993), pp. 1371–1383.

[V4] G. Vinnicombe, *Uncertainty and Feedback*, Imperial College Press, London, 2001.

[ZE] G. Zames and A. K. El-Sakkary, *Unstable systems and feedback: The gap metric*, in Proceedings of the Allerton Conference, University of Illinois, Monticello, IL, 1980, pp. 380–385.

[ZDG] K. Zhou, J. Doyle, and K. Glover, *Robust and Optimal Control*, Prentice Hall, Upper Saddle River, NJ, 1996.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 2, pp. 628–660

MULTIOBJECTIVE $\mathcal{H}_2/\mathcal{H}_\infty$ CONTROL DESIGN*

XIANG CHEN† AND KEMIN ZHOU‡

Abstract. An important question in feedback control design is how to achieve desired performance for dynamical systems subject to both model uncertainties and white noise. For this purpose, it is desirable to develop a systematic design technique that combines the good aspects of both $\mathcal{H}_2(LQG)$ and $\mathcal{H}_\infty$ methods, which provides the motivation for the development of multiobjective design framework in this paper. Encouraged by a time domain game approach for $\mathcal{H}_2/\mathcal{H}_\infty$ control design, two multiobjective control design problems are formulated and solved in the time domain. Results for both the finite time horizon and the infinite time horizon are presented. It is shown that all the results can be obtained by solving the corresponding set of coupled Riccati equations.

Key words. multiobjective control, $\mathcal{H}_2/\mathcal{H}_\infty$ control, $\mathcal{H}_\infty$ Gaussian control

AMS subject classifications. 49K15, 93C05, 93C15, 93C35

PII. S0363012998346372

1. Introduction.

It is probably fair to say that the most important objective of any control design is to achieve certain desired performance specifications in spite of external disturbances and noises, system parameter variations, and variations of system operating conditions. What a judicially designed feedback can usually achieve is to improve the system performance in one aspect by sacrificing the system performance in another aspect. Thus a feedback control design is a process of making trade-offs between conflicting objectives. Two prominent conflicting objectives in most feedback control designs are good transient response and robustness with respect to disturbances and system uncertainties. Usually a very robust control law tends to make the system's transient response poor [35, 36]. On the other hand, a system with an extremely good transient response for a nominal operation condition (or model) could be very sensitive to external disturbances and parameter variations [9]. In this case a good control design should be a compromise between good transient performance and robustness. It is generally agreed that a suitable linear quadratic Gaussian (LQG) or $\mathcal{H}_2$ criterion can be a good measure for transient performance, while the $\mathcal{H}_\infty$ optimal control design framework is developed primarily because of the robustness consideration. Thus it is natural to consider a design framework that can systematically make the design tradeoffs between these two design objectives. The development of such a multiobjective design framework is the main topic of this paper.

The multiobjective control problem has received much attention from the control research community in the past decade [3, 4, 6, 7, 8, 11, 12, 14, 17, 20, 22, 25, 26, 28, 27, 29, 30, 31, 34, 36]. Though there are many multiobjective approaches [32], the $\mathcal{H}_2/\mathcal{H}_\infty$ approach has a better physical interpretation and clearer motivation, as discussed above, and attracts a great number of researchers. It should be pointed out that the term $\mathcal{H}_2/\mathcal{H}_\infty$ usually refers to any multiobjective optimal design of which the

*Received by the editors October 19, 1998; accepted for publication (in revised form) April 5, 2001; published electronically August 29, 2001. This research was supported in part by NSERC (Research grant 217351), AFOSR (F49620-94-1-0415), ARO (DAAH04-96-1-0193), LEQSF (DOD/LEQSF(1996–99)-04, LEQSF(1995–98)-RD-A-14).

http://www.siam.org/journals/sicon/40-2/34637.html

†Department of Electrical and Computer Engineering, University of Windsor, 401 Sunset Avenue, Windsor, Ontario, Canada N9B 3P4 (xchen@uwindsor.ca).

‡Department of Electrical and Computer Engineering, Louisiana State University, Baton Rouge, LA 70803-5901 (kemin@ee.lsu.edu).

performance measures have both $\mathcal{H}_2(LQG)$ and $\mathcal{H}_\infty$ interpretations. Thus, under the big framework of $\mathcal{H}_2/\mathcal{H}_\infty$, there are many different ways to formulate the problem. It is more appropriate to call them multiobjective $\mathcal{H}_2/\mathcal{H}_\infty$ designs as is done in this paper.

Some major results about multiobjective control with an $\mathcal{H}_2/\mathcal{H}_\infty$ interpretation are summarized as follows (see also [32]):

1. Fixed-order controller design by minimizing an auxiliary cost functional [3, 17, 18]: This formulation minimizes an upper bound on the $\mathcal{H}_2$ norm of the closed-loop transfer function as the index functional which is subject to an $\mathcal{H}_\infty$ norm constraint and designs a fixed order controller for the formulated problem.

2. Convex optimization [20, 14, 19, 12, 8, 4]: There are several approaches in this category. In general, each approach uses (different) matrix inequalities as performance measures to characterize a convex optimization problem. The advantage of the convex optimization approach is that there exist effective and powerful algorithms for the solutions of these problems. However, it is difficult to generalize this approach to a nonlinear system or to apply this approach to systems with stochastic disturbances.

3. Optimizing an entropy cost functional [24, 25, 15]: This approach designs a controller to minimize the so-called closed-loop entropy which provides an upper bound of $\mathcal{H}_2$ cost, while guaranteeing the $\mathcal{H}_\infty$ performance. It turns out that this approach is equivalent to the approach of minimizing an auxiliary cost functional in [3] for the single external input case [24].

4. Bounded power characterization [11, 36]: This approach can treat systems with both white noise and bounded power disturbances. The design objective is to minimize the power of the output error signal. The problem solved by this approach is a dual to that solved in [3] in some sense [34]. This approach is interesting because it uses norms of power signals instead of norms of transfer functions to characterize the problem, which positions it as a time domain approach.

5. Nash game approach [21, 22, 28]: This approach uses the Nash equilibrium strategy [2] as a performance measure to characterize the problem with a very clear $\mathcal{H}_2/\mathcal{H}_\infty$ interpretation. It is also possible to generalize this approach to a nonlinear system [23]. However, only a state feedback problem was solved in [22], and the output feedback problem turns out to be very difficult [28].

As pointed out before, the clear $\mathcal{H}_2/\mathcal{H}_\infty$ interpretation plus the solvability through some standard algorithm makes the Nash game approach very attractive. Hence how to generalize this approach to the output feedback case is very interesting and important, considering that, in general, information of system state variables may not be available. The desire of developing output feedback mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control using the game approach triggered our development of multiobjective control design in this paper. We shall adopt the observer-based controller structure for the proposed multiobjective control problems. As a result of this, we have to take the estimation of system states into account in the output feedback framework; hence the inclusion of the noise (usually Gaussian white noise) effect becomes necessary for the purpose of optimization as it does in classical LQG control. Because of this similarity, we call the second multiobjective control problem formulated in this paper "$\mathcal{H}_\infty$ Gaussian Control."

The paper is organized as follows: in section 2 the formulation of our problems is given; section 3 provides definitions and preliminary results about signals, systems, and constrained optimization required for sections 4 and 5, which are devoted to solving the main problems in both finite and infinite time horizons; some comments

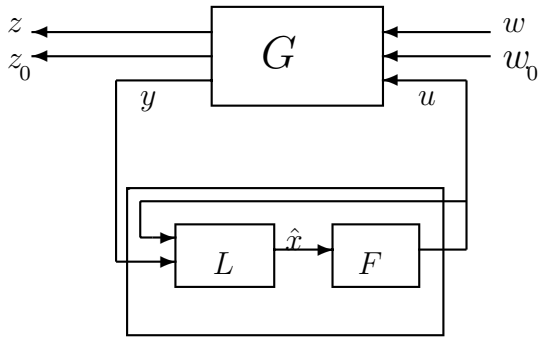


FIG. 1. *Multiobjective control design.*

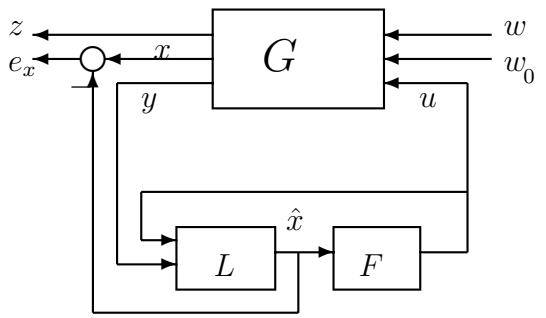


FIG. 2. H_∞ *Gaussian control design.*

and the conclusion can be found in sections 6 and 7.

The notations used in this paper are standard: I always denotes the identity matrix with dimensions determined in context; $E\{\cdot\}$ is the expectation operator; if A is a matrix or a vector, then A^T is its transpose and A^* is its conjugate transpose; if x is a complex number, then $Re(x)$ and $Im(x)$ are, respectively, its real and imaginary parts; $\mathbf{R}$ is the set of real numbers; $\mathcal{RH}_\infty$ is the space of all proper and real rational stable matrix transfer functions; if A is a square matrix, then trace (A) is the trace of A; $\|\cdot\|$ represents the Euclidean norm of a vector.

2. Formulation of main problems. As stated in the introduction, in this paper, we are interested in control design problems as shown in Figures 1 and 2.

From now on, we call the problem in Figure 1 *multiobjective control design* and the problem in Figure 2 $\mathcal{H}_\infty$ *Gaussian control design.* We shall first provide the definition of signal and system norms. Then we present the formulation of our problems.

2.1. Norms of signals and systems. There are two classes of stochastic signals which are used in this paper for the development of multiobjective control design: bounded power signals and white noise signals. It should be pointed out that a deterministic version of bounded power signals can also be defined (see [36, 13]), and the results obtained in this paper can be derived correspondingly.

Given a real vector stochastic signal $u(t)$,

$$u(t) = \left[\begin{array}{cccc} u_1(t) & u_2(t) & \cdots & u_m(t) \end{array}\right]^T \in \mathbf{R}^m \quad \forall t \in R,$$

where $u_i(t)$, $i = 1, \ldots, m$ are real stationary random processes, define the *mean* and

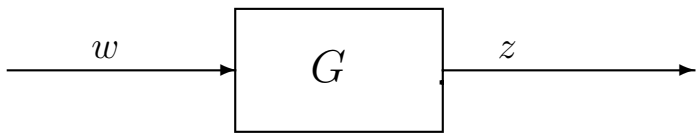


FIG. 3. *A stable system.*

autocorrelation matrices of $u(t)$, respectively, as follows:

$$E\{u\} := \left[\begin{array}{cccc} E\{u_1(t)\} & E\{u_2(t)\} & \cdots & E\{u_m(t)\} \end{array}\right]^T,$$

$$R_{uu}(\tau) := \lim_{T\to\infty} \frac{1}{T} \int_0^T E\{u(t+\tau)u^T(t)\}dt.$$

The Fourier transform of $R_{uu}(\tau)$, if exists, is $S_{uu} := \frac{1}{2\pi}\int_{-\infty}^{\infty} R_{uu}(\tau)e^{-j\omega\tau}d\tau$. The so-called bounded power signal is defined as follows.

DEFINITION 1. *A vector stationary stochastic signal u is said to have bounded power if*

1. *both R_{uu} and S_{uu} exist;*
2. $\lim_{T\to\infty} \frac{1}{T}\int_0^T E\|u(t)\|^2 dt < \infty$.

Let $\mathcal{P}$ be the space of all signals with bounded power. A seminorm can be defined on $\mathcal{P}$:

$$\|u\|_{\mathcal{P}} := \sqrt{\lim_{T\to\infty} \frac{1}{T} \int_0^T E\{\|u(t)\|^2\}dt} = \sqrt{\text{trace}[R_{uu}(0)]} \quad \forall u \in \mathcal{P}.$$

The well-known Gaussian white noise $w_0(t)$ is a stationary random process that satisfies $E\{w_0(t)\} \equiv 0$ and $E\{w_0(t)w_0^T(\tau)\} = Q(t)\delta(t-\tau)$, where $\delta(t)$ is Dirac δ function and $Q(t)$ is a positive definite matrix. In this paper, we shall assume, without loss of generality, that $Q(t) = I$, where I is an identity matrix, i.e., $w_0(t)$ is a zero mean stationary process with an identity power spectrum. A more rigorous description of the white noise process can be found in standard textbooks for stochastic processes (see, e.g., [33]). Independence between stochastic signals is defined as follows.

DEFINITION 2. *Two vector stationary stochastic signals $w_1(t)$ and $w_2(t)$ are said to be independent if, for any $t_1 \geq 0$ and $t_2 \geq 0$, we have*

$$E\{w_1(t_1)w_2^T(t_2)\} = E\{w_1(t_1)\}E\{w_2^T(t_2)\}.$$

A close relation exists between these two types of signals and the $\mathcal{H}_\infty$ and $\mathcal{H}_2$ norms of systems. Recall that, given a system $G(s) \in \mathcal{RH}_\infty$ in Figure 3 with a state space realization (A, B, C, D) and denoted by $G(s) = D + C(sI - A)^{-1}B := \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right]$, the $\mathcal{H}_\infty$ and $\mathcal{H}_2$ norms of this system are defined as

$$\|G(s)\|_\infty := \sup_\omega \bar{\sigma}\{G(j\omega)\}, \quad \|G(s)\|_2 := \sqrt{\frac{1}{2\pi}\int_{-\infty}^{\infty} \text{trace}[G^*(j\omega)G(j\omega)]d\omega},$$

where $\bar{\sigma}$ is the largest singular value of $G(s)$.

Let w be a bounded power signal. It can be proved that $\|G(s)\|_\infty = \sup_w \frac{\|z\|_{\mathcal{P}}}{\|w\|_{\mathcal{P}}}$ (see [36]). On the other hand, if $w = w_0(t)$ is a white noise signal, then it can be

proved that $\|G(s)\|_2 = \|z\|_{\mathcal{P}}$ (see [36]). On finite time horizon $[0,T]$, we define the 2-norm of a signal u as $\|u\|_{[0,T]} := \sqrt{\int_0^T E\|u\|^2 dt}$. Accordingly, we define the system norms

$$\|G(s)\|_{\infty,[0,T]} := \sup_w \frac{\|z\|_{[0,T]}}{\|w\|_{[0,T]}}, \quad \text{where } w \text{ is any bounded power signal,}$$

and

$$\|G(s)\|_{2,[0,T]} := \|z\|_{[0,T]}, \quad \text{where } w \text{ is a white noise signal.}$$

These definitions will be helpful for formulating our problems on the finite time horizon. Finally, note that it is easy to verify the following equivalencies:

$$\|G(s)\|_\infty < \gamma \iff 0 < \gamma^2\|w\|_{\mathcal{P}}^2 - \|z\|_{\mathcal{P}}^2 \quad \forall w \neq 0,$$

$$\|G(s)\|_{\infty,[0,T]} < \gamma \iff 0 < \gamma^2\|w\|_{[0,T]}^2 - \|z\|_{[0,T]}^2 \quad \forall w \neq 0.$$

2.2. Formulation of multiobjective control design. Consider a linear control system G in Figure 1 described by

$$\dot{x} = Ax + B_0 w_0 + B_1 w + B_2 u, \quad x(0) = 0, \tag{1}$$

$$y = C_2 x + D_{20} w_0, \quad R_0 := D_{20} D_{20}^T > 0, \tag{2}$$

$$z = C_1 x + D_{12} u, \quad R_1 = D_{12}^T D_{12}, \tag{3}$$

$$z_0 = C_0 x + D_{02} u, \quad R_{02} := D_{02}^T D_{02} > 0, \tag{4}$$

where w is a bounded power signal and w_0 is a white noise signal. Define the performance index functionals

$$J_1(u,w,w_0) := \gamma^2\|w\|_{[0,T]}^2 - \|z\|_{[0,T]}^2, \quad J_2(u,w,w_0) := \|z_0\|_{[0,T]}^2,$$

and

$$J_3(u,w,w_0) := \gamma^2\|w\|_{\mathcal{P}}^2 - \|z\|_{\mathcal{P}}^2, \quad J_4(u,w,w_0) := \|z_0\|_{\mathcal{P}}^2,$$

where u is the output feedback control law to be designed.

Then the *multiobjective control design* is formulated as follows: *Find an output feedback control law u_* such that it achieves*

$$J_1(u_*,w_*,w_0) \le J_1(u_*,w,w_0) \ \forall w \neq w_*, \ \ J_2(u_*,w_*,w_0) \le J_2(u,w_*,w_0)$$

for the finite time horizon case and

$$J_3(u_*,w_*,w_0) \le J_3(u_*,w,w_0) \ \ \forall w \neq w_*, \ \ J_4(u_*,w_*,w_0) \le J_4(u,w_*,w_0)$$

for the infinite time horizon case, where w_* is the worst possible disturbance signal to be determined.

It must be pointed out that the control law u (hence the optimal control law u_*) is supposed to have the following observer form:

$$\dot{\hat{x}} = \hat{A}\hat{x} + B_2 u - Ly, \hat{x}(0) = 0,$$
$$u = F\hat{x}.$$

2.3. Formulation of $\mathcal{H}_\infty$ Gaussian control. Consider a linear system in Figure 2 described by

$$\dot{x} = Ax + B_0 w_0 + B_1 w + B_2 u, \quad x(0) = 0, \tag{5}$$

$$z = C_1 x + D_{12} u, \quad R_1 := D_{12}^T D_{12} > 0, \tag{6}$$

$$y = C_2 x + D_{20} w_0, \quad R_0 := D_{20} D_{20}^T > 0, \tag{7}$$

where w is a bounded power signal and w_0 is a white noise signal. w and w_0 are independent. Let $\mathcal{P}_s$ be the subset of $\mathcal{P}$ consisting of all bounded power signals independent from w_0; hence $w \in \mathcal{P}_s$. The control law is supposed to take the observer form

$$\dot{\hat{x}} = \hat{A}\hat{x} + B_2 u - Ly, \hat{x}(0) = 0,$$
$$u = F\hat{x}.$$

Let $e_x := x - \hat{x}$, and define the following performance index functionals:

$$J_1(u, w, w_0) := \gamma^2 \|w\|_{[0,T]}^2 - \|z\|_{[0,T]}^2, \quad J_2(u, w, w_0) := \|e_x\|_{[0,T]}^2,$$

and

$$J_3(u, w, w_0) := \gamma^2 \|w\|_{\mathcal{P}}^2 - \|z\|_{\mathcal{P}}^2, \quad J_4(u, w, w_0) := \|e_x\|_{\mathcal{P}}^2.$$

Then the $\mathcal{H}_\infty$ *Gaussian control design* is formulated as follows: *Find a feedback control law u_* in the given form such that it achieves*

$$J_1(u_*, w_*, w_0) \le J_1(u_*, w, w_0) \quad \forall w \in \mathcal{P}_s, \quad J_2(u_*, w_*, w_0) \le J_2(u, w_*, w_0)$$

for the finite time horizon case and

$$J_3(u_*, w_*, w_0) \le J_3(u_*, w, w_0) \quad \forall w \in \mathcal{P}_s, \quad J_4(u_*, w_*, w_0) \le J_4(u, w_*, w_0)$$

for the infinite time horizon case, where w_* is called the worst disturbance signal.

Note that w_* may not be necessarily in $\mathcal{P}_s$, but it does guarantee that the performance index J_1 or J_3 has a lower bound.

3. Preliminaries. In this section, we present some preliminary results which will be applied to proving the main results in the subsequent sections.

LEMMA 3. *Let $R_{zw}(s)$ be the stable matrix transfer function from w (a bounded power signal) to z defined by*

$$\dot{x} = Ax + Bw, \quad x(0) = 0,$$
$$z = Cx.$$

Then the following two statements are equivalent:

1. $\|R_{zw}(s)\|_{\infty,[0,T]} < \gamma$, *i.e.*, $0 < \gamma^2 \|w\|_{[0,T]}^2 - \|z\|_{[0,T]}^2 \ \ \forall w \neq 0$;
2. *the solution of the Riccati equation*

$$-\dot{P}(t) = A^T P(t) + P(t)A - \gamma^{-2} P(t) B B^T P(t) - C^T C, \quad P(T) = 0,$$

has no finite escape time on $[0, T]$.

A proof of this lemma can be found in [22, Lemma 2].

LEMMA 4. *Consider the system*

$$\dot{x} = Ax + B_0 w_0 + B_1 w + B_2 u, \quad x(0) = 0,$$
$$y = C_2 x + D_{20} w_0,$$

where w_0 is a white noise and w is a stationary signal. Suppose w_0 and w are independent. Let the controller $K(s) = \bar{C}(sI - \bar{A})^{-1}\bar{B}$. Then we have

$$E\{x(t)w_0^T(t_1)\} = (e_{11}B_0 + e_{12}\bar{B}D_{20})/2, \ \ t_1 \le t, \quad or \quad E\{x(t)w_0^T(t_1)\} = 0, \quad t_1 > t,$$

where $\hat{A} = \begin{bmatrix} A & B_2\bar{C} \\ \bar{B}C_2 & \bar{A} \end{bmatrix}$ *and* $e^{\hat{A}(t-t_1)} = \begin{bmatrix} e_{11} & e_{12} \\ e_{21} & e_{22} \end{bmatrix}$.

Proof. Since $K(s) = \bar{C}(sI - \bar{A})^{-1}\bar{B}$, the closed-loop system becomes

$$\dot{\hat{x}} = \hat{A}\hat{x} + \hat{B}_0 w_0 + \hat{B}_1 w,$$

where $\hat{x} = [\ x^T \quad \bar{x}^T\ ]^T$, $\hat{B}_0 = [\ B_0^T \quad (\bar{B}D_{20})^T\ ]^T$, and $\hat{B}_1 = [\ B_1^T \quad 0\ ]^T$. Hence

$$\hat{x} = \int_0^t e^{\hat{A}(t-\tau)}[\hat{B}_0 w_0(\tau) + \hat{B}_1 w(\tau)]d\tau,$$

and

$$E[\ w_0(t_1)x^T(t) \quad w_0(t_1)\bar{x}^T(t)\ ]^T = E\left\{\int_0^t e^{\hat{A}(t-\tau)}[\hat{B}_0 w_0(\tau) + \hat{B}_1 w(\tau)]w_0^T(t_1)d\tau\right\}$$

$$= \int_0^t e^{\hat{A}(t-\tau)}\hat{B}_0 E\{w_0(\tau)w_0^T(t_1)\}d\tau = \int_0^t e^{\hat{A}(t-\tau)}\hat{B}_0\delta(\tau - t_1)d\tau,$$

which gives $E\{x(t)w_0^T(t_1)\} = (e_{11}B_0 + e_{12}\bar{B}D_{20})/2$ for $t_1 \le t$, or $E\{x(t)w_0^T(t_1)\} = 0$ for $t_1 > t$. ∎

LEMMA 5. *Consider the system*

$$\dot{x} = Ax + B_0 w_0 + B_1 w + B_2 u, \quad x(0) = 0,$$
$$y = C_2 x + D_{20} w_0,$$

where w_0 is a white noise and the controller $K(s)$ is of the form

$$\dot{\bar{x}} = \bar{A}\bar{x} + \bar{B}y, \quad \bar{x}(0) = 0,$$
$$u = \bar{C}\bar{x}.$$

Let $w = F\hat{x}$, where $\hat{x} = [\ x^T \quad \bar{x}^T\ ]^T$. Then we have

$$E\{x(t)w_0^T(t_1)\} = (e_{11}B_0 + e_{12}\bar{B}D_{20})/2, \ \ t_1 \le t, \quad or \quad E\{x(t)w_0^T(t_1)\} = 0, \quad t_1 > t,$$

where $\hat{A} = \begin{bmatrix} A & B_2\bar{C} \\ \bar{B}C_2 & \bar{A} \end{bmatrix} + B_1 F$ *and* $e^{\hat{A}(t-t_1)} = \begin{bmatrix} e_{11} & e_{12} \\ e_{21} & e_{22} \end{bmatrix}$.

Proof. The proof is similar to that of Lemma 4 and is omitted. ∎

A constrained optimization problem solved below plays the key role in the solution of our multiobjective control design. The sufficient conditions for the constrained optimization problem will be proved in this section, and the proof of necessary conditions will be given in the appendix of this paper.

Given $A \in \mathbf{R}^{n\times n}$, $B \in \mathbf{R}^{n\times r}$, $C \in \mathbf{R}^{p\times n}$, $D \in \mathbf{R}^{p\times r}$, and $R = DD^T > 0$, we define two index functionals:

1. $J_1(L(t)) = \text{trace}(\int_0^T Q(t)P(t)Q^T(t)dt)$, $T \geq 0$, where $Q(t)$ is any time-varying weighting matrix, and $P(t) = P^T(t) \geq 0$ on $[0,T]$ with $P(0) = 0$ satisfies

$$(A + L(t)C)P(t) + P(t)(A + L(t)C)^T + (B + L(t)D)(B + L(t)D)^T = \dot{P}(t). \tag{8}$$

2. $J_2(L) = \text{trace}(QPQ^T)$, where Q is any constant weighting matrix, $A + LC$ is Hurwitz, and $P = P^T \geq 0$ satisfies

$$(A + LC)P + P(A + LC)^T + (B + LD)(B + LD)^T = 0. \tag{9}$$

The constrained optimization problems on finite time and infinite time horizons can be stated as follows:

1. $\min_{L(t)} J_1(L(t)) = \min_{P(t)} \text{trace}(\int_0^T Q(t)P(t)Q^T(t)dt)$, where $L(t)$ and $P(t)$ are subject to the constraint (8).
2. $\min_L J_2(L) = \min_P \text{trace}(QPQ^T)$, where L and P are subject to the constraint (9).

The next two theorems provide solutions to these two optimization problems.

THEOREM 6. *Consider the constrained optimization problem defined on the finite time horizon; if there is a solution $P_*(t) \geq 0$ on $[0,T]$ with $P_*(0) = 0$ for*

$$(A - BD^TR^{-1}C)P_*(t) + P_*(t)(A - BD^TR^{-1}C)^T - P_*(t)C^TR^{-1}CP_*(t) + B(I - D^TR^{-1}D)B^T = \dot{P}_*(t),$$

then $J_1(L(t))$ achieves the minimum value at $L_*(t) = -(P_*(t)C^T + BD^T)R^{-1}$.

Conversely, if there are $L(t)$ and $P(t)$ such that

$$(A + L(t)C)P(t) + P(t)(A + L(t)C)^T + (B + L(t)D)(B + L(t)D)^T = \dot{P}(t)$$

and $J_1(L(t))$ is the minimum value, then there is a solution $P_*(t) \geq 0$ on $[0,T]$ with $P_*(0) = 0$ to

$$(A - BD^TR^{-1}C)P_*(t) + P_*(t)(A - BD^TR^{-1}C)^T - P_*(t)C^TR^{-1}CP_*(t) + B(I - D^TR^{-1}D)B^T = \dot{P}_*(t),$$

and the minimum value of $J(L(t))$ is also achieved at $L_*(t) = -(P_*(t)C^T + BD^T)R^{-1}$.

Proof (sufficiency). Take $\Delta P(t) = P(t) - P_*(t)$. Then

$$\Delta\dot{P}(t) = (A + L(t)C)\Delta P(t) + \Delta P(t)(A + L(t)C)^T + (L(t) - L_*(t))R(L(t) - L_*(t))^T,$$

where $L_*(t) = -(P_*(t)C^T + BD^T)R^{-1}$. Now let $\Phi(t,\tau)$ be the transition matrix of $A + L(t)C$. Then

$$\Delta P(t) = \int_0^t \Phi(t,s)(L(s) - L_*(s))R(L(s) - L_*(s))^T\Phi^T(t,s)ds \geq 0$$

for any $L(t)$ and $\Delta P(t) = 0$ if $L(t) = L_*(t)$. Therefore, $J(L(t)) - J(L_*(t)) \geq 0$ for any $L(t)$, which means that $J(L(t))$ achieves the minimum value at $L_*(t)$.

For the proof of necessity, please see the appendix. □

THEOREM 7. *Consider the constrained optimization problem, defined for infinite time horizon case; if there is a stabilizing solution $P_* \geq 0$ for*

$$(A-BD^TR^{-1}C)P_*+P_*(A-BD^TR^{-1}C)^T-P_*C^TR^{-1}CP_*+B(I-D^TR^{-1}D)B^T=0$$

i.e., $A-BD^TR^{-1}C-P_*C^TR^{-1}C$ is stable, then $J_2(L)$ achieves the minimum value at $L_*=-(P_*C^T+BD^T)R^{-1}$.

Conversely, let (C,A) be detectable. If there are L_1 and $P_1 \geq 0$, where $A+L_1C$ is stable and P_1 solves

$$(A+L_1C)P_1+P_1(A+L_1C)^T+(B+L_1D)(B+L_1D)^T=0$$

such that $J_2(L)$ is minimized, then there is a $P_* \geq 0$ solving

$$(A-BD^TR^{-1}C)P_*+P_*(A-BD^TR^{-1}C)^T-P_*C^TR^{-1}CP_*+B(I-D^TR^{-1}D)B^T=0.$$

Moreover, an optimal L_* can be obtained as $L_*=-(P_*C^T+BD^T)R^{-1}$ if $A+L_*C$ is Hurwitz.

Proof (sufficiency). Since P_* is a stabilizing solution, $A+L_*C$ is Hurwitz, where $L_*=-(P_*C^T+BD^T)R^{-1}$. For any L for which $A+LC$ is stable, we have $P \geq 0$ solving

$$(A+LC)P+P(A+LC)^T+(B+LD)(B+LD)^T=0.$$

Now take $\Delta P=P-P_*$. Then

$$(A+LC)\Delta P+\Delta P(A+LC)^T+(L-L_*)R(L-L_*)^T=0.$$

By the standard property of the Lyapunov equation, we have $\Delta P \geq 0$ and $\Delta P=0$ if and only if $L=L_*$. Hence $J(L)-J(L_*) \geq 0$ for any L, which means that $J(L)$ achieves the minimum value at L_*.

For the proof of necessity, please see the appendix. □

4. Multiobjective control design. In this section, we give the main theorems for the multiobjective control design problems formulated in section 2.2. To simplify the notations, the following abbreviations are assumed:

$$A_s=A-B_2R_{02}^{-1}D_{02}^TC_0, \quad A_f=A-B_0D_{20}^TR_0^{-1}C_2,$$

$$P:=B_0(I-D_{20}^TR_0^{-1}D_{20})B_0^T, \quad Q:=C_0^T(I-D_{02}R_{02}^{-1}D_{02}^T)C_0.$$

4.1. Multiobjective control design—finite time horizon. To motivate our output feedback design, we shall first present the state feedback design results, a simplified version of which can be found in [22].

THEOREM 8. *For the system G described by* (1)–(4) *and the associated cost functionals $J_1(u,w,w_0)$ and $J_2(u,w,w_0)$, there exist linear memoryless state feedback strategies (Nash equilibrium strategies) u_* and w_* such that*

$$0 \leq J_1(u_*,w_*,0) \leq J_1(u_*,w,0), \quad J_2(u_*,w_*,0) \leq J_2(u,w_*,0)$$

if and only if the coupled Riccati differential equations

$$-\dot{P}_1(t)=(A_s-B_2R_{02}^{-1}B_2^TP_2(t))^TP_1(t)+P_1(t)(A_s-B_2R_{02}^{-1}B_2^TP_2(t))+\gamma^{-2}P_1(t)B_1B_1^TP_1(t)$$

$$+[C_1 - D_{12}R_{02}^{-1}(D_{02}^T C_0 + B_2^T P_2(t))]^T[C_1 - D_{12}R_{02}^{-1}(D_{02}^T C_0 + B_2^T P_2(t))],$$

$$-\dot{P}_2(t) = (A_s + \gamma^{-2}B_1B_1^T P_1(t))^T P_2(t) + P_2(t)(A_s + \gamma^{-2}B_1B_1^T P_1(t))$$

$$-P_2(t)B_2R_{02}^{-1}B_2^T P_2(t) + Q$$

have solutions $P_1(t) \geq 0$ *and* $P_2(t) \geq 0$ *on* $[0,T]$ *with* $P_1(T) = 0$ *and* $P_2(T) = 0$. *Furthermore, if the solutions exist, we have* $w_* = \gamma^{-2}B_1^T P_1(t)x$ *and* $u_* = F_*(t)x$ *with* $F_* = -R_{02}^{-1}(D_{02}^T C_0 + B_2^T P_2(t))$.

The output feedback design are established in the following theorems for multiobjective control in finite time horizon.

THEOREM 9. *There exist a* w_* *and an output feedback control law* u_* *in the form of*

$$\begin{aligned}\dot{\hat{x}} &= (A + \gamma^{-2}B_1B_1^T P_1(t) + B_2F_*(t))\hat{x} + L_*(t)(C_2\hat{x} - y),\\ u_* &= F_*(t)\hat{x},\end{aligned}$$

such that $J_1(u_*, w_*, w_0) \leq J_1(u_*, w, w_0), J_2(u_*, w_*, w_0) \leq J_2(u, w_*, w_0)$ *if the coupled differential Riccati equations*

$$-\dot{P}_1(t) = (A_s - B_2R_{02}^{-1}B_2^T P_2(t))^T P_1(t) + P_1(t)(A_s - B_2R_{02}^{-1}B_2^T P_2(t)) + \gamma^{-2}P_1(t)B_1B_1^T P_1(t)$$

$$+[C_1 - D_{12}R_{02}^{-1}(D_{02}^T C_0 + B_2^T P_2(t))]^T[C_1 - D_{12}R_{02}^{-1}(D_{02}^T C_0 + B_2^T P_2(t))],$$

$$-\dot{P}_2(t) = (A_s + \gamma^{-2}B_1B_1^T P_1)^T P_2(t) + P_2(t)(A_s + \gamma^{-2}B_1B_1^T P_1(t)) - P_2(t)B_2R_{02}^{-1}B_2^T P_2(t) + Q,$$

$$\dot{P}_3(t) = (A_f + \gamma^{-2}B_1B_1^T P_1(t))P_3(t) + P_3(t)(A_f + \gamma^{-2}B_1B_1^T P_1(t))^T - P_3(t)C_2^T R_{20}^{-1}C_2P_3(t) + P$$

have solutions $P_1(t) \geq 0$, $P_2(t) \geq 0$, *and* $P_3(t) \geq 0$ *on* $[0,T]$ *with* $P_1(T) = 0$, $P_2(T) = 0$, *and* $P_3(0) = 0$.

If the solutions exist, then

$$w_* = \gamma^{-2}B_1^T P_1(t)x, \quad F_*(t) = -R_{02}^{-1}(D_{02}^T C_0 + B_2^T P_2(t)), \quad L_*(t) = -(B_0D_{20}^T + P_3(t)C_2^T)R_{20}^{-1}.$$

Proof. Suppose there exist solutions $P_1(t) \geq 0$, $P_2(t) \geq 0$, and $P_3(t) \geq 0$, $t \in [0,T]$, with $P_1(T) = 0$, $P_2(T) = 0$, and $P_3(0) = 0$ for the three differential Riccati equations. We have

$$J_1(u, w, w_0) = \gamma^2\|w\|_{[0,T]}^2 - \|z\|_{[0,T]}^2 = E\left\{\int_0^T (\gamma^2\|w\|^2 - \|z\|^2)dt\right\}$$

$$= E\left\{\int_0^T [\gamma^2\|w\|^2 - \|z\|^2 - \frac{d}{dt}(x^T P_1(t)x)]dt\right\}$$

$$= E\left\{\int_0^T [\gamma^2\|w\|^2 - \|z\|^2 - \dot{x}^T P_1(t)x - x^T\dot{P}_1(t)x - x^T P_1(t)\dot{x}]dt\right\}.$$

Using the equation for $\dot{P}_1(t)$, we get

$$J_1(u,w,w_0) = E\left\{\int_0^T [\gamma^2 \|w-w_*\|^2 - u^T R_1 u + \bar{u}_*^T R_1 \bar{u}_*\right.$$

$$\left. -2x^T(P_1(t)B_2 + C_1^T D_{12})(u-\bar{u}_*) - 2x^T P_1(t) B_0 w_0]dt\right\},$$

where $w_* = \gamma^{-2} B_1^T P_1(t) x$ and $\bar{u}_* = -R_{02}^{-1}(D_{02}^T C_0 + B_2^T P_2(t))x$, i.e., the optimal strategies for state feedback case. Clearly, if we take $w = w_*$, then for any u, hence for the optimally designed output feedback control law u_*, we have $J_1(u_*, w_*, w_0) \le J_1(u_*, w, w_0)$.

Now we design u_* to minimize $J_2(u, w_*, w_0)$. By substituting the w_* into the system equation, we get

$$\begin{aligned}
&\dot{x} = (A + \gamma^{-2} B_1 B_1^T P_1(t))x + B_0 w_0 + B_2 u, \quad x(0) = 0,\\
&y = C_2 x + D_{20} w_0, \quad R_0 := D_{20} D_{20}^T > 0,\\
&z_0 = C_0 x + D_{02} u, \quad R_{02} := D_{02}^T D_{02} > 0.
\end{aligned}$$

This is a standard LQG problem [1, 16]. Thus the optimal control law is given by

$$\begin{aligned}
&\dot{\hat{x}} = (A + \gamma^{-2} B_1 B_1^T P_1(t))\hat{x} + B_2 u_* + L_*(t)(C_2\hat{x} - y),\\
&u_* = F_*(t)\hat{x},
\end{aligned}$$

where $F_*(t) = -R_{02}^{-1}(D_{02}^T C_0 + B_2^T P_2(t))$, $L_*(t) = -(B_0 D_{20}^T + P_3(t) C_2^T) R_{20}^{-1}$, $P_2(t) > 0$, and $P_3(t) > 0$ solve

$$-\dot{P}_2(t) = (A_s + \gamma^{-2} B_1 B_1^T P_1(t))^T P_2(t) + P_2(t)(A_s + \gamma^{-2} B_1 B_1^T P_1(t))$$

$$-P_2(t) B_2 R_{02}^{-1} B_2^T P_2(t) + Q,$$

$$\dot{P}_3(t) = (A_f + \gamma^{-2} B_1 B_1^T P_1(t)) P_3(t) + P_3(t)(A_f + \gamma^{-2} B_1 B_1^T P_1(t))^T$$

$$-P_3(t) C_2^T R_{20}^{-1} C_2 P_3(t) + P,$$

with $P_2(T) = 0$ and $P_3(0) = 0$, and u_* achieves $J_2(u_*, w_*, w_0) \le J_2(u, w_*, w_0)$. □

THEOREM 10. *Suppose the state feedback control problem is solvable, i.e., there are $P_1(t) \ge 0$ and $P_2(t) \ge 0$ with $P_1(T) = 0$ and $P_2(T) = 0$ solving*

$$-\dot{P}_1(t) = (A_s - B_2 R_{02}^{-1} B_2^T P_2(t))^T P_1(t) + P_1(t)(A_s - B_2 R_{02}^{-1} B_2^T P_2(t)) + \gamma^{-2} P_1(t) B_1 B_1^T P_1(t)$$

$$+[C_1 - D_{12} R_{02}^{-1}(D_{02}^T C_0 + B_2^T P_2(t))]^T [C_1 - D_{12} R_{02}^{-1}(D_{02}^T C_0 + B_2^T P_2(t))],$$

$$-\dot{P}_2(t) = (A_s + \gamma^{-2} B_1 B_1^T P_1(t))^T P_2(t) + P_2(t)(A_s + \gamma^{-2} B_1 B_1^T P_1(t))$$

$$-P_2(t) B_2 R_{02}^{-1} B_2^T P_2(t) + Q,$$

and there is a w_* and an output feedback control u_* (hence an $L_*(t)$) in the given form of

$$\begin{aligned}\dot{\hat{x}} &= (A+\gamma^{-2}B_1B_1^TP_1(t)+B_2F_*(t))\hat{x}+L_*(t)(C_2\hat{x}-y),\\ u_* &= F_*(t)\hat{x},\end{aligned}$$

where $F_*(t)=-R_{02}^{-1}(D_{02}^TC_0+B_2^TP_2(t))$, such that

$$J_1(u_*,w_*,w_0)\le J_1(u_*,w,w_0),\quad J_2(u_*,w_*,w_0)\le J_2(u,w_*,w_0);$$

then there is a $P_3(t)\ge 0$ with $P_3(0)=0$ solving

$$\begin{aligned}\dot{P}_3(t) = (A_f+\gamma^{-2}B_1B_1^TP_1(t))P_3(t)+P_3(t)(A_f+\gamma^{-2}B_1B_1^TP_1(t))^T\\ -P_3(t)C_2^TR_{20}^{-1}C_2P_3(t)+P,\end{aligned}$$

and $L_*(t)$ can be chosen as $L_*(t)=-(B_0D_{20}^T+P_3(t)C_2^T)R_{20}^{-1}$.

Proof. Let $P_1(t)\ge 0$ and $P_2(t)\ge 0$ with $P_1(T)=0$ and $P_2(T)=0$ solve

$$\begin{aligned}-\dot{P}_1(t) &= (A_s-B_2R_{02}^{-1}B_2^TP_2(t))^TP_1(t)\\ &\quad +P_1(t)(A_s-B_2R_{02}^{-1}B_2^TP_2(t))+\gamma^{-2}P_1(t)B_1B_1^TP_1(t)\\ &\quad +[C_1-D_{12}R_{02}^{-1}(D_{02}^TC_0+B_2^TP_2(t))]^T[C_1-D_{12}R_{02}^{-1}(D_{02}^TC_0+B_2^TP_2(t))],\\ -\dot{P}_2(t) &= (A_s+\gamma^{-2}B_1B_1^TP_1(t))^TP_2(t)+P_2(t)(A_s+\gamma^{-2}B_1B_1^TP_1(t))\\ &\quad -P_2(t)B_2R_{02}^{-1}B_2^TP_2(t)+Q.\end{aligned}$$

Since

$$\begin{aligned}J_1(u_*,w,w_0) &= E\left\{\int_0^T(\gamma^2\|w\|^2-\|z\|^2)dt\right\}\\ &= E\left\{\int_0^T\left[\gamma^2\|w\|^2-\|z\|^2-\frac{d}{dt}(x^TP_1(t)x)\right]dt\right\}\\ &= E\left\{\int_0^T[\gamma^2\|w\|^2-\|z\|^2-\dot{x}^TP_1(t)x-x^T\dot{P}_1(t)x-x^TP_1(t)\dot{x}]dt\right\},\end{aligned}$$

using the equation for $\dot{P}_1(t)$, we get

$$\begin{aligned}E\left\{\int_0^T(\gamma^2\|w\|^2-\|z\|^2)dt\right\} = E\left\{\int_0^T[\gamma^2\|w-w_*\|^2-u^TR_1u+\bar{u}_*^TR_1\bar{u}_*\right.\\ \left.-2x^T(P_1(t)B_2+C_1^TD_{12})(u-\bar{u}_*)-2x^TP_1(t)B_0w_0]dt\right\},\end{aligned}$$

where $w_*=\gamma^{-2}B_1^TP_1(t)x$, $\bar{u}_*=-R_{02}^{-1}(D_{02}^TC_0+B_2^TP_2(t))x=F_*(t)x$. Hence u_* and w_* achieve $J_1(u_*,w_*,w_0)\le J_1(u_*,w,w_0)$.

Substituting w_* into the system equation, we get

$$\begin{aligned}\dot{x} &= (A+\gamma^{-2}B_1B_1^TP_1(t))x+B_0w_0+B_2u_*,\\ y &= C_2x+D_{20}w_0,\\ z_0 &= C_0x+D_{02}u_*.\end{aligned}$$

Thus

$$J_2(u_*, w_*, w_0) = E\int_0^T \|z_0\|^2 dt$$
$$= E\int_0^T \left[x^T C_0^T C_0 x + 2x^T C_0^T D_{02} u_* + u_*^T R_{02} u_* + \frac{d}{dt}(x^T P_2(t) x)\right] dt$$

$$= E\int_0^T [x^T C_0^T C_0 x + 2x^T C_0^T D_{02} u_* + u_*^T R_{02} u_* + \dot{x}^T P_2(t) x + x^T \dot{P}_2(t) x + x^T P_2(t) \dot{x}] dt.$$

Using the equation about $P_2(t)$, we get

$$J_2(u_*, w_*, w_0) = E\int_0^T (u_* - \bar{u}_*)^T R_{02}(u_* - \bar{u}_*) dt + 2E\int_0^T x^T P_2(t) B_0 w_0 dt$$

$$= E\int_0^T (u_* - \bar{u}_*)^T R_{02}(u_* - \bar{u}_*) dt + 2\mathrm{trace}\int_0^T P_2(t) B_0 E\{w_0 x^T\} dt.$$

By Lemma 5, we have $E\{x w_0^T\} = B_0/2$. Hence

$$J_2(u_*, w_*, w_0) = E\int_0^T (u_* - \bar{u}_*)^T R_{02}(u_* - \bar{u}_*) dt + \mathrm{trace}\int_0^T P_2(t) B_0 B_0^T dt$$
$$= E\int_0^T (u_* - \bar{u}_*)^T R_{02}(u_* - \bar{u}_*) dt + \mathrm{trace}\int_0^T B_0^T P_2(t) B_0 dt.$$

We need only to consider the first term. Define $e_x = x - \hat{x}$; then

$$E\int_0^T (u_* - \bar{u}_*)^T R_{02}(u_* - \bar{u}_*) dt = E\int_0^T e_x^T F_*^T(t) R_{02} F_*(t) e_x dt.$$

Since we have

$$\dot{\hat{x}} = (A + \gamma^{-2} B_1 B_1^T P_1(t) + B_2 F_*(t))\hat{x} + L_*(t)(C_2 \hat{x} - y),$$
$$u_* = F_*(t)\hat{x},$$

clearly, e_x satisfies

$$\dot{e}_x = \dot{x} - \dot{\hat{x}} = (A + \gamma^{-2} B_1 B_1^T P_1(t) + L_*(t) C_2) e_x + (B_0 + L_*(t) D_{20}) w_0 := A_{L_*} e_x + B_{L_*} w_0.$$

Let $\Phi(t, \tau)$ be the transition matrix of $A + \gamma^{-2} B_1 B_1^T P_1(t) + L_*(t) C_2$; then

$$e_x = \int_0^t \Phi(t, \tau)(B_0 + L_*(t) D_{20}) w_0 d\tau.$$

This gives

$$
\begin{aligned}
&E\int_0^T e_x^T F_*^T(t)R_{02}F_*(t)e_x dt\\
&\quad = E\left\{\int_0^T\int_0^t\int_0^t w^T(\tau)B_{L_*}^T\Phi^T(t,\tau)F_*^T(t)R_{02}F_*(t)\Phi(t,\tau)B_{L_*}w(s)d\tau\, ds\, dt\right\}\\
&\quad = \operatorname{trace}\left\{\int_0^T\int_0^t\int_0^t F_*^T(t)R_{02}F_*(t)\Phi(t,\tau)B_{L_*}E\{w(s)w(\tau)\}B_{L_*}^T\Phi^T(t,\tau)d\tau\, ds\, dt\right\}\\
&\quad = \operatorname{trace}\left\{\int_0^T\int_0^t\int_0^t F_*^T(t)R_{02}F_*(t)\Phi(t,\tau)B_{L_*}\delta(\tau-s)B_{L_*}^T\Phi^T(t,\tau)d\tau\, ds\, dt\right\}\\
&\quad = \operatorname{trace}\left\{\int_0^T D_{02}F_*(t)Y(t)F_*^T(t)D_{02}^T dt\right\},
\end{aligned}
$$

where $Y(t)=\int_0^t\Phi(t,s)B_{L_*}B_{L_*}^T\Phi^T(t,s)ds\geq 0$ satisfies

$$\dot{Y}(t)=A_{L_*}Y(t)+Y(t)A_{L_*}^T+B_{L_*}B_{L_*}^T,\quad Y(0)=0.$$

Since

$$J_2(u_*,w_*,w_0)=\operatorname{trace}\left\{\int_0^T D_{02}F_*(t)Y(t)F_*^T(t)D_{02}^T dt\right\}+\operatorname{trace}\int_0^T B_0^T P_2(t)B_0 dt$$

is the minimum value by assumption, by Theorem 6 there is a $P_3(t)\geq 0,\ P_3(t)\leq Y(t),\ \ \forall t\in[0,T]$, with $P_3(0)=0$ solving

$$\dot{P}_3(t)=(A_f+\gamma^{-2}B_1B_1^TP_1(t))P_3(t)+P_3(t)(A_f+\gamma^{-2}B_1B_1^TP_1(t))^T$$

$$-P_3(t)C_2^TR_{20}^{-1}C_2P_3(t)+P,$$

and, besides, $L_*(t)$ can be chosen as $L_*(t)=-(B_0D_{20}^T+P_3(t)C_2^T)R_{20}^{-1}$ since

$$J_2(u_*,w_*,w_0)=\operatorname{trace}\left\{\int_0^T D_{02}F_*(t)P_3(t)F_*^T(t)D_{02}^T dt\right\}+\operatorname{trace}\int_0^T B_0^TP_2(t)B_0dt.$$

This concludes the proof. □

4.2. Multiobjective control design—infinite time horizon. For multiobjective control design in infinite time horizon, we need to add the following standard assumptions:

(A1) (A,B_2) is stabilizable and (C_2,A) is detectable,

(A2) $\begin{bmatrix} A-j\omega I & B_2\\ C_0 & D_{02}\end{bmatrix}$ has full column rank for all ω,

(A3) $\begin{bmatrix} A-j\omega I & B_0\\ C_2 & D_{20}\end{bmatrix}$ has full row rank for all ω.

The state feedback design result [22] is presented in the next theorem.

THEOREM 11. *There exist linear memoryless state feedback strategies (Nash equilibrium strategies) u_* and w_* such that*

$$0\leq J_3(u_*,w_*,0)\leq J_3(u_*,w,0),\ \ J_4(u_*,w_*,0)\leq J_4(u,w_*,0)$$

if and only if the coupled Riccati equations

$$(A_s - B_2R_{02}^{-1}B_2^TP_2)^TP_1 + P_1(A_s - B_2R_{02}^{-1}B_2^TP_2) + \gamma^{-2}P_1B_1B_1^TP_1$$

$$+[C_1 - D_{12}R_{02}^{-1}(D_{02}^TC_0 + B_2^TP_2)]^T[C_1 - D_{12}R_{02}^{-1}(D_{02}^TC_0 + B_2^TP_2)] = 0$$

and

$$(A_s + \gamma^{-2}B_1B_1^TP_1)^TP_2 + P_2(A_s + \gamma^{-2}B_1B_1^TP_1) - P_2B_2R_{02}^{-1}B_2^TP_2 + Q = 0$$

have stabilizing solutions $P_1 \geq 0$ and $P_2 \geq 0$, i.e., $A_s + \gamma^{-2}B_1B_1^TP_1 - B_2R_{02}^{-1}B_2^TP_2$ is stable.

Furthermore, if the solutions exist, we have $w_* = \gamma^{-2}B_1^TP_1x$ and $u_* = F_*x$ with $F_* := -R_{02}^{-1}(D_{02}^TC_0 + B_2^TP_2)$.

Now we prove the results of output feedback design.

THEOREM 12. *There exist a w_* and an output feedback control law u_* such that*

$$J_3(u_*, w_*, w_0) \leq J_3(u_*, w, w_0), \quad J_4(u_*, w_*, w_0) \leq J_4(u, w_*, w_0)$$

if the coupled Riccati algebraic equations

$$(A_s - B_2R_{02}^{-1}B_2^TP_2)^TP_1 + P_1(A_s - B_2R_{02}^{-1}B_2^TP_2) + \gamma^{-2}P_1B_1B_1^TP_1$$

$$+[C_1 - D_{12}R_{02}^{-1}(D_{02}^TC_0 + B_2^TP_2)]^T[C_1 - D_{12}R_{02}^{-1}(D_{02}^TC_0 + B_2^TP_2)] = 0,$$

$$(A_s + \gamma^{-2}B_1B_1^TP_1)^TP_2 + P_2(A_s + \gamma^{-2}B_1B_1^TP_1) - P_2B_2R_{02}^{-1}B_2^TP_2 + Q = 0$$

and

$$(A_f + \gamma^{-2}B_1B_1^TP_1)P_3 + P_3(A_f + \gamma^{-2}B_1B_1^TP_1)^T - P_3C_2^TR_{20}^{-1}C_2P_3 + P = 0$$

have stabilizing solutions $P_1 \geq 0$, $P_2 \geq 0$, and $P_3 \geq 0$, i.e., both $A_s + \gamma^{-2}B_1B_1^TP_1 - B_2R_{02}^{-1}B_2^TP_2$ and $A_f + \gamma^{-2}B_1B_1^TP_1 - P_3C_2^TR_{20}^{-1}C_2$ are stable.

If the solutions exist, we have $w_* = \gamma^{-2}B_1^TP_1x$, and u_* is of the form

$$\begin{aligned}\dot{\hat{x}} &= (A + \gamma^{-2}B_1B_1^TP_1 + B_2F_*)\hat{x} + L_*(C_2\hat{x} - y),\\ u_* &= F_*\hat{x},\end{aligned}$$

where $F_* = -R_{02}^{-1}(D_{02}^TC_0 + B_2^TP_2)$ and $L_* = -(B_0D_{20}^T + P_3C_2^T)R_0^{-1}$.

Conversely, if the state feedback control problem is solvable, i.e., there are stabilizing solutions $P_1 \geq 0$ and $P_2 \geq 0$ for

$$(A_s - B_2R_{02}^{-1}B_2^TP_2)^TP_1 + P_1(A_s - B_2R_{02}^{-1}B_2^TP_2) + \gamma^{-2}P_1B_1B_1^TP_1$$

$$+[C_1 - D_{12}R_{02}^{-1}(D_{02}^TC_0 + B_2^TP_2)]^T[C_1 - D_{12}R_{02}^{-1}(D_{02}^TC_0 + B_2^TP_2)] = 0,$$

$$(A_s + \gamma^{-2}B_1B_1^TP_1)^TP_2 + P_2(A_s + \gamma^{-2}B_1B_1^TP_1) - P_2B_2R_{02}^{-1}B_2^TP_2 + Q = 0,$$

and there is a w_* and an optimal control u_* in the form of

$$\begin{aligned}\dot{\hat{x}} &= (A + \gamma^{-2}B_1B_1^TP_1 + B_2F_*)\hat{x} + L_*(C_2\hat{x} - y),\\ u_* &= F_*\hat{x},\end{aligned}$$

where $F_* = -R_{02}^{-1}(D_{02}^T C_0 + B_2^T P_2)$ *such that*

$$J_3(u_*, w_*, w_0) \le J_3(u_*, w, w_0), \quad J_4(u_*, w_*, w_0) \le J_4(u, w_*, w_0),$$

then there is a $P_3 \ge 0$ *solving*

$$(A_f + \gamma^{-2} B_1 B_1^T P_1) P_3 + P_3 (A_f + \gamma^{-2} B_1 B_1^T P_1)^T - P_3 C_2^T R_0^{-1} C_2 P_3 + P = 0.$$

Moreover, if $A + \gamma^{-2} B_1 B_1^T P_1 - (B_0 D_{20}^T + P_3 C_2^T) R_0^{-1} C_2$ *is stable, then* $L_* = -(B_0 D_{20}^T + P_3 C_2^T) R_0^{-1}$.

Proof. Suppose there exist stabilizing solutions $P_1 \ge 0$, $P_2 \ge 0$, and $P_3 \ge 0$ to the following Riccati equations:

$$(A_s - B_2 R_{02}^{-1} B_2^T P_2)^T P_1 + P_1 (A_s - B_2 R_{02}^{-1} B_2^T P_2) + \gamma^{-2} P_1 B_1 B_1^T P_1$$

$$+ [C_1 - D_{12} R_{02}^{-1} (D_{02}^T C_0 + B_2^T P_2)]^T [C_1 - D_{12} R_{02}^{-1} (D_{02}^T C_0 + B_2^T P_2)] = 0,$$

$$(A_s + \gamma^{-2} B_1 B_1^T P_1)^T P_2 + P_2 (A_s + \gamma^{-2} B_1 B_1^T P_1) - P_2 B_2 R_{02}^{-1} B_2^T P_2 + Q = 0,$$

and

$$(A_f + \gamma^{-2} B_1 B_1^T P_1) P_3 + P_3 (A_f + \gamma^{-2} B_1 B_1^T P_1)^T - P_3 C_2^T R_0^{-1} C_2 P_3 + P = 0.$$

Let u be any stabilizing control law. Since

$$\begin{aligned} J_3(u, w, w_0) &= \gamma^2 \|w\|_{\mathcal{P}}^2 - \|z\|_{\mathcal{P}}^2 \\ &= \lim_{T \to \infty} E \left\{ \frac{1}{T} \int_0^T (\gamma^2 \|w\|^2 - \|z\|^2) dt \right\} \\ &= \lim_{T \to \infty} E \left\{ \frac{1}{T} \int_0^T (\gamma^2 \|w\|^2 - x^T C_1^T C_1 x - 2 x^T C_1^T D_{12} u - u^T R_{12} u) dt \right\}, \end{aligned}$$

using the equation for P_1, we get

$$J_3(u, w, w_0) = E \left\{ \lim_{T \to \infty} \int_0^T [\gamma^2 \|w - w_*\|^2 - u^T R_{12} u + \bar{u}_*^T R_1 \bar{u}_* \right.$$

$$\left. - 2 x^T (P_1 B_2 + C_1^T D_{12})(u - \bar{u}_*) - 2 x^T P_1 B_0 w_0] dt \right\},$$

where $R_1 := D_{12}^T D_{12}$, $w_* = \gamma^{-2} B_1^T P_1 x$, and $\bar{u}_* = -R_{02}^{-1}(D_{02}^T C_0 + B_2^T P_2) x$, i.e., the optimal strategies for state feedback case. Clearly, if we take $w = w_*$, then for any u, and hence for the optimally designed output feedback control law u_*, we have $J_3(u_*, w_*, w_0) \le J_3(u_*, w, w_0)$. Now we design u_* to minimize J_4. By substituting the w_* into the system equation, we get

$$\begin{aligned} \dot{x} &= (A + \gamma^{-2} B_1 B_1^T P_1) x + B_0 w_0 + B_2 u, \\ y &= C_2 x + D_{20} w_0, \\ z_0 &= C_0 x + D_{02} u. \end{aligned}$$

Clearly, for the index functional J_4, it is a standard LQG problem [1, 16]. Thus the optimal control law is given by

$$\begin{aligned}\dot{\hat{x}} &= (A+\gamma^{-2}B_1B_1^TP_1)\hat{x}+B_2u_*+L_*(C_2\hat{x}-y),\\ u_* &= F_*\hat{x},\end{aligned}$$

where $F_* = -R_{02}^{-1}(D_{02}^TC_0+B_2^TP_2)$, $L_* = -(B_0D_{20}^T+P_3C_2^T)R_0^{-1}$, and u_* achieves $J_4(u_*,w_*,w_0)\le J_4(u,w_*,w_0)$. Note that $J_4(u_*,w_*,w_0)$ can be calculated as

$$J_4(u_*,w_*,w_0) = \operatorname{trace}\left\{B_0B_0^TP_2\right\}+\operatorname{trace}\left\{F_*^TR_{02}F_*P_3\right\}.$$

Conversely, suppose the state feedback control problem is solvable, i.e., there are stabilizing solutions to

$$(A_s-B_2R_{02}^{-1}B_2^TP_2)^TP_1+P_1(A_s-B_2R_{02}^{-1}B_2^TP_2)+\gamma^{-2}P_1B_1B_1^TP_1$$

$$+[C_1-D_{12}R_{02}^{-1}(D_{02}^TC_0+B_2^TP_2)]^T[C_1-D_{12}R_{02}^{-1}(D_{02}^TC_0+B_2^TP_2)]=0,$$

$$(A_s+\gamma^{-2}B_1B_1^TP_1)^TP_2+P_2(A_s+\gamma^{-2}B_1B_1^TP_1)-P_2B_2R_{02}^{-1}B_2^TP_2+Q=0.$$

Let u_* be in the form of

$$\begin{aligned}\dot{\hat{x}} &= (A+\gamma^{-2}B_1B_1^TP_1+B_2F_*)\hat{x}+L_*(C_2\hat{x}-y),\\ u_* &= F_*\hat{x},\end{aligned}$$

where $F_* = -R_{02}^{-1}(D_{02}^TC_0+B_2^TP_2)$ and u_* and w_* achieve

$$J_3(u_*,w_*,w_0)\le J_3(u_*,w,w_0),\quad J_4(u_*,w_*,w_0)\le J_4(u,w_*,w_0).$$

From the proof of the sufficiency, we get $w_* = \gamma^{-2}B_1^TP_1x$. Substituting w_* and u_* into the system equations, we have

$$\begin{aligned}\dot{x} &= (A+\gamma^{-2}B_1B_1^TP_1)x+B_0w_0+B_2u_*,\\ y &= C_2x+D_{20}w_0,\\ z_0 &= C_0x+D_{02}u_*.\end{aligned}$$

Since

$$J_4(u_*,w_*,w_0) = \|z_0\|_{\mathcal{P}}^2 = \lim_{T\to\infty}E\left\{\frac{1}{T}\int_0^T\|z_0\|^2dt\right\}$$

$$=\lim_{T\to\infty}E\left\{\frac{1}{T}\int_0^T[x^TC_0^TC_0x+2x^TC_0^TD_{02}u_*+u_*^TR_{02}u_*]dt\right\},$$

using the equation about P_2, we get

$$J_4(u_*,w_*,w_0) = \lim_{T\to\infty}E\left\{\frac{1}{T}\int_0^T(u_*-\bar{u}_*)^TR_{02}(u_*-\bar{u}_*)dt\right\}$$

$$
\begin{aligned}
&+2\lim_{T\to\infty} E\left\{\frac{1}{T}\int_0^T x^T P_2 B_0 w_0 dt\right\}\\
&=\lim_{T\to\infty} E\left\{\frac{1}{T}\int_0^T (u_*-\bar{u}_*)^T R_{02}(u_*-\bar{u}_*)dt\right\}\\
&\quad+2\mathrm{trace}\left\{\lim_{T\to\infty}\frac{1}{T}\int_0^T P_2 B_0 E\{w_0 x^T\}dt\right\}\\
&=\lim_{T\to\infty} E\left\{\frac{1}{T}\int_0^T (u_*-\bar{u}_*)^T R_{02}(u_*-\bar{u}_*)dt\right\}\\
&\quad+\mathrm{trace}(B_0^T P_2 B_0),
\end{aligned}
$$

where $\bar{u}_* = -R_{02}^{-1}(D_{02}^T C_0 + B_2^T P_2)x$. We need only to consider the first term. Define $e_x = x - \hat{x}$; then

$$
\lim_{T\to\infty} E\left\{\frac{1}{T}\int_0^T (u_*-\bar{u}_*)^T R_{02}(u_*-\bar{u}_*)dt\right\} = \lim_{T\to\infty} E\left\{\frac{1}{T}\int_0^T e_x^T F_*^T R_{02} F_* e_x dt\right\},
$$

where e_x satisfies

$$
\dot{e}_x = \dot{x} - \dot{\hat{x}} = (A + \gamma^{-2} B_1 B_1^T P_1 + L_* C_2) e_x + (B_0 + L_* D_{20}) w_0 = A_{L_*} e_x + B_{L_*} w_0.
$$

Hence $e_x = \int_0^t e^{A_{L_*}(t-\tau)} B_{L_*} w_0 \ d\tau$. This gives

$$
\lim_{T\to\infty} E\left\{\frac{1}{T}\int_0^T e_x^T F_*^T R_{02} F_* e_x dt\right\} = \mathrm{trace}(D_{02} F_* Y F_*^T D_{02}^T),
$$

where $Y = \int_0^\infty \exp(A_{L_*} s) B_{L_*} B_{L_*}^T \exp(A_{L_*}^T s) ds \geq 0$ satisfies $A_{L_*} Y + Y A_{L_*}^T + B_{L_*} B_{L_*}^T = 0$. Since

$$
J_4(u_*, w_*, w_0) = \mathrm{trace}(D_{02} F_* Y F_*^T D_{02}^T) + \mathrm{trace}(B_0^T P_2 B_0)
$$

is the minimum value, by Theorem 7, there is a $P_3 \geq 0,\ P_3 \leq Y$ solving

$$
(A_f + \gamma^{-2} B_1 B_1^T P_1) P_3 + P_3 (A_f + \gamma^{-2} B_1 B_1^T P_1)^T - P_3 C_2^T R_0^{-1} C_2 P_3 + P = 0.
$$

If $A + \gamma^{-2} B_1 B_1^T P_1 - (B_0 D_{20}^T + P_3 C_2^T) R_0^{-1} C_2$ is stable, then L_* can be chosen as $L_* = -(B_0 D_{20}^T + P_3 C_2^T) R_0^{-1}$ because $J_4(u_*, w_*, w_0) = \mathrm{trace}(D_{02} F_* P_3 F_*^T D_{02}^T) + \mathrm{trace}(B_0^T P_2 B_0)$. This concludes the proof. □

5. $\mathcal{H}_\infty$ Gaussian control design. In this section, we give the solution to the problem formulated in section 2.3. To simplify the notations, we shall introduce the following abbreviations:

$$
A_x := A - B_2 R_1^{-1} D_{12}^T C_1, \quad A_y := A - B_0 D_{20}^T R_0^{-1} C_2,
$$

$$
P := B_0 (I - D_{20}^T R_0^{-1} D_{20}) B_0^T, \quad Q := C_1^T (I - D_{12} R_1^{-1} D_{12}^T) C_1.
$$

5.1. $\mathcal{H}_\infty$ Gaussian control design—finite time horizon. The design results for $\mathcal{H}_\infty$ Gaussian control in the finite time horizon are summarized in the next theorem.

THEOREM 13. *Let the dynamical system G be described by* (5)–(7). *If there are solutions $P_1(t) \geq 0$, $P_2(t) \geq 0$, and $P_3(t) \geq 0$ with $P_1(T) = 0$, $P_2(T) = 0$, and $P_3(0) = 0$ solving the differential Riccati equations*

$$A_x^T P_1(t) + P_1(t)A_x + P_1(t)(B_1 B_1^T/\gamma^2 - B_2 R_1^{-1} B_2^T)P_1(t) + Q = -\dot{P}_1(t),$$

$$P_2(t)(A_y+\gamma^{-2}B_1B_1^TP_1(t)-P_3(t)C_2^TR_0^{-1}C_2)+(A_y+\gamma^{-2}B_1B_1^TP_1(t)-P_3(t)C_2^TR_0^{-1}C_2)^TP_2(t)$$

$$+\gamma^{-2}P_2(t)B_1B_1^TP_2(t) + (D_{12}^TC_1 + B_2^TP_1(t))^TR_1^{-1}(D_{12}^TC_1 + B_2^TP_1(t)) = -\dot{P}_2(t),$$

$$[A_y + \gamma^{-2}B_1B_1^T(P_1(t) + P_2(t))]P_3(t) + P_3(t)[A_y + \gamma^{-2}B_1B_1^T(P_1(t) + P_2(t))]^T$$

$$-P_3(t)C_2^TR_0^{-1}C_2P_3(t) + P = \dot{P}_3(t),$$

then there exist an optimal control law u_* and a worst disturbance signal w_* such that

$$J_1(u_*, w_*, w_0) \leq J_1(u_*, w, w_0), \quad J_2(u_*, w_*, w_0) \leq J_2(u, w_*, w_0).$$

If the solutions exist, we have the optimal controller

$$\begin{aligned}\dot{\hat{x}} &= (A + \gamma^{-2}B_1B_1^TP_1(t) + L_*(t)C_2 + B_2F_*(t))\hat{x} - L_*(t)y, \quad \hat{x}(0) = 0,\\ u_* &= F_*(t)\hat{x},\end{aligned}$$

where $F_*(t) := -R_1^{-1}(D_{12}^TC_1 + B_2^TP_1(t))$, $L_*(t) = -(B_0D_{20}^T + P_3(t)C_2^T)R_0^{-1}$, and $w_* = \gamma^{-2}B_1^T(P_1(t)x + P_2(t)e_x)$, $e_x = x - \hat{x}$.

Conversely, let $P_1(t) \geq 0$ with $P_1(T) = 0$ solve

$$A_x^T P_1(t) + P_1(t)A_x + P_1(t)(B_1 B_1^T/\gamma^2 - B_2 R_1^{-1} B_2^T)P_1(t) + Q = -\dot{P}_1(t).$$

Suppose there exist a w_*' and an optimal control u_* (hence an $L_*(t)$) in the form

$$\begin{aligned}\dot{\hat{x}} &= (A + \gamma^{-2}B_1B_1^TP_1(t) + L_*(t)C_2 + B_2F_*(t))\hat{x} - L_*(t)y, \quad \hat{x}(0) = 0,\\ u_* &= F_*(t)\hat{x},\end{aligned}$$

where $F_*(t) = -R_1^{-1}(D_{12}^TC_1 + B_2^TP_1(t))$ such that $0 < J_1(u_*, w_*', 0) \leq J_1(u_*, w, 0)$. Then, there is a w_* plus the same u_* achieving $J_1(u_*, w_*, w_0) \leq J_1(u_*, w, w_0)$.

If, furthermore, u_* and w_* also achieve $J_2(u_*, w_*, w_0) \leq J_2(u, w_*, w_0)$, then there exist $P_2(t) \geq 0$, $P_3(t) \geq 0$ with $P_2(T) = 0$ and $P_3(0) = 0$ solving

$$P_2(t)(A_y+\gamma^{-2}B_1B_1^TP_1(t)-P_3(t)C_2^TR_0^{-1}C_2)+(A_y+\gamma^{-2}B_1B_1^TP_1(t)-P_3(t)C_2^TR_0^{-1}C_2)^TP_2(t)$$

$$+\gamma^{-2}P_2(t)B_1B_1^TP_2(t) + (D_{12}^TC_1 + B_2^TP_1(t))^TR_1^{-1}(D_{12}^TC_1 + B_2^TP_1(t)) = -\dot{P}_2(t),$$

$$[A_y + \gamma^{-2}B_1B_1^T(P_1(t) + P_2(t))]P_3(t) + P_3(t)[A_y + \gamma^{-2}B_1B_1^T(P_1(t) + P_2(t))]^T$$

$$-P_3(t)C_2^T R_0^{-1} C_2 P_3(t) + P = \dot{P}_3(t),$$

and $J_2(L(t))$ *is minimized by* $L_*(t) = -(B_0 D_{20}^T + P_3(t)C_2^T)R_0^{-1}$.

Proof (sufficiency). Suppose there exist solutions $P_1(t) \geq 0$, $P_2(t) \geq 0$, and $P_3(t) \geq 0$ $\forall t \in [0,T]$ with $P_1(T) = 0$, $P_2(T) = 0$, and $P_3(0) = 0$ solving those three differential Riccati equations.

Using the first Riccati equation and Lemma 5 to complete square for $J_1(u, w, w_0)$, we get

$$\begin{aligned} J_1(u,w,w_0) &= \gamma^2 \|w\|_{[0,T]}^2 - \|z\|_{[0,T]}^2 \\ &= \gamma^2 \|w - \tilde{w}_*\|_{[0,T]}^2 - \|D_{12}(u - \tilde{u}_*)\|_{[0,T]}^2 - \int_0^T \operatorname{trace}\{B_0^T P_1(t) B_0\} dt, \end{aligned}$$

where $\tilde{w}_* = \gamma^{-2} B_1^T P_1(t) x$ and $\tilde{u}_* = -R_1^{-1}(D_{12}^T C_1 + B_2^T P_1(t))x$. Define

$$r := w - \gamma^{-2} B_1^T P_1(t) x, \quad v := D_{12}\left\{u + R_1^{-1}(D_{12}^T C_1 + B_2^T P_1(t))x\right\}.$$

Then the system equations can be rewritten as

$$\begin{aligned} \dot{x} &= (A + \gamma^{-2} B_1 B_1^T P_1(t))x + B_0 w_0 + B_1 r + B_2 u, \quad x(0) = 0, \\ v &= D_{12}\left\{R_1^{-1}(D_{12}^T C_1 + B_2^T P_1(t))x + u\right\}, \\ y &= C_2 x + D_{20} w_0, \end{aligned}$$

and the performance index $J_1(u, w, w_0)$ becomes

$$J_1(u,w,w_0) = \gamma^2 \|r\|_{[0,T]}^2 - \|v\|_{[0,T]}^2 - \int_0^T \operatorname{trace}\{B_0^T P_1(t) B_0\} dt.$$

Using $L_* = -R_1^{-1}(D_{12}^T C_1 + B_2^T P_1(t))$ to construct a state estimator,

$$\dot{\hat{x}} = (A + \gamma^{-2} B_1 B_1^T P_1(t))\hat{x} + B_2 u + L_*(t)(C_2 \hat{x} - y), \quad \hat{x}(0) = 0,$$

a natural choice of the optimal control law $u = u_*$ would be $u_* = -R_1^{-1}(D_{12}^T C_1 + B_2^T P_1(t))\hat{x}$. Denote $e_x = x - \hat{x}$. Accordingly, the system can be further simplified into

$$\begin{aligned} \dot{e}_x &= (A + \gamma^{-2} B_1 B_1^T P_1(t) + L_*(t) C_2) e_x + (B_0 + L_*(t) D_{20}) w_0 + B_1 r, \quad e_x(0) = 0, \\ v &= D_{12} R_1^{-1}(D_{12}^T C_1 + B_2^T P_1(t)) e_x. \end{aligned}$$

Now we can complete square for $J_1(u, w, w_0)$ again by using the second Riccati equation and Lemma 5 to get

$$\begin{aligned} J_1(u_*, w, w_0) = \gamma^2 \|r - \gamma^{-2} B_1^T P_2(t) e_x\|_{[0,T]}^2 &- \int_0^T \operatorname{trace}\{B_0^T P_1(t) B_0\} dt \\ &- \int_0^T trace\{(B_0 + L_*(t) D_{20})^T P_2(t)(B_0 + L_*(t) D_{20})\} dt. \end{aligned}$$

We claim that the worst signal w_* can be taken as

$$w_* = \gamma^{-2} B_1^T (P_1(t) x + P_2(t) e_x) \quad \text{or} \quad r_* = \gamma^{-2} B_1^T P_2(t) e_x.$$

Indeed, substituting $w_* = \gamma^{-2}B_1^T(P_1(t)x+P_2(t)e_x)$ and $u_* = -R_1^{-1}(D_{12}^TC_1+B_2^TP_1(t))(x-e_x)$ into the system equation (5)–(7), we get

$$\dot{\bar{x}} = \bar{A}\bar{x} + \bar{B}w_0, \quad z = \bar{C}\bar{x},$$

where $\bar{x} = [\ x^T \quad e_x^T\ ]^T$,

$$\bar{A} = \begin{bmatrix} A+\gamma^{-2}B_1B_1^TP_1(t) - B_2R_1^{-1}(D_{12}^TC_1+B_2^TP_1(t)) & \gamma^{-2}B_1B_1^TP_2(t) + B_2R_1^{-1}(D_{12}^TC_1+B_2^TP_1(t)) \\ 0 & A+\gamma^{-2}B_1B_1^T(P_1(t)+P_2(t))+L_*(t)C_2 \end{bmatrix},$$

$$\bar{B} = [\ B_0^T \quad (B_0+L_*(t)D_{20})^T\ ]^T,$$
$$\bar{C} = [\ C_1 - D_{12}R_1^{-1}(D_{12}^TC_1+B_2^TP_1(t)) \quad D_{12}R_1^{-1}(D_{12}^TC_1+B_2^TP_1(t))\ ].$$

Hence $\bar{x} = \int_0^t e^{\bar{A}(t-\tau)}\bar{B}w_0 d\tau$. By Lemma 5, we have

$$E\{x(t)w_0^T(t)\} = \frac{1}{2}B_0, \quad E\{e_x(t)w_0^T(t)\} = \frac{1}{2}(B_0+L_*(t)D_{20}).$$

The first completed square gives

$$\begin{aligned} J_1(u_*,w_*,w_0) &= \gamma^2\|w_*\|_{[0,T]}^2 - \|z\|_{[0,T]}^2 \\ &= \gamma^2\|\gamma^{-2}B_1^TP_2(t)e_x\|_{[0,T]}^2 - \|D_{12}R_1^{-1}(D_{12}^TC_1+B_2^TP_1(t))e_x\|_{[0,T]}^2 \\ &\qquad -2\int_0^T \operatorname{trace}\{B_0^TP_1(t)E(xw_0^T)\}dt \\ &= \gamma^2\|\gamma^{-2}B_1^TP_2(t)e_x\|_{[0,T]}^2 - \|D_{12}R_1^{-1}(D_{12}^TC_1+B_2^TP_1(t))e_x\|_{[0,T]}^2 \\ &\qquad -\int_0^T \operatorname{trace}\{B_0^TP_1(t)B_0\}dt. \end{aligned}$$

Note that the second Riccati equation can be written as

$$P_2(t)(A+\gamma^{-2}B_1B_1^T(P_1(t)+P_2(t))+L_*(t)C_2)+(A+\gamma^{-2}B_1B_1^T(P_1(t)+P_2(t))+L_*(t)C_2)^TP_2(t)$$

$$-\gamma^{-2}P_2(t)B_1B_1^TP_2(t) + (D_{12}^TC_1+B_2^TP_1(t))^TR_1^{-1}(D_{12}^TC_1+B_2^TP_1(t)) = -\dot{P}_2(t).$$

Hence

$$\gamma^2\|\gamma^{-2}B_1^TP_2(t)e_x\|_{[0,T]}^2 - \|D_{12}R_1^{-1}(D_{12}^TC_1+B_2^TP_1(t))e_x\|_{[0,T]}^2$$

$$= E\int_0^T e_x^T[\gamma^{-2}P_2(t)B_1B_1^TP_2(t) - (D_{12}^TC_1+B_2^TP_1(t))^TR_1^{-1}(D_{12}^TC_1+B_2^TP_1(t))]e_x dt$$

$$= E\int_0^T e_x^T[P_2(t)(A+\gamma^{-2}B_1B_1^T(P_1(t)+P_2(t))+L_*(t)C_2)$$

$$+(A+\gamma^{-2}B_1B_1^T(P_1(t)+P_2(t))+L_*(t)C_2)^TP_2(t)+\dot{P}_2(t)]e_x dt$$

$$= E\int_0^T [e_x^T P_2(t)\dot{e}_x + \dot{e}_x^T P_2(t)e_x + e_x^T \dot{P}_2(t)e_x - 2e_x P_2(t)(B_0 + L_*(t)C_2)w_0]dt$$

$$= -2\int_0^T \text{trace}\{(B_0 + L_*(t)C_2)^T P_2(t)E(e_x w_0^T)\}dt$$

$$= -\int_0^T \text{trace}\{(B_0 + L_*(t)C_2)^T P_2(t)(B_0 + L_*(t)C_2)\}dt.$$

Therefore,

$$J_1(u_*, w_*, w_0) = -\int_0^T \{\text{trace}\{(B_0 + L_*(t)C_2)^T P_2(t)(B_0 + L_* C_2(t))\} + \text{trace}\{B_0^T P_1(t)B_0\}\}dt.$$

Clearly, $J_1(u_*, w_*, w_0) \le J_1(u_*, w, w_0)$ for all w which is independent from w_0. Hence w_* is the worst signal.

Next, it is shown that u_* does minimize the index $J_2(u, w_*, w_0)$ under the worst disturbance w_*. Let $L(t)$ be any filter gain. Substituting w_* (or r_*) into the system equations, we get

$$\dot{e}_x = (A + \gamma^{-2}B_1B_1^T P_1(t) + L(t)C_2 + \gamma^{-2}B_1B_1^T P_2(t))e_x + (B_0 + L(t)D_{20})w_0, \quad e_x(0) = 0,$$
$$:= A_L e_x + B_L w_0.$$

Let $\Phi(t,\tau)$ be the transition matrix for A_L; then $e_x = \int_0^t \Phi(t,\tau)B_L w_0(\tau)d\tau$ and

$$J_2(u, w_*, w_0) = \|e_x\|^2_{[0,T]} = E\left\{\int_0^T \int_0^t \int_0^t w_0^T(\tau)B_L^T\Phi^T(t,\tau)\Phi(t,s)B_L w_0(s)d\tau\, ds\, dt\right\}$$

$$= \text{trace}\left\{\int_0^T \int_0^t \int_0^t \Phi(t,s)B_L E\{w_0(s)w_0^T(\tau)\}B_L^T\Phi^T(t,\tau)d\tau\, ds\, dt\right\}$$

$$= \text{trace}\left\{\int_0^T \int_0^t \int_0^t \Phi(t,s)B_L\delta(\tau - s)B_L^T\Phi^T(t,\tau)d\tau\, ds\, dt\right\}$$

$$= \text{trace}\left\{\int_0^T \int_0^t \Phi(t,s)B_L B_L^T\Phi^T(t,s)ds\, dt\right\} = \text{trace}\{\int_0^T Y(t)dt\},$$

where $Y(t) = \int_0^t \Phi(t,s)B_L B_L^T\Phi^T(t,s)ds \ge 0$ satisfies $A_L Y(t) + Y(t)A_L^T + B_L B_L^T = \dot{Y}(t)$. By Theorem 6 and applying the third Riccati equation, $J_2(u, w*, w_0)$ achieves the minimum value at $L(t) = L_*(t)$, which means that u_* is the desired optimal control. Thus u_* and w_* achieve

$$J_1(u_*, w_*, w_0) \le J_1(u_*, w, w_0), \quad J_2(u_*, w_*, w_0) \le J_2(u, w_*, w_0).$$

Proof (necessity). Let $P_1(t) \ge 0\ \forall t \in [0,T]$ with $P_1(T) = 0$ solve

$$A_x^T P_1(t) + P_1(t)A_x + P_1(t)(B_1B_1^T/\gamma^2 - B_2R_1^{-1}B_2^T)P_1(t) + Q = -\dot{P}_1(t).$$

Suppose there are a w_*' and an optimal control u_* in the form of

$$\dot{\hat{x}} = (A + \gamma^{-2}B_1B_1^T P_1(t) + L_*(t)C_2 + B_2F_*(t))\hat{x} - L_*(t)y, \quad \hat{x}(0) = 0,$$
$$u_* = F_*(t)\hat{x},$$

where $F_*(t) = -R_1^{-1}(D_{12}^T C_1 + B_2^T P_1(t))$, achieving $0 < J_1(u_*, w'_*, 0) \le J_1(u_*, w, 0)$. This suggests that the system without the white noise,

$$\begin{aligned}
&\dot{x} = Ax + B_1 w + B_2 u_*, \quad x(0) = 0,\\
&z = C_1 x + D_{12} u_*,\\
&y = C_2 x + D_{20} w_0,
\end{aligned}$$

achieves the $\mathcal{H}_\infty$ performance on finite support $[0, T]$. Now define

$$e_x := x - \hat{x}, \quad r := w - \gamma^{-2} B_1^T P_1(t) x, \quad v_* := D_{12}\left\{u_* + R_1^{-1}(D_{12}^T C_1 + B_2^T P_1(t))x\right\}.$$

The system without the white noise can be converted into

$$\begin{aligned}
&\dot{e}_x = (A + \gamma^{-2} B_1 B_1^T P_1(t) + L_*(t) C_2) e_x + B_1 r, \quad e_x(0) = 0,\\
&v_* = D_{12}\left\{R_1^{-1}(D_{12}^T C_1 + B_2^T P_1(t)) e_x\right\},
\end{aligned}$$

and $J_1(u_*, w, 0)$ can be converted into

$$J_1(u_*, w, 0) = \gamma^2 \|w - \tilde{w}_*\|_{[0,T]}^2 - \|D_{12}(u_* - \tilde{u}_*)\|_{[0,T]}^2 = \gamma^2 \|r\|_{[0,T]}^2 - \|v_*\|_{[0,T]}^2,$$

where $\tilde{w}_* = \gamma^{-2} B_1^T P_1(t) x$ and $\tilde{u}_* = -R_1^{-1}(D_{12}^T C_1 + B_2^T P_1(t))x$. Hence, by Lemma 3, there is a $P_2(t) \ge 0\ \forall t \in [0, T]$ with $P_2(T) = 0$ solving

$$P_2(t)(A + \gamma^{-2} B_1 B_1^T P_1(t) + L_*(t) C_2) + (A_y + \gamma^{-2} B_1 B_1^T P_1(t) + L_*(t) C_2)^T P_2(t)$$

$$+\gamma^{-2} P_2(t) B_1 B_1^T P_2(t) + (D_{12}^T C_1 + B_2^T P_1(t))^T R_1^{-1}(D_{12}^T C_1 + B_2^T P_1(t)) = -\dot{P}_2(t),$$

and, clearly, w'_* can be obtained as

$$w'_* = r'_* + \gamma^{-2} B_1^T P_1(t) x = \gamma^{-2} B_1^T (P_1(t) x + P_2(t) e_x), \quad r'_* = \gamma^{-2} B_1^T P_2(t) e_x.$$

Consider the system with white noise

$$\begin{aligned}
&\dot{x} = Ax + B_0 w_0 + B_1 w + B_2 u_*, \quad x(0) = 0,\\
&z = C_1 x + D_{12} u_*,\\
&y = C_2 x + D_{20} w_0.
\end{aligned}$$

Under the same transformation, we have

$$\begin{aligned}
&\dot{e}_x = (A + \gamma^{-2} B_1 B_1^T P_1(t) + L_*(t) C_2) e_x + (B_0 + L_*(t) D_{20}) w_0 + B_1 r, \quad e_x(0) = 0,\\
&v_* = D_{12}\left\{R_1^{-1}(D_{12}^T C_1 + B_2^T P_1(t)) e_x\right\}.
\end{aligned}$$

$J_1(u_*, w, w_0)$ becomes (see the proof for the sufficiency)

$$J_1(u_*, w, w_0) = \gamma^2 \|w - \tilde{w}_*\|_{[0,T]}^2 - \|D_{12}(u_* - \tilde{u}_*)\|_{[0,T]}^2 - \int_0^T \operatorname{trace}\{B_0^T P_1(t) B_0\} dt$$

$$= \gamma^2 \|r - \gamma^{-2} B_1^T P_2(t) e_x\|_{[0,T]}^2 - \int_0^T \operatorname{trace}\{B_0^T P_1(t) B_0\} dt$$

$$-\int_0^T \text{trace}\{(B_0 + L_*(t)D_{20})^T P_2(t)(B_0 + L_*(t)D_{20})\}dt.$$

Choosing

$$r_* = \gamma^{-2}B_1^T P_2(t)e_x \quad \text{or} \quad w_* = r_* + \gamma^{-2}B_1^T P_1(t)x = \gamma^{-2}B_1^T(P_1(t)x + P_2(t)e_x),$$

it is easy to see that $J_1(u_*, w_*, w_0) \le J_1(u_*, w, w_0)$. If, furthermore, this w_* together with u_* achieves $J_2(u_*, w_*, w_0) \le J_2(u, w_*, w_0)$, by substituting w_* into the system equations, we get

$$\begin{aligned}\dot{e}_x &= (A + \gamma^{-2}B_1B_1^T P_1(t) + L_*(t)C_2 + \gamma^{-2}B_1B_1^T P_2(t))e_x + (B_0 + L_*(t)D_{20})w_0,\\ &:= A_{L_*}e_x + B_{L_*}w_0.\end{aligned}$$

Let $\Phi(t,\tau)$ be the transition matrix of A_{L_*}; then $e_x = \int_0^t \Phi^T(t,\tau)B_{L_*}w_0(\tau)d\tau$, and

$$J_2(u_*, w_*, w_0) = \|e_x\|_{[0,T]} = \text{trace}\left\{\int_0^T Y(t)dt\right\}$$

is the minimum value, where $Y(t) = \int_0^t \Phi(t,s)B_{L_*}B_{L_*}^T\Phi^T(t,s)ds \ge 0$, $Y(0) = 0$, satisfies $A_{L_*}Y(t) + Y(t)A_{L_*}^T + B_{L_*}B_{L_*}^T = \dot{Y}(t)$. Thus, by Theorem 6, there is a $P_3(t) \ge 0\ \forall t \in [0,T]$ with $P_3(0) = 0$ solving

$$[A_y + \gamma^{-2}B_1B_1^T(P_1(t) + P_2(t))]P_3(t) + P_3(t)[A_y + \gamma^{-2}B_1B_1^T(P_1(t) + P_2(t))]^T$$

$$-P_3(t)C_2^T R_0^{-1}C_2P_3(t) + P = \dot{P}_3(t),$$

and $L_*(t)$ can be chosen as $L_*(t) = -(B_0D_{20}^T + P_3(t)C_2^T)R_0^{-1}$. Substituting L_* back into the Riccati equation about $P_2(t)$, clearly, $P_2(t) \ge 0\ \forall t \in [0,T]$ with $P_2(T) = 0$ solves

$$P_2(t)(A_y+\gamma^{-2}B_1B_1^T P_1(t)-P_3(t)C_2^T R_0^{-1}C_2)+(A_y+\gamma^{-2}B_1B_1^T P_1(t)-P_3(t)C_2^T R_0^{-1}C_2)^T P_2(t)$$

$$+\gamma^{-2}P_2(t)B_1B_1^T P_2(t) + (D_{12}^TC_1 + B_2^T P_1(t))^T R_1^{-1}(D_{12}^TC_1 + B_2^T P_1(t)) = -\dot{P}_2(t). \qquad \square$$

5.2. $\mathcal{H}_\infty$ Gaussian control design—infinite time horizon. For $\mathcal{H}_\infty$ Gaussian design on infinite time horizon, we need to add the following standard assumptions:

(A1) (A, B_2) is stabilizable and (C_2, A) is detectable;

(A2) $\begin{bmatrix} A - j\omega I & B_2 \\ C_1 & D_{12} \end{bmatrix}$ has full column rank for all ω;

(A3) $\begin{bmatrix} A - j\omega I & B_0 \\ C_2 & D_{20} \end{bmatrix}$ has full row rank for all ω.

The infinite time $\mathcal{H}_\infty$ Gaussian control design is presented in the next theorem.

THEOREM 14. *Let the dynamical system G be described by* (5)–(7). *Suppose that w and w_0 are independent. If there are stabilizing solutions $P_1 \ge 0$, $P_2 \ge 0$, and $P_3 \ge 0$ solving the Riccati equations*

$$A_x^T P_1 + P_1A_x + P_1(B_1B_1^T/\gamma^2 - B_2R_1^{-1}B_2^T)P_1 + Q = 0,$$

$$P_2(A_y + \gamma^{-2}B_1B_1^T P_1 - P_3C_2^T R_0^{-1}C_2) + (A_y + \gamma^{-2}B_1B_1^T P_1 - P_3C_2^T R_0^{-1}C_2)^T P_2$$

$$+\gamma^{-2}P_2B_1B_1^TP_2+(D_{12}^TC_1+B_2^TP_1)^TR_1^{-1}(D_{12}^TC_1+B_2^TP_1)=0,$$

$$[A_y+\gamma^{-2}B_1B_1^T(P_1+P_2)]P_3+P_3[A_y+\gamma^{-2}B_1B_1^T(P_1+P_2)]^T$$

$$-P_3C_2^TR_0^{-1}C_2P_3+P=0,$$

i.e., if $A_x+(B_1B_1^T/\gamma^2-B_2R_1^{-1}B_2^T)P_1$ *and* $A_y+\gamma^{-2}B_1B_1^T(P_1+P_2)-P_3C_2^TR_0^{-1}C_2$ *are both stable, then there exist an optimal control law* u_* *and a worst disturbance signal* w_* *such that*

$$J_3(u_*,w_*,w_0)\le J_3(u_*,w,w_0),\quad J_4(u_*,w_*,w_0)\le J_4(u,w_*,w_0).$$

If the solutions exist, then $w_*=\gamma^{-2}B_1^T(P_1x+P_2e)$*, and an optimal controller is given by*

$$\begin{aligned}\dot{\hat{x}}&=(A+\gamma^{-2}B_1B_1^TP_1+L_*C_2+B_2F_*)\hat{x}-L_*y,\\ u_*&=F_*\hat{x},\quad \hat{x}(0)=0,\end{aligned}$$

where $F_*:=-R_1^{-1}(D_{12}^TC_1+B_2^TP_1)$ *and* $L_*=-(B_0D_{20}^T+P_3C_2^T)R_0^{-1}$.

Conversely, let $P_1\ge 0$ *be a stabilizing solution to*

$$A_x^TP_1+P_1A_x+P_1(B_1B_1^T/\gamma^2-B_2R_1^{-1}B_2^T)P_1+Q=0.$$

Suppose there exist a w_*' *and a controller* u_* *(hence, an* L_**)*

$$\begin{aligned}\dot{\hat{x}}&=(A+\gamma^{-2}B_1B_1^TP_1+L_*C_2+B_2F_*)\hat{x}-L_*y,\\ u_*&=F_*\hat{x},\quad F_*=-R_1^{-1}(D_{12}^TC_1+B_2^TP_1),\end{aligned}$$

achieving $0<J_3(u_*,w_*',0)\le J_3(u_*,w,0)$*. Then there exists a* w_* *achieving* $J_3(u_*,w_*,w_0)\le J_3(u_*,w,w_0)$*. If, furthermore, this* w_* *also achieves* $J_4(u_*,w_*,w_0)\le J_4(u,w_*,w_0)$*, then there exist* $P_2\ge 0$ *and* $P_3\ge 0$ *solving*

$$P_2(A_y+\gamma^{-2}B_1B_1^TP_1-P_3C_2^TR_0^{-1}C_2)+(A_y+\gamma^{-2}B_1B_1^TP_1-P_3C_2^TR_0^{-1}C_2)^TP_2$$

$$+\gamma^{-2}P_2B_1B_1^TP_2+(D_{12}^TC_1+B_2^TP_1)^TR_1^{-1}(D_{12}^TC_1+B_2^TP_1)=0,$$

$$[A_y+\gamma^{-2}B_1B_1^T(P_1+P_2)]P_3+P_3[A_y+\gamma^{-2}B_1B_1^T(P_1+P_2)]^T-P_3C_2^TR_0^{-1}C_2P_3+P=0.$$

Moreover, if $A+\gamma^{-2}B_1B_1^T(P_1+P_2)-(B_0D_{20}^T+P_3C_2^T)R_0^{-1}C_2$ *is stable, then* $L_*=-(B_0D_{20}^T+P_3C_2^T)R_0^{-1}$.

Proof (sufficiency). Suppose that there are $P_1\ge 0$, $P_2\ge 0$, and $P_3\ge 0$ solving

$$A_x^TP_1+P_1A_x+P_1(B_1B_1^T/\gamma^2-B_2R_1^{-1}B_2^T)P_1+Q=0,$$

$$P_2(A_y+\gamma^{-2}B_1B_1^TP_1-P_3C_2^TR_0^{-1}C_2)+(A_y+\gamma^{-2}B_1B_1^TP_1-P_3C_2^TR_0^{-1}C_2)^TP_2$$

$$+\gamma^{-2}P_2B_1B_1^TP_2+(D_{12}^TC_1+B_2^TP_1)^TR_1^{-1}(D_{12}^TC_1+B_2^TP_1)=0,$$

$$[A_y+\gamma^{-2}B_1B_1^T(P_1+P_2)]P_3+P_3[A_y+\gamma^{-2}B_1B_1^T(P_1+P_2)]^T-P_3C_2^TR_0^{-1}C_2P_3+P=0.$$

First, it is claimed that $A_y+\gamma^{-2}B_1B_1^TP_1-P_3C_2^TR_0^{-1}C_2=A+\gamma^{-2}B_1B_1^TP_1+L_*C_2$ is stable, where $L_*=-(B_0D_{20}^T+P_3C_2^T)R_0^{-1}$. The reason is as follows: if $A+\gamma^{-2}B_1B_1^TP_1+L_*C_2$ is not stable, then at least one of its eigenvalues λ is on the closed right-half plane, i.e., $\mathrm{Re}(\lambda)\geq 0$. Let x be the eigenvector corresponding to λ; then

$$x^TP_2(A_y+\gamma^{-2}B_1B_1^TP_1-P_3C_2^TR_0^{-1}C_2)x+x^T(A_y+\gamma^{-2}B_1B_1^TP_1-P_3C_2^TR_0^{-1}C_2)^TP_2x$$

$$\gamma^{-2}x^TP_2B_1B_1^TP_2x+x^T(D_{12}^TC_1+B_2^TP_1)^TR_1^{-1}(D_{12}^TC_1+B_2^TP_1)x=0$$

or

$$2\mathrm{Re}(\lambda)x^TP_2x+x^T(D_{12}^TC_1+B_2^TP_1)^TR_1^{-1}(D_{12}^TC_1+B_2^TP_1)x+\gamma^{-2}x^TP_2B_1B_1^TP_2x=0,$$

which gives $B_1^TP_2x=0$ and $(D_{12}^TC_1+B_2^TP_1)x=0$. Thus

$$[A_y+\gamma^{-2}B_1B_1^T(P_1+P_2)-P_3C_2^TR_0^{-1}C_2]x=[A+\gamma^{-2}B_1B_1^TP_1+L_*C_2]x=\lambda x,$$

which means that $A_y+\gamma^{-2}B_1B_1^T(P_1+P_2)-P_3C_2^TR_0^{-1}C_2$ is not stable, which is a contradiction.

Now consider the index $J_3(u,w,w_0)$. Let u be any stabilizing control law. Define $r:=w-\gamma^{-2}B_1^TP_1x$ and $v:=D_{12}\left\{u+R_1^{-1}(D_{12}^TC_1+B_2^TP_1)x\right\}$. Then the system equations can be rewritten as

$$\begin{aligned}\dot{x}&=(A+\gamma^{-2}B_1B_1^TP_1)x+B_0w_0+B_1r+B_2u,\\ v&=D_{12}\left\{R_1^{-1}(D_{12}^TC_1+B_2^TP_1)x+u\right\},\\ y&=C_2x+D_{20}w_0,\end{aligned}$$

and the performance index $J_3(u,w,w_0)$ becomes

$$J_3(u,w,w_0)=\gamma^2\|w\|_{\mathcal{P}}^2-\|z\|_{\mathcal{P}}^2=\gamma^2\|r\|_{\mathcal{P}}^2-\|v\|_{\mathcal{P}}^2-\mathrm{trace}\{B_0^TP_1B_0\}.$$

Note that the first Riccati equation and Lemma 4 are used to derive this equation.

We can use L_* to construct a state estimator:

$$\dot{\hat{x}}=(A+\gamma^{-2}B_1B_1^TP_1)\hat{x}+B_2u+L_*(C_2\hat{x}-y),\quad \hat{x}(0)=0.$$

It is pointed out that the control law $u_*=-R_1^{-1}(D_{12}^TC_1+B_2^TP_1)\hat{x}$ comes naturally when the state information is not available. Let $e_x=x-\hat{x}$. The system can be further simplified into

$$\begin{aligned}\dot{e}_x&=(A+\gamma^{-2}B_1B_1^TP_1+L_*C_2)e_x+(B_0+L_*D_{20})w_0+B_1r,\\ v&=D_{12}R_1^{-1}(D_{12}^TC_1+B_2^TP_1)e_x.\end{aligned}$$

Now $J_3(u,w,w_0)$ becomes (by using the second Riccati equation and Lemma 5)

$$\begin{aligned}J_3(u_*,w,w_0)=\gamma^2\|r-\gamma^{-2}B_1^TP_2e_x\|_{\mathcal{P}}^2&-\mathrm{trace}\{B_0^TP_1B_0\}\\ &-\mathrm{trace}\{(B_0+L_*D_{20})^TP_2(B_0+L_*D_{20})\}.\end{aligned}$$

Similar to finite time horizon case, we can take

$$r_* = \gamma^{-2} B_1^T P_2 e_x \quad \text{or} \quad w_* = r_* + \gamma^{-2} B_1^T P_1 x = \gamma^{-2} B_1^T (P_1 x + P_2 e_x).$$

Then we have $J_3(u_*, w_*, w_0) \le J_3(u_*, w, w_0)$.

Next, it is shown that u_* does minimize the index $J_4(u, w_*, w_0)$. Let L be any filter gain such that both $A + \gamma^{-2} B_1 B_1^T P_1 + LC_2$ and $A + \gamma^{-2} B_1 B_1^T P_1 + LC_2 + \gamma^{-2} B_1 B_1^T P_2$ are stable. Substituting w_* (or r_*) into the system equations, we get $\dot{e}_x = A_L e_x + B_L w_0$, where

$$A_L = A + \gamma^{-2} B_1 B_1^T P_1 + LC_2 + \gamma^{-2} B_1 B_1^T P_2, \quad B_L = B_0 + LD_{20}.$$

Note that $e_x = \int_0^t e^{A_L(t-\tau)} B_L w_0(\tau) d\tau$ and

$$\begin{aligned} J_4(u, w_*, w_0) &= \|e_x\|_{\mathcal{P}}^2 \\ &= \lim_{T\to\infty} E \left\{ \frac{1}{T} \int_0^T \int_0^t \int_0^t w_0^T(\tau) B_L^T e^{A_L^T(t-\tau)} e^{A_L(t-s)} B_L w_0(s) d\tau \, ds \, dt \right\} \\ &= \operatorname{trace} \left\{ \lim_{T\to\infty} \frac{1}{T} \int_0^T \int_0^t \int_0^t e^{A_L(t-s)} B_L \delta(\tau - s) B_L^T e^{A_L^T(t-\tau)} d\tau \, ds \, dt \right\} \\ &= \operatorname{trace} \left\{ \lim_{T\to\infty} \frac{1}{T} \int_0^T \int_0^t e^{A_L(t-s)} B_L B_L^T e^{A_L^T(t-s)} ds \, dt \right\} = \operatorname{trace}\{Y\}, \end{aligned}$$

where $Y = \int_0^\infty e^{A_L s} B_L B_L^T e^{A_L^T s} ds \ge 0$ satisfies $A_L Y + Y A_L^T + B_L B_L^T = 0$. By Theorem 7 and using the third Riccati equation, it can be seen that $J_4(u, w_*, w_0)$ achieves the minimum value at $L = L_*$, where $L_* = -(B_0 D_{20}^T + P_3 C_2^T) R_0^{-1}$, which means that u_* is indeed the desired optimal control. Thus u_* and w_* achieve

$$J_3(u_*, w_*, w_0) \le J_3(u_*, w, w_0), \quad J_4(u_*, w_*, w_0) \le J_4(u, w_*, w_0).$$

Proof (necessity). Let $P_1 \ge 0$ solve

$$A_x^T P_1 + P_1 A_x + P_1 (B_1 B_1^T / \gamma^2 - B_2 R_1^{-1} B_2^T) P_1 + Q = 0.$$

Suppose the controller u_*,

$$\begin{aligned} \dot{\hat{x}} &= (A + \gamma^{-2} B_1 B_1^T P_1 + L_* C_2 + B_2 F_*) \hat{x} - L_* y, \\ u_* &= F_* \hat{x}, \quad F_* = -R_1^{-1} (D_{12}^T C_1 + B_2^T P_1), \end{aligned}$$

and a w_*' achieve $0 < J_3(u_*, w_*', 0) \le J_3(u_*, w, 0)$. This suggests that the system without the white noise

$$\begin{aligned} \dot{x} &= Ax + B_1 w + B_2 u_*, \quad x(0) = 0, \\ z &= C_1 x + D_{12} u_*, \\ y &= C_2 x + D_{20} w_0, \end{aligned}$$

achieves the $\mathcal{H}_\infty$ performance. Define

$$e_x := x - \hat{x}, \quad r := w - \gamma^{-2} B_1^T P_1 x, \quad v_* := D_{12} \left\{ u_* + R_1^{-1} (D_{12}^T C_1 + B_2^T P_1) x \right\}.$$

The system without the white noise can be converted into

$$\begin{aligned}\dot{e}_x &= (A+\gamma^{-2}B_1B_1^TP_1+L_*C_2)e_x+B_1r, \quad e_x(0)=0,\\ v_* &= D_{12}\left\{R_1^{-1}(D_{12}^TC_1+B_2^TP_1)e_x\right\},\end{aligned}$$

and $J_3(u_*,w,0)$ can be converted into

$$J_3(u_*,w,0)=\gamma^2\|w-\tilde{w}_*\|_{\mathcal{P}}^2-\|D_{12}(u_*-\tilde{u}_*)\|_{\mathcal{P}}^2=\gamma^2\|r\|_{\mathcal{P}}^2-\|v_*\|_{\mathcal{P}}^2,$$

where $\tilde{w}_*=\gamma^{-2}B_1^TP_1x$ and $\tilde{u}_*=-R_1^{-1}(D_{12}^TC_1+B_2^TP_1)x$. Hence, by the bounded real lemma, there is a $P_2\geq 0$ solving

$$P_2(A+\gamma^{-2}B_1B_1^TP_1+L_*C_2)+(A_y+\gamma^{-2}B_1B_1^TP_1+L_*C_2)^TP_2$$

$$+\gamma^{-2}P_2B_1B_1^TP_2+(D_{12}^TC_1+B_2^TP_1)^TR_1^{-1}(D_{12}^TC_1+B_2^TP_1)=0,$$

and, accordingly, w_*' can be obtained as

$$w_*'=r_*'+\gamma^{-2}B_1^TP_1x=\gamma^{-2}B_1^T(P_1x+P_2e_x), \quad r_*'=\gamma^{-2}B_1^TP_2e_x.$$

Now consider the system with white noise

$$\begin{aligned}\dot{x} &= Ax+B_0w_0+B_1w+B_2u_*, \quad x(0)=0,\\ z &= C_1x+D_{12}u_*,\\ y &= C_2x+D_{20}w_0.\end{aligned}$$

Under the same transformation, we have

$$\begin{aligned}\dot{e}_x &= (A+\gamma^{-2}B_1B_1^TP_1+L_*C_2)e_x+(B_0+L_*D_{20})w_0+B_1r, \quad e_x(0)=0,\\ v_* &= D_{12}\left\{R_1^{-1}(D_{12}^TC_1+B_2^TP_1)e_x\right\},\end{aligned}$$

and $J_3(u_*,w,w_0)$ becomes (see the proof for the sufficiency)

$$J_3(u_*,w,w_0)=\gamma^2\|w-\tilde{w}_*\|_{\mathcal{P}}^2-\|D_{12}(u_*-\tilde{u}_*)\|_{\mathcal{P}}^2-\operatorname{trace}\{B_0^TP_1B_0\}$$

$$=\gamma^2\|r-\gamma^{-2}B_1^TP_2e_x\|_{\mathcal{P}}^2-\operatorname{trace}\{B_0^TP_1B_0\}-\operatorname{trace}\{(B_0+L_*D_{20})^TP_2(B_0+L_*D_{20})\}.$$

Therefore, if we choose

$$w_*=r_*+\gamma^{-2}B_1^TP_1x=\gamma^{-2}B_1^T(P_1x+P_2e_x), \quad r_*=\gamma^{-2}B_1^TP_2e_x,$$

then we have $J_3(u_*,w_*,w_0)\leq J_3(u_*,w,w_0)$.

If, furthermore, this w_* together with u_* achieves $J_4(u_*,w_*,w_0)\leq J_4(u,w_*,w_0)$, by substituting w_* into the system equations, we get

$$\begin{aligned}\dot{e}_x &= (A+\gamma^{-2}B_1B_1^TP_1+L_*C_2+\gamma^{-2}B_1B_1^TP_2)e_x+(B_0+L_*D_{20})w_0,\\ &:= A_{L_*}e_x+B_{L_*}w_0,\end{aligned}$$

where $A_{L_*}=A+\gamma^{-2}B_1B_1^TP_1+L_*C_2+\gamma^{-2}B_1B_1^TP_2$ and $B_{L_*}=B_0+L_*D_{20}$. So $e_x=\int_0^t e^{A_{L_*}(t-\tau)}B_{L_*}w_0(\tau)d\tau$ and $J_4(u_*,w_*,w_0)=\operatorname{trace}\{Y\}$ is the minimum value,

where $Y = \int_0^\infty e^{A_{L_*}s} B_{L_*} B_{L_*}^T e^{A_{L_*}^T s} ds \geq 0$ satisfies $A_{L_*} Y + Y A_{L_*}^T + B_{L_*} B_{L_*}^T = 0$. Thus, by Theorem 7, there are $P_3 \geq 0$ and $P_3 \leq Y$ solving

$$[A_y + \gamma^{-2} B_1 B_1^T (P_1 + P_2)] P_3 + P_3 [A_y + \gamma^{-2} B_1 B_1^T (P_1 + P_2)]^T - P_3 C_2^T R_0^{-1} C_2 P_3 + P = 0.$$

In the case in which $A + \gamma^{-2} B_1 B_1^T (P_1 + P_2) - (B_0 D_{20}^T + P_3 C_2^T) R_0^{-1} C_2$ is stable, L_* can be chosen as $L_* = -(B_0 D_{20}^T + P_3 C_2^T) R_0^{-1}$. Substituting L_* back into the Riccati equation about P_2, clearly, P_2 solves

$$P_2 (A_y + \gamma^{-2} B_1 B_1^T P_1 - P_3 C_2^T R_0^{-1} C_2) + (A_y + \gamma^{-2} B_1 B_1^T P_1 - P_3 C_2^T R_0^{-1} C_2)^T P_2$$

$$+ \gamma^{-2} P_2 B_1 B_1^T P_2 + (D_{12}^T C_1 + B_2^T P_1)^T R_1^{-1} (D_{12}^T C_1 + B_2^T P_1) = 0.$$

This concludes the proof. □

6. Comments. We have the following comments on the results in this paper.

1. It must be pointed out that, in either of our developments, while we can achieve quadratic optimization ($\mathcal{H}_2$ performance) under the worst disturbance, the optimal γ may or may not be achievable. As seen from the results, we can only show that the $\mathcal{H}_\infty$ performance index has a lower bound, but this bound may or may not be the optimal γ as achieved by a pure $\mathcal{H}_\infty$ control. This means that, naturally and reasonably, we should not expect a multiobjective control law to be as robust as an $\mathcal{H}_\infty$ control law. Similarly, we can only expect that a multiobjective control law achieves the quadratic optimal performance when the disturbance signal is in its "worst" form. And all of these facts exactly reflect the design trade-off.

2. For the solvability of the solutions in this paper, we point out that, although our results are characterized by three Riccati equations, only two of them are really coupled, while the third one can be solved independently. Hence the computation of our results should be no more difficult than that of the state feedback case, where two coupled Riccati equations can be solved by standard numerical integration, as pointed out in [22].

7. Conclusion. In this paper, we have generalized the state feedback results of mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control [22] to the output feedback case. We have also solved a newly formulated $\mathcal{H}_\infty$ Gaussian control problem which provides a natural design trade-off between the LQG performance and the robust performance. Both sufficient and necessary conditions were given for the existence of the control law in both the finite time horizon and the infinite time horizon. Since these conditions are in the form of coupled Riccati equations, differential or algebraic, which are deemed as solvable by standard algorithm, the design procedures developed in this paper provide computable solutions, which strongly enhances the possibility of potential engineering applications.

Appendix: Proof for necessary conditions in Theorems 6 and 7. Note that we shall only give the proof in detail for necessary conditions in Theorem 6 because the proof for Theorem 7 can be done in very similar way.

We shall first establish some preliminary results for the proof.

Let $\{P_i(t), \quad i = 1, 2, \ldots, P_i(t) = P_i^T(t), t \geq 0\}$ be a sequence in $\mathbf{R^{n \times n}}$. Correspondingly, we define a sequence $\{L_i(t),\ i = 2, 3 \ldots, t \geq 0\}$ in $\mathbf{R^{n \times p}}$ with $L_{i+1}(t) = -(P_i(t) C^T + B D^T) R^{-1}\ \forall t \geq 0$, for some $R > 0$. The limits of $\{P_i(t)\}$ and $\{L_i(t)\}$ are defined as follows.

DEFINITION 15. *We say that $P_*(t)$ and $L_*(t)$ are the limits of sequences $\{P_i(t)\}$ and $\{L_i(t)\}$ if*

$$x^T P_*(t)x = \lim_{i\to\infty} x^T P_i(t)x, \quad L_*(t) = -(P_*(t)C^T + BD^T)R^{-1} \ \ \forall x \in \mathbf{R}^n, \forall t \geq 0.$$

If these limits exist, we denote

$$P_*(t) = \lim_{i\to\infty} P_i(t), \quad L_*(t) = \lim_{i\to\infty} L_{i+1}(t) = -\lim_{i\to\infty}(P_i(t)C^T + BD^T)R^{-1}.$$

It is easy to see that $L_i(t)$ has a limit if $P_i(t)$ does.

PROPOSITION 16. *A sequence $\{P_i(t)\}$ converges to some $P_*(t)$ if and only if the convergence is entrywise, i.e., if $p^i_{kj}(t)$ and $p_{kj*}(t)$ are entries of $P_i(t)$ and $P_*(t)$, then*

$$p_{kj*}(t) = \lim_{i\to\infty} p^i_{kj}(t), \qquad k, j = 1, 2, \ldots, n \qquad \forall t \geq 0.$$

Proof. If the convergence is entrywise, i.e.,

$$p_{kj*}(t) = \lim_{i\to\infty} p^i_{kj}(t), \quad k, j = 1, 2, \ldots, n,$$

then we have $\forall x \in \mathbf{R}^n$ and $\forall t \geq 0$

$$\lim_{i\to\infty} x^T P_i(t)x = \lim_{i\to\infty} \sum_{k,j} p^i_{kj}(t)x_k x_j = \sum_{k,j} \lim_{i\to\infty} p^i_{kj}(t)x_k x_j = \sum_{k,j} p_{kj*}(t)x_k x_j = x^T P_*(t)x.$$

So $P_*(t) = \lim_{i\to\infty} P_i(t)$. Conversely, if $P_i(t)$ converges to $P_*(t)$ for any $t \geq 0$, i.e., $x^T P_*(t)x = \lim_{i\to\infty} x^T P_i(t)x \ \forall x \in \mathbf{R}^n, \ \forall t \geq 0$, or

$$\sum_{j,q} p_{jq*}(t)x_j x_q = \lim_{i\to\infty} \sum_{j,q} p^i_{jq}(t)x_j x_q = \sum_{j,q} \lim_{i\to\infty} p^i_{jq}(t)x_j x_q.$$

Comparing coefficients on both sides (considering x is arbitrary), we obtain

$$p_{jq*}(t) = \lim_{i\to\infty} p^i_{jq}(t).$$

That is, $P_i(t)$ converges to $P_*(t)$ entrywise. □

We are interested in a pair of special sequences $\{P_i(t)\}$ and $\{L_i(t)\}$, which are generated by the following procedures.

Procedures.

1. Choose $L_1(t) \in \mathbf{R}^{n\times p}$ for any $t \geq 0$.
2. Solve $P_i(t), i = 1, 2, \ldots, \ P_i(0) = 0$ from

$$(A + L_i(t)C)P_i(t) + P_i(t)(A + L_i(t)C)^T + (B + L_i(t)D)(B + L_i(t)D)^T = \dot{P}_i(t).$$

3. Set $L_{i+1}(t) = -(P_i(t)C^T + BD^T)R^{-1}$, $i = 1, 2, \ldots$, for some $R > 0$.

PROPOSITION 17. *Sequences $P_i(t)$ and $L_i(t)$ generated by the above procedures* 1–3 *always have limits $P_*(t)$ and $L_*(t)$.*

Proof. We need only to prove that $P_i(t)$ has a limit $P_*(t)$. Note that we have, for $i = 1, 2, \ldots,$

$$(A + L_i(t)C)P_i(t) + P_i(t)(A + L_i(t)C)^T + (B + L_i(t)D)(B + L_i(t)D)^T = \dot{P}_i(t),$$

$$(A+L_{i+1}(t)C)P_{i+1}+P_{i+1}(A+L_{i+1}(t)C)^T+(B+L_{i+1}(t)D)(B+L_{i+1}(t)D)^T=\dot{P}_{i+1}(t).$$

Define $\Delta P_i(t)=P_{i+1}-P_i(t)$ and $\Delta L_i(t)=L_{i+1}(t)-L_i(t)$; then

$$(A+L_{i+1}(t)C)\Delta P_i(t)+\Delta P_i(t)(A+L_{i+1}(t)C)^T-\Delta L_i(t)R\Delta L_i^T(t)=\Delta\dot{P}_i(t).$$

Let $\Phi(t,\tau)$ be the transition matrix of $A+L_{i+1}(t)C$; then we have

$$\Delta P_i(t)=-\int_0^t \Phi(t,s)\Delta L_i(s)R\Delta L_i^T(s)\Phi^T(t,s)ds,$$

which gives that $\Delta P_i(t)\le 0\ \forall t\ge 0$. This means that for any $x\in\mathbf{R}^n$

$$0\le\cdots\le x^TP_{i+1}(t)x\le x^TP_i(t)x\le\cdots\le x^TP_1(t)x\ \ \forall t\ge 0.$$

Hence $\lim_{i\to\infty}x^TP_i(t)x$ exists and

$$\begin{aligned}\lim_{i\to\infty}x^TP_i(t)x&=\lim_{i\to\infty}\sum_{k,j}p_{kj}^i(t)(t)x_kx_j=\sum_{k,j}\lim_{i\to\infty}p_{kj}^i(t)x_kx_j\\&=\sum_{k,j}p_{kj*}(t)x_kx_j=x^TP_*(t)x,\end{aligned}$$

where $p_{kj*}(t)=\lim_{i\to\infty}p_{kj}^i(t)\ \forall t\ge 0$ and $P_*(t)=[p_{kj*}(t)]$. Therefore, $P_i(t)$ has a limit and so does $L_i(t)$ with

$$L_*(t)=\lim_{i\to\infty}L_{i+1}(t)=-\lim_{i\to\infty}(P_i(t)C^T+BD^T)R^{-1}=-(P_*(t)C^T+BD^T)R^{-1}.\qquad\square$$

LEMMA 18. *For sequences $P_i(t)$ and $L_i(t)$ generated by the procedures* 1–3, *if $P_*(t)$ and $L_*(t)$ are the limit points of these sequences, then $P_*(t)\ge 0$ solves*

$$(A+L_*(t)C)P_*(t)+P_*(t)(A+L_*(t)C)^T+(B+L_*(t)D)(B+L_*(t)D)^T=\dot{P}_*(t),$$

where $L_*(t)=-(P_*(t)C^T+BD^T)R^{-1}$.

Proof. Suppose $P_*(t)$ and $L_*(t)$ are the limit points of sequences $P_i(t)$ and $L_i(t)$. Let $p_{kj}^i(t)$ and $p_{kj*}(t)$ be entries of $P_i(t)$ and $P_*(t)$. Let $l_{mq}^i(t)$ and $l_{mq*}(t)$ be entries of $L_i(t)$ and $L_*(t)$. By Proposition 16, we have entrywise convergence $p_{kj*}(t)=\lim_{i\to\infty}p_{kj}^i(t),\ k,j=1,2,\dots,n$, and, consequently, for any $m=1,\dots,n$, and $q=1,\dots,p$,

$$l_{mq*}(t)=\lim_{i\to\infty}l_{mq}^i(t)(p_{kj}^i(t),\ k,j=1,2,\dots,n)=l_{mq}^i(t)(\lim_{i\to\infty}p_{kj}^i(t),\ k,j=1,2,\dots,n)$$

since $l_{mq}^i(t)$ is a continuous function of $p_{kj}^i(t),\ k,j=1,2,\dots,n$.

Next we define

$$F(P_i(t),L_i(t))=\dot{P}_i(t)-(A+L_i(t)C)P_i(t)+P_i(t)(A+L_i(t)C)^T+(B+L_i(t)D)(B+L_i(t)D)^T.$$

Obviously, $F(P_i(t),L_i(t))=0\ \forall i=1,2,\dots,\forall t\ge 0$. Let $f_{kj}^i(t),\ k,j=1,\dots,n$ be entries of $F(P_i(t),L_i(t))$; then they are continuous about all $p_{kj}^i(t)$, $\dot{p}_{kj}^i(t)$, and $l_{mq}^i(t)$. Therefore,

$$f_{kj*}(t)=\lim_{i\to\infty}f_{kj}^i(t)(p_{kj}^i(t),l_{mq}^i(t))=0,\quad k,j=1,\dots,n,\ \ \forall t\ge 0.$$

This shows that $F(P_*(t), L_*(t)) = 0$ or

$$(A + L_*(t)C)P_*(t) + P_*(t)(A + L_*(t)C)^T + (B + L_*(t)D)(B + L_*(t)D)^T = \dot{P}_*(t),$$

where $L_*(t) = -(P_*(t)C^T + BD^T)R^{-1}$. □

Now we are in the position to prove the necessity of Theorem 6.

Proof. If there are $L(t)$ and a $P(t), P(0) = 0$, such that

$$(A + L(t)C)P(t) + P(t)(A + L(t)C)^T + (B + L(t)D)(B + L(t)D)^T = \dot{P}(t),$$

and J_1 achieves the minimum value at $L(t)$, take $L_1(t) = L(t)$ as the initial value and generate the sequences $P_i(t)$ and $L_i(t)$ using the procedures 1–3. Then the following claims can be made (see Proposition 17):

1. $0 \leq \cdots \leq P_{i+1}(t) \leq P_i(t) \leq \cdots \leq P_1(t)$.
2. $\{P_i(t),\ i = 1, 2, \ldots,\}$ and $\{L_i(t),\ i = 1, 2, \ldots,\}$ have limit points $P_*(t)$ and $L_*(t)$ and $P_*(t) \leq P_1(t)\ \forall t \geq 0$.

Hence, by Lemma 18, $P_*(t)$ and $L_*(t) = -(P_*(t)C^T + BD^T)R^{-1}$ solve

$$(A + L_*(t)C)P_*(t) + P_*(t)(A + L_*(t)C)^T + (B + L_*(t)D)(B + L_*(t)D)^T = \dot{P}_*(t),$$

and, clearly, $J_1(L)$ achieves the minimum value at $L_*(t) = -(P_*(t)C^T + BD^T)R^{-1}$. □

REFERENCES

[1] B. D. O. Anderson and J. B. Moore (1989), *Optimal Control: Linear Quadratic Methods*, Prentice Hall, Englewood Cliffs, NJ.

[2] T. Başar and G. J. Olsder (1982), *Dynamic Non-Cooperative Game Theory*, Academic Press, New York.

[3] D. S. Bernstein and W. M. Haddad (1989), *LQG control with an $\mathcal{H}_\infty$ performance bound: A Riccati equation approach*, IEEE Trans. Automat. Control, 34, pp. 293–305.

[4] M. Chilali, P. Gahinet, and C. Scherer (1996), *Multiobjective output-feedback control via LMI optimization*, in Proceedings of the IFAC World Congress, Vol. D, San Francisco, CA, pp. 249–254.

[5] X. Chen (1998), *Multiobjective Optimal Filtering and Control*, Ph.D. dissertation, Louisiana State University, Baton Rouge, LA.

[6] X. Chen and K. Zhou (1998), *$\mathcal{H}_\infty$ Gaussian control design*, in Proceedings of the 37th IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, pp. 638–643.

[7] X. Chen, G. Saloman, and K. Zhou (1998), *Multiobjective output feedback control*, in Proceedings of the 37th IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, pp. 1810–1811.

[8] R. D'Andrea (1996), *LMI approach to mixed H_2 and H_∞ performance objective controller design*, in Proceedings of the IFAC World Congress, Vol.G, San Francisco, CA, pp. 327–332.

[9] J. C. Doyle (1978), *Guaranteed margins for LQG regulators*, IEEE Trans. Automat. Control, 23, pp. 756–757.

[10] J. C. Doyle, K. Glover, P. P. Khargonekar, and B. A. Francis (1989), *State-space solutions to standard $\mathcal{H}_2$ and $\mathcal{H}_\infty$ control problems*, IEEE Trans. Automat. Control, 34, pp. 831–847.

[11] J. C. Doyle, K. Zhou, K. Glover, and B. Bodenheimer (1994), *Mixed H_2 and H_∞ performance objectives* II: *Optimal control*, IEEE Trans. Automat. Control, 39, pp. 1575–1587.

[12] Y. Fujisaki and T. Yoshida (1996), *A linear matrix inequality approach to mixed H_2/H_∞ control*, in Proceedings of the IFAC World Congress, Vol. C, San Francisco, CA, pp. 361–366.

[13] W. A. Gardner (1988), *Statistical Spectral Analysis: A Nonprobabilistic Theory*, Prentice Hall, Englewood Cliffs, NJ.

[14] J. C. Geromel, P. L. D. Peres, and S. R. Souza (1992), *Mixed H_2/H_∞ control for continuous-time linear system*, in Proceedings of the IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, pp. 3717–3722.

[15] K. Glover and D. Mustafa (1989), *Derivation of the maximum entropy $\mathcal{H}_\infty$ controller and a state-space formula for its entropy*, Internat. J. Control, 50, pp. 899–916.

[16] M. Green and D. J. N. Limebeer (1995), *Linear Robust Control*, Prentice Hall, Englewood Cliffs, NJ.

[17] W. M. Haddad, D. S. Bernstein, and D. Mustafa (1991), *Mixed-norm H_2/H_∞ regulation and estimation: The discrete-time case*, Systems Control Lett., 16, pp. 235–247.

[18] W. M. Haddad and D. S. Bernstein (1990), *Generalized Riccati equations for the full and reduced-order mixed-norm H_2/H_∞ standard problem*, Systems Control Lett., 14, pp. 185–197.

[19] G. D. Halikias (1994), *The hierarchical $\mathcal{H}_\infty/\mathcal{H}_2$-optimal control problem*, in Proceedings of the IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, pp. 3165–3166.

[20] P. P. Khargonekar and M. A. Rotea (1991), *Mixed H_2/H_∞ control: A convex optimization approach*, IEEE Trans. Automat. Control, 36, pp. 824–837.

[21] D. J. N. Limebeer, B. D. O. Anderson, P. P. Khargonekar, and M. Green (1992), *A game theoretic approach to $\mathcal{H}_\infty$ control for time varying systems*, SIAM J. Control Optim., 30, pp. 262–283.

[22] D. J. N. Limebeer, B. D. O. Anderson, and B. Hendel (1994), *A Nash game approach to mixed H_2/H_∞ control*, IEEE Trans. Automat. Control, 39, pp. 687–692.

[23] W. Lin (1995), *Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control of nonlinear systems*, in Proceedings of the IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, pp. 333–338.

[24] D. Mustafa (1989), *Relations between maximum entropy/$\mathcal{H}_\infty$ control and combined $\mathcal{H}_\infty/LQG$ control*, Systems Control Lett., 12, pp. 193–203.

[25] D. Mustafa and K. Glover (1990), *Minimum Entropy $\mathcal{H}_\infty$ Control*, Lecture Notes in Control and Inform. Sci. 146, Springer-Verlag, Berlin.

[26] M. A. Rotea and P. P. Khargonekar (1991), *H_2-optimal control with and H_∞ constraint: The state feedback case*, Automatica J. IFAC, 27, pp. 307–316.

[27] C. Scherer, P. Gahinet, and M. Chilali (1997), *Multiobjective output-feedback control via LMI optimization*, IEEE Trans. Automat. Control, 42, pp. 896–910.

[28] G. D. Sweriduk and A. J. Calise (1997), *Differential game approach to the mixed $H_2 - H_\infty$ problem*, J. Guidance, Control, and Dynamics, 20, pp. 1229–1234.

[29] M. Steinbuch and O. H. Bosgra (1994), *Robust performance H_2/H_∞ optimal control*, in Proceedings of the IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, pp. 3167–3172.

[30] A. A. Stoorvogel (1993), *The robust H_2 control problem: A worst-case design*, IEEE Trans. Automat. Control, 38, pp. 1358–1370.

[31] G. Tadmor (1990), *Worst-case design in the time domain: The maximum principle and the standard $\mathcal{H}_\infty$ problem*, Math. Control Signals Systems, 3, pp. 301–324.

[32] B. Vroemen and B. de Jager (1997), *Multiobjective control: An overview*, in Proceedings of the 36th IEEE Conference on Decision and Control, San Diego, CA, pp. 440–445.

[33] E. Wong and B. Hajek (1985), *Stochastic Processes in Engineering Systems*, 2nd ed., Springer-Verlag, New York.

[34] H. Yeh, S. Banda, and B. Chang (1992), *Necessary and sufficient conditions for mixed $\mathcal{H}_2$ and $\mathcal{H}_\infty$ control*, IEEE Trans. Automat. Control, 37, pp. 355–358.

[35] K. Zhou, J. C. Doyle, and K. Glover (1996), *Robust and Optimal Control*, Prentice Hall, Upper Saddle River, NJ.

[36] K. Zhou, K. Glover, B. Bodenheimer, and J. Doyle (1994), *Mixed $\mathcal{H}_2$ and $\mathcal{H}_\infty$ performance objectives* I: *Robust performance analysis*, IEEE Trans. Automat. Control, 39, pp. 1564–1574.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 3, pp. 661–680

STABILITY OF PERTURBED DELAY DIFFERENTIAL EQUATIONS AND STABILIZATION OF NONLINEAR CASCADE SYSTEMS*

W. MICHIELS†, R. SEPULCHRE‡, AND D. ROOSE†

Abstract. In this paper the effect of bounded input perturbations on the stability of nonlinear globally asymptotically stable delay differential equations is analyzed. We investigate under which conditions global stability is preserved and if not, whether semiglobal stabilization is possible by controlling the size or shape of the perturbation. These results are used to study the stabilization of partially linear cascade systems with partial state feedback.

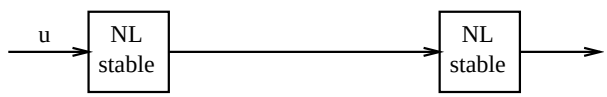
Key words. cascade systems, delay equations, nonlinear control

AMS subject classifications. 34K20, 93A20, 93D15

PII. S0363012999365042

1. Introduction.

The stability analysis of the series (cascade) interconnection of two stable nonlinear systems described by ODEs is a classical subject in system theory [8, 9, 11].



Contrary to the linear case, the zero-input global asymptotic stability of each subsystem does not imply the zero-input global asymptotic stability of the interconnection. The output of the first subsystem acts as a transient input disturbance which can be sufficient to destabilize the second subsystem. In the ODE case, such destabilizing mechanisms are well understood since the seminal work by Sussmann and Kokotovic [10]. They can be subtle but are almost invariably associated to a finite escape time in the second subsystem. (Some states blow up to infinity in a finite time.) The present paper explores similar instability mechanisms generated by the series interconnection of nonlinear delay differential equations (DDEs). In particular, we consider the situation where the destabilizing effect of the interconnection is delayed and examine the difference from the ODE situation.

Instrumental to the stability analysis of cascades, we first study the effect of external (affine) perturbations w on the stability of nonlinear time delay systems

$$\dot{z} = f(z, z(t-\tau)) + \Psi(z, z(t-\tau))w, \quad z \in \mathbb{R}^n, \ w \in \mathbb{R}, \tag{1.1}$$

where we assume that the equilibrum $z = 0$ of $\dot{z} = f(z, z(t-\tau))$ is globally asymptotically stable (GAS). We consider perturbations $w = \eta(t)$ which belong to both L_1 and L_∞ and investigate the region in the space of initial conditions which give rise to bounded solutions under various assumptions on the system and the perturbation.

*Received by the editors December 13, 1999; accepted for publication (in revised form) March 11, 2001; published electronically September 7, 2001.
http://www.siam.org/journals/sicon/40-3/36504.html
†Department of Computer Science, Katholieke Universiteit Leuven, B-3001, Leuven, Belgium (Wim.Michiels@cs.kuleuven.ac.be, Dirk.Roose@cs.kuleuven.ac.be).
‡Institut Montefiore, B28, University of Liege, 4000 Liege Sart-Tilman, Belgium (r.sepulchre@ulg.ac.be).

These results are strengthened to asymptotic stability results when the perturbation is generated by a GAS DDE.

We consider both global and semiglobal results. In the ODE-case, an obstruction to global stability is formed by the fact that arbitrarily small input perturbations can cause the state to escape to infinity in a finite time, for instance, when the interconnection term $\Psi(z)$ is nonlinear in z. This is studied extensively in the literature in the context of the stability of cascades, see, e.g., [10, 8] and the references therein. Even though delayed perturbations do not cause a finite escape time, we explain a similar mechanism giving rise to unbounded solutions, caused by nonlinear delayed interconnection terms.

In situations where unbounded solutions are inevitable for large initial conditions, we investigate under which conditions trajectories can be bounded semiglobally in the space of initial conditions, in case the perturbation is parametrized, i.e., $\eta = \eta(t, a)$. Hereby we let the parameter a control the L_1 or L_∞ norm of the perturbation. We also consider the effect of concentrating the perturbation in an arbitrarily small time-interval. The study of controlled perturbations is motivated by the situation where the perturbation is the output of a controlled system; see Figure 1.1.

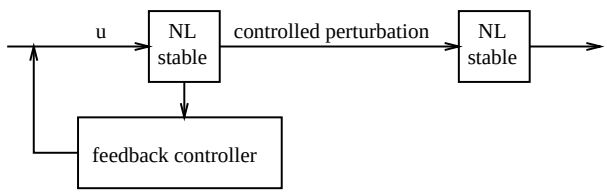


FIG. 1.1. *Partial state feedback as a way of controlling the input perturbation to the second subsystem.*

In particular, we will be interested in the stabilization of the cascade

$$\left\{ \begin{array}{l} \dot{z} = f(z, z(t-\tau)) + \Psi(z, z(t-\tau))y, \\ \dot{\xi} = A\xi + Bu, \\ y = C\xi, \qquad \xi \in \mathbb{R}^\mu, \quad u, y \in \mathbb{R}, \end{array} \right. \tag{1.2}$$

where the perturbation w in (1.1) is now the output of a linear system. We assume that the pair (A, B) is controllable and will control the "perturbation" y with linear state feedback

$$u = F\xi. \tag{1.3}$$

In the ODE-case this stabilization problem has been extensively studied in the literature, for instance, in [12, 1, 10, 6]. Because the output of the linear subsystem, which acts as a destabilizing disturbance to the nonlinear subsystem, can cause trajectories to escape to infinity in a finite time, one typically tries to drive the "perturbation" y quickly to zero. However, a high-gain control, placing all observable eigenvalues far into the left half plane, will not necessarily result in large stability regions because of the fast peaking phenomenon [10, 8]. Peaking is a structural property of the ξ-subsystem whereby achieving faster convergence implies larger overshoots which can in turn destabilize the cascade. Semiglobal stability results are obtained when imposing structural assumptions on the ξ-subsystem (a nonpeaking system) or by imposing

conditions on the z-subsystem and the growth of the interconnection term Ψ: for instance, in [8, Theorem 4.41] one requires a nonpeaking linear subsystem, and the conditions of [10, Theorem 9.1] are a trade-off between peaking and growth.

The structure of the paper is as follows. After some preliminaries (section 2), we study the effect of bounded input perturbations in sections 3 and 4 and use the obtained results to study the stabilization of partially linear cascades with partial state feedback in section 5.

2. Preliminaries. The state of the delay equation (1.1) at time t can be described as a vector $z(t) \in \mathbb{R}^n$ or as a function segment z_t defined by

$$z_t(\theta) = z(t+\theta),\ \theta \in [-\tau,\ 0].$$

Therefore, delay equations form a special class of functional differential equations [2, 4, 5].

We assume that the right-hand side of (1.1) is continuous in all of its arguments and Lipschitz in z and $z(t-\tau)$. Then a solution is uniquely defined by specifying as an initial condition a function segment z_0 whereby $z_0 \in \mathcal{C}([-\tau,\ 0],\ \mathbb{R}^n)$, the Banach space of continuous functions mapping the delay-interval $[-\tau,\ 0]$ into $\mathbb{R}^n$ and equipped with the supremum-norm $\|.\|_s$.

Sufficient conditions for stability of a functional differential equation are provided by the theory of Lyapunov functionals [2, 5], a generalization of the classical Lyapunov theory for ODEs. For functional differential equations of the form

$$\dot{z} = F(z_t), \tag{2.1}$$

according to [2, Definition V.5.3], a mapping $V : \mathcal{C}([-\tau,\ 0],\ \mathbb{R}^n) \to \mathbb{R}$ is called a Lyapunov functional on a set G if V is continuous on $\overline{G}$ and $\dot{V} \leq 0$ on G. Here $\dot{V}$ is the upper-right-hand derivative of V along the solutions of (2.1), i.e.,

$$\dot{V}(z_t) = \limsup_{h \to 0+} \frac{1}{h} \left[V(z_{t+h}) - V(z_t) \right].$$

The following theorem, taken from [2, Corollary V.3.1], provides sufficient conditions for global asymptotic stability.

THEOREM 2.1. *Suppose $z = 0$ is a solution of (2.1) and $V : \mathcal{C}([-\tau,\ 0],\ \mathbb{R}^n) \to \mathbb{R}$ is continuous with $V(0) = 0$. When there exist nonnegative functions $a(r)$ and $b(r)$ with $a(r) > 0$ as $r > 0$ and $a(r) \to \infty$ as $r \to \infty$ such that*

$$a(\|z(t)\|) \leq V(z_t), \quad \dot{V}(z_t) \leq -b(\|z(t)\|),$$

then the zero solution is stable, and every solution is bounded. If, in addition, $b(r)$ is positive definite, then every solution approaches zero as $t \to \infty$.

Instead of working with functionals, it is also possible to use classical Lyapunov functions when relaxing the condition $\dot{V} \leq 0$. This approach, leading to the so-called Lyapunov–Razumikhin theorems [5], is not considered in this paper.

In most of the theorems of the paper, the condition of global asymptotic stability for the unperturbed system ((1.1) with $\eta = 0$) is not sufficient. When the dimension of the system is higher than one, we sometimes need precise information about the interaction of different components of the state $z(t)$. This information is captured in the Lyapunov functional, associated with the unperturbed system. Therefore, when necessary, we will restrict ourselves to the class of DDEs, satisfying the following assumption.

ASSUMPTION 2.2. *The unperturbed system* $\dot{z} = f(z, z(t-\tau))$ *is* delay-independent *globally asymptotically stable (i.e., GAS for all values of the delay) with a Lyapunov functional of the form*

$$V(z_t) = k(z) + \int_{t-\tau}^{t} l(z(\theta))d\theta \tag{2.2}$$

with $k(z) > 0$, $l(z) \geq 0$, *and* $k(z)$ *radially unbounded, such that the conditions of Theorem* 2.1 *(with* $b(r)$ *positive definite) are satisfied.*

This particular choice is motivated by the fact that such functionals are used for a class of linear time-delay systems [2, 5]. Furthermore, choosing a delay-independent stable unperturbed system also allows us to investigate whether the results obtained in the presence of perturbations are still global in the delay. Note that in the ODE-case (2.2) reduces to $V = k(z)$ and hardly forms any restriction because under mild conditions its existence is guaranteed by converse theorems.

The perturbation $\eta(t) \in L_p([0,\ \infty))$ when $\exists M$ such that $\|\eta\|_p = \left[\int_0^\infty |\eta(s)|^p ds\right]^{\frac{1}{p}} = M < \infty$, $\eta(t) \in L_\infty$, when $\|\eta\|_\infty = \sup_{t\geq 0} |\eta(t)| < \infty$.

We assume η in (1.1) to be continuous and to belong to both L_1 and L_∞. When the perturbation is generated by an autonomous DDE, $\dot{\xi} = a(\xi, \xi(t-\tau))$, $\eta = b(\xi, \xi(t-\tau))$ with a and b continuous, locally Lipschitz and $b(0,0) = 0$, which is GAS and locally exponentially stable (LES), these assumptions are satisfied.

In the paper we show that when the unperturbed system is delay-independent stable and the initial conditions are bounded (i.e., $\|z_0\|_s \leq R < \infty$), arbitrarily small perturbations may cause unbounded trajectories provided the delay is large enough; hence arbitrarily small perturbations may destroy the delay-independent stability property. For such cases it is instructive to investigate whether semiglobal stabilization in the delay is possible: with a parametrized perturbation $\eta(t, a)$, we say that the trajectories of (1.1) can be bounded semiglobally in z and semiglobally in the delay if for each compact region $\Omega \subset \mathbb{R}^n$ and $\forall \bar{\tau} \in \mathbb{R}^+$ there exists a positive number $\bar{a}$ such that for every delay value $\tau \leq \bar{\tau}$, all initial conditions $z_0 \in \mathcal{C}([-\tau,\ 0],\ \mathbb{R}^n)$, with $z_0(\theta) \in \Omega,\ \theta \in [-\tau,\ 0]$, give rise to bounded trajectories when $a \geq \bar{a}$.

A C^0 function $\gamma : \mathbb{R} \to \mathbb{R}$ belongs to class κ if it is strictly increasing and $\gamma(0) = 0$. The symbol $\|.\|$ is used for the Euclidean norm in $\mathbb{R}^n$, and by $\|x, y\|$ we mean $\left(\|x\|^2 + \|y\|^2\right)^{\frac{1}{2}}$.

3. The mechanism of destabilizing perturbations. In contrast to linear systems, small perturbations (in the L_1 or L_∞ sense) are sufficient to destabilize nonlinear differential equations. In the ODE-case, the nonlinear mechanism for instability is well known: small perturbations suffice to make solutions escape to infinity in a finite time, for instance, when the interconnection term Ψ is nonlinear in z. This is illustrated with the following example:

$$\begin{aligned} \dot{z} &= -z + z^2\eta, \\ \dot{\eta} &= -a\eta, \end{aligned} \tag{3.1}$$

which can be solved analytically for z to give

$$z(t) = \frac{e^{-t}}{\frac{1}{z(0)} - \int_0^t e^{-s}\eta(s)ds} = \frac{e^{-t}}{\frac{1}{z(0)} - \eta(0)\int_0^t e^{-(1+a)s}ds}. \tag{3.2}$$

If $z(0)\eta(0) > 1 + a$, z escapes to infinity in a finite time t_e, which is given by

$$t_e = \frac{1}{1+a} \log \left(\frac{z(0)\eta(0)}{z(0)\eta(0) - (1+a)} \right). \tag{3.3}$$

This last expression shows that the escape time becomes smaller as the initial conditions are chosen larger, and, as a consequence, however fast $\eta(t)$ would be driven to zero in the first equation of (3.1), $z(0)$ could always be chosen large enough for the solution to escape to infinity in finite time.

In the simple example (3.1), the perturbation is the output of a stable linear system. Its initial condition $\eta(0)$ dictates the L_∞ norm of the perturbation, while the parameter a controls its L_1 norm. Making these norms arbitrarily small does not result in global stability. This is due to the nonlinear growth of the interconnection term.

One may wonder whether the instability mechanism encountered in the ODE situation (3.1) will persist in the DDE situation

$$\begin{cases} \dot{z} = -bz + z(t-\tau)^2\eta, \\ \dot{\eta} = -a\eta. \end{cases} \tag{3.4}$$

In contrast to (3.1), system (3.4) exhibits no finite escape time. This can be proven by application of the method of steps, i.e., from the boundedness of $z(\theta)$, $\theta \in [(k-1)\tau, k\tau]$, we conclude boundedness in $[k\tau, (k+1)\tau]$ of $\dot{z}(\theta)$ and thus of $z(\theta)$. Nevertheless the exponentially decaying input η still causes unbounded solutions in (3.4): this particular system is easily seen to have an exponential solution $z_e(t) = \frac{a+b}{\eta(0)} e^{2a\tau} e^{at}$. The instability mechanism can be explained by the superlinear divergence of the solutions of $\dot{z} = z^\alpha(t-\tau)$ for $\alpha > 1$.

PROPOSITION 3.1.

$$\dot{z} = z(t-\tau)^\alpha, \quad \alpha > 1,$$

has solutions which diverge faster than any exponential function.

Proof. Take as an initial condition a strictly positive solution segment z_0 over $[-\tau,\ 0]$ with $z(0) > 1$. For $t \geq 0$, the trajectory is monotonically increasing. This means that in the interval $[k\tau,\ (k+1)\tau]$ for $k \geq 1$,

$$z((k-1)\tau)^\alpha \leq \dot{z} \leq z(k\tau)^\alpha.$$

The solution at point $k\tau, k \geq 1$ is bounded below by the sequence satisfying

$$z_{k+1} = z_k + \tau z_{k-1}^\alpha, \quad z_0 = z(0),\ z_1 = z(\tau),$$

which has limit $+\infty$. The ratio $R_k = \frac{z_k}{z_{k-1}}$ satisfies

$$R_{k+1}R_k = R_k + \tau z_{k-1}^{\alpha-1},$$

and, consequently, $(R_{k+1} - 1)R_k$ tends to infinity. However, for an exponential function e^{at}, $R = e^{a\tau}$ and $(R-1)R$ is constant. ∎

Because of the faster than exponential growth of z in (3.4), the arbitrarily fast exponential decay of η cannot counter the blow-up caused by the nonlinearity in $z(t-\tau)$, and hence the system is not GAS.

The instability mechanism illustrated by (3.1) and (3.4) can be avoided by imposing suitable growth restrictions on the interconnection term Ψ. When the unperturbed system is scalar, it is sufficient to restrict the interconnection term to have linear growth in both of its arguments, i.e.,

$$\exists c_1, c_2 > 0 \text{ such that } \|\Psi(z, z(t-\tau))\| \leq c_1 + c_2\|z, z(t-\tau)\|. \tag{3.5}$$

This linear growth condition is by itself not sufficient, however, if the unperturbed system has dimension greater than one. In that case, the interaction of the different components of the state $z(t)$ can still cause "nonlinear" effects leading to unbounded solutions. An illustration of this phenomenon is given by the system

$$\left\{ \begin{array}{rcl} \dot{z}_1 &=& -z_1 + z_2\eta(t), \\ \dot{z}_2 &=& -z_2 + z_1^2 z_2, \\ \dot{\eta} &=& -\eta, \end{array} \right. \tag{3.6}$$

which was shown in [8] to have unbounded solutions, despite the linearity of the interconnection. The instability is caused by the mutual interaction between z_1 and z_2 when $\eta \neq 0$.

The following theorem, inspired by Theorem 4.7 in [8], provides sufficient conditions for bounded solutions. To prevent the instability mechanism due to interacting states, conditions are put on the Lyapunov functional of the unperturbed system.

THEOREM 3.2. *Assume that the system $\dot{z} = f(z, z(t-\tau)) + \Psi(z, z(t-\tau))\eta$ satisfies Assumption* 2.2 *and that the interconnection term $\Psi(z, z(t-\tau))$ grows linearly in its arguments, i.e., satisfies* (3.5)*. Furthermore, if the perturbation $\eta(t) \in L_1([0, \infty))$ and $k(z)$ satisfies*

(i) $\alpha_1\|z\|^\gamma \leq k(z) \leq \alpha_2\|z\|^\gamma, \quad 0 < \alpha_1 < \alpha_2 < \infty, \quad 1 \leq \gamma < \infty,$

(ii) $||\frac{dk}{dz}||\ ||z|| \leq ck(z),$

then all trajectories of the perturbed system are bounded for all values of the time delay.

Condition (ii) is sometimes called a *polynomial growth condition* because it is satisfied if $k(z)$ is polynomial in z but not satisfied if $k(z)$ is exponential in z.

Proof. Along a trajectory $z(t)$ we have

$$\begin{aligned} \dot{V} &\leq \left\| \frac{dk}{dz} \right\| \|\Psi(z, z(t-\tau))\||\eta| \\ &\leq c\frac{k(z)}{\|z\|}(c_1 + c_2\sqrt{\|z\|^2 + \|z(t-\tau)\|^2})|\eta| \\ &\leq c\alpha_2^{1/\gamma}k(z)^{1-1/\gamma}\left(c_1 + c_2\sqrt{\tfrac{k(z)^{2/\gamma}}{\alpha_1^{2/\gamma}} + \tfrac{k(z(t-\tau))^{2/\gamma}}{\alpha_1^{2/\gamma}}}\right)|\eta| \\ &\leq c\alpha_2^{1/\gamma}(c_1 k(z)^{1-1/\gamma} + c_2\alpha_1^{-1/\gamma}\sqrt{k(z)^2 + k(z)^{2-2/\gamma}k(z(t-\tau))^{2/\gamma}})|\eta| \\ &\leq c\alpha_2^{1/\gamma}(c_1 V^{1-1/\gamma} + c_2\alpha_1^{-1/\gamma}\sqrt{V^2 + V^{2-2/\gamma}k(z(t-\tau))^{2/\gamma}})|\eta|. \end{aligned} \tag{3.7}$$

For $t \in [0,\ \tau]$, $z(t)$ cannot escape to infinity because $k(z(t-\tau))$ is bounded (calculated from the initial condition), and the above estimate can be integrated over the interval since the right-hand side is linear in V and $\eta \in L_1$.

For $t \geq \tau$ we can use the estimate $k(z(t-\tau)) \leq V(z(t-\tau))$:

$$\dot{V} \leq c\alpha_2^{1/\gamma}\left(c_1 V^{1-1/\gamma} + c_2\alpha_1^{-1/\gamma}\sqrt{V^2 + V^{2-2/\gamma}V(t-\tau)^{2/\gamma}}\right)|\eta|.$$

Because this estimate for $\dot{V}$ is increasing in both of its arguments, an upper bound for V along the trajectory is described by

$$\dot{W} = c\alpha_2^{1/\gamma}\left(c_1 W^{1-1/\gamma} + c_2\alpha_1^{-1/\gamma}\sqrt{W^2 + W^{2-2/\gamma}W(t-\tau)^{2/\gamma}}\right)|\eta|$$

with $W(z_\tau) = V(z_\tau)$ as the initial condition. Via the method of steps, it is clear that W cannot escape to infinity in a finite time. From $t = \tau$ on, W is monotonically increasing. As a consequence, for $t \geq 2\tau$, $W(t) \geq W(t-\tau)$ and

$$\dot{W} \quad \leq c\alpha_2^{1/\gamma}\left(c_1 W^{1-1/\gamma} + c_2\alpha_1^{-1/\gamma}\sqrt{2}W\right)|\eta(t)|,$$

and this estimate can be integrated leading to uniform boundedness of $W(t)$ and $V(t)$ in $[0,\ \infty)$ since $\eta(t) \in L_1$. Hence the trajectory $z(t)$ is bounded. ∎

Remark 3.3. When the interconnection term is undelayed, i.e., Ψ depends only on the argument z, condition (i) in Theorem 3.2 can be dropped, and as a special case (f also undelayed), Theorem 4.7 of [8] is recovered. The presence of a delay in the unperturbed system does not provide additional complications compared to the ODE-case, and the proof is analogous to the proof of Theorem 4.7 of [8].

Along a trajectory, we now have

$$\dot{V} \leq \left\|\frac{dk}{dz}\right\|(c_1 + c_2\|z\|)|\eta|.$$

When $\|z\| \geq 1$ it follows from $\|\frac{dk}{dz}\| \leq \frac{ck(z)}{\|z\|}$ that $\dot{V} \leq c(c_1 + c_2)V|\eta|$. When $\|z\| \leq 1$, we have $\dot{V} \leq M|\eta|$ with $M = \sup_{\|z\|\leq 1}\|\frac{dk}{dz}\|(c_1 + c_2\|z\|)$, and when, in addition, $V \geq 1$, we have $\dot{V} \leq MV|\eta|$.

Hence the following estimate holds whenever $V \geq 1$:

$$\dot{V} \leq \max\left(c(c_1 + c_2), M\right)V|\eta|.$$

From the explicit integration of this estimate the boundedness of V and the trajectory are proven. ∎

4. Semiglobal results for controlled perturbations. Although no global results can be guaranteed in the absence of growth conditions, the examples in the previous section suggest that one should be able to bound the solutions semiglobally in the space of initial conditions by decreasing the size of the perturbation. Therefore, we assume that the perturbation is parametrized:

$$\eta = \eta(t, a).$$

We will consider two cases: (a) parameter a controls the L_1 or the L_∞ norm of η, and (b) a regulates the shape of a perturbation with fixed L_1 norm.

4.1. Controlling the L_1 norm and the L_∞ norm of the perturbation. We first assume that the L_1 norm of η is controlled. We have the following result.

THEOREM 4.1. *Consider the system*

$$\dot{z} = f(z, z(t-\tau)) + \Psi(z, z(t-\tau))\eta(t, a),$$

and suppose that the unperturbed system is GAS with the Lyapunov functional $V(z_t)$ satisfying Assumption 2.2. *If, furthermore, $\|\eta(t,a)\|_1 \to 0$ as $a \to \infty$, then the trajectories can be bounded semiglobally both in z and the delay τ by increasing a.*

Proof. Let $\tau \geq 0$ be fixed, and denote by Ω the desired stability domain in $\mathbb{R}^n$, i.e., such that all trajectories $z(t)$ with initial condition $z_0(\theta) \in \Omega$ for all $\theta \in [-\tau,\ 0]$ are bounded. Let $V_c = \sup_{z_0 \in \Omega} V(z_0)$. Along a trajectory starting in Ω, we have

$$\begin{aligned}\dot{V} &= \frac{dk}{dz} f(z, z(t-\tau)) + l(z(t)) - l(z(t-\tau)) + \frac{dk}{dz}\Psi(z, z(t-\tau))\eta(t,a) \\ &\leq |\frac{dk}{dz}\Psi(z, z(t-\tau))|.|\eta(t,a)|.\end{aligned}$$

As long as $V(t) \leq 2V_c$, both $z(t)$ and $z(t-\tau)$ belong to a compact set. Hence $\exists M > 0$ such that $|\frac{dk}{dz}\Psi(z, z(t-\tau)| \leq M$ and

$$V(t) - V(0) \leq M \int_0^\infty |\eta(s,a)| ds = M\|\eta(t,a)\|_1.$$

When $a \to \infty$, the maximal increase of V tends to zero. Hence there exists a number $\bar{a} >$ such that for $a \geq \bar{a}$, the assumption $V(t) \leq 2V_c$ is valid for all $t \geq 0$ and the trajectories with initial condition in Ω are bounded.

Note that for a fixed region $\Omega \subset \mathbb{R}^n$, V_c increases with τ and this influences both the value of M in the estimation of $|\frac{dk}{dz}\Psi(z, z(t-\tau)|$ and the critical value $\bar{a}$ of a. However, for any delay value τ in a compact interval $[0,\ \bar{\tau}]$, the trajectories starting in Ω are bounded when taking $a \geq \sup_{\tau \in [0,\ \bar{\tau}]} \bar{a}(\tau)$. Hence the trajectories can be bounded semiglobally in both the state and the delay. ∎

This result above is natural because for a given initial condition, a certain amount of energy is needed for destabilization, expressed mathematically by $\|\eta\|_1$. However, global stability in the state is not possible because the required energy can become arbitrarily small for large initial conditions; see, for instance, example (3.1). Later we will discuss why the trajectories cannot be bounded globally in the delay.

Now we consider the case whereby the L_∞ norm of the perturbation is controlled.

THEOREM 4.2. *Consider the system*

$$\dot{z} = f(z, z(t-\tau)) + \Psi(z, z(t-\tau))\eta(t,a).$$

Suppose that the unperturbed system is GAS with the Lyapunov functional $V(z_t)$ satisfying Assumption 2.2. *If $\|\eta(t,a)\|_\infty \to 0$ as $a \to \infty$, the trajectories of the perturbed system can be bounded semiglobally in both z and the delay τ.*

Proof. As in the proof of Theorem 4.1, it is sufficient to prove semiglobal boundedness in the state for a fixed $\tau \geq 0$. Let Ω and V_c be defined as in the proof of Theorem 4.1. Define $\Omega_2 = \{z \in \mathbb{R}^n : k(z) \leq 4V_c\}$, and for some (small) $\epsilon > 0$, $\Omega_\epsilon = \{z \in \mathbb{R}^n :\ \|z\| \leq \epsilon\} \subset \Omega$.

Because of Assumption 2.2, the time derivative of V along a trajectory satisfies

$$\begin{aligned}\dot{V} &= \frac{dk}{dz} f(z, z(t-\tau)) + l(z(t)) - l(z(t-\tau)) + \frac{dk}{dz}\Psi(z, z(t-\tau))\eta(t,a) \\ &\leq -b(\|z\|) + \left|\frac{dk}{dz}\Psi(z, z(t-\tau))\right| \eta(t,a).\end{aligned} \tag{4.1}$$

Let $M = \sup_{z,y \in \Omega_2} |\frac{dk}{dz}\Psi(z,y)|$.

When $z(t) \in \Omega_2 \backslash \Omega_\epsilon$ we have, since b is positive definite, $\dot{V} \leq -b(\|z\|) + M\|\eta\|_\infty \leq -N$ for some number $N > 0$, provided $\|\eta\|_\infty$ is small. For $z(t) \in \Omega_\epsilon$, we have the estimate $\dot{V} \leq M\|\eta\|_\infty$.

Now we prove by contradiction that all trajectories with initial condition in Ω are bounded for small $\|\eta\|_\infty$. Suppose that a solution starting in Ω (hence $V(z_0) \leq V_c$) is unbounded. Then it has to cross the "level set" $2V_c$. Assume that this happens for the first time at t^*. Note that for small $\|\eta\|_\infty$, t^* is large. During the interval $[t^* - \tau,\ t^*]$, V can both increase and decrease. When V increases in this time-interval, $z(t) \in \Omega_\epsilon$, and the increase ΔV is limited: $\Delta V \leq M\|\eta\|_\infty \tau$. When $z(t)$ is outside Ω_ϵ during a time-interval $\Delta t \subset [t^* - \tau,\ t^*]$, V always decreases because $\dot{V} \leq -N$. Since $V(t^*) > V(t^* - \tau)$, we have

$$N\Delta t \leq M\tau\|\eta\|_\infty. \tag{4.2}$$

Hence by reducing $\|\eta(t,a)\|_\infty$ we can make the time-interval Δt arbitrarily small. On the other hand, there exists a constant L, independently of a, such that

$$\left\| \frac{dz}{dt} \right\| \leq \|f(z, z(t-\tau)) + \Psi(z, z(t-\tau))\eta(t,a)\| \leq L < \infty$$

when z_t is inside Ω_2, because f and Ψ map bounded sets into bounded sets. Hence with $|t_2 - t_1| \leq \Delta t$ we have $\|z(t_1) - z(t_2)\| \leq L\Delta t$. Because of (4.2) we can increase a (reduce $\|\eta(t,a)\|_\infty$) such that $L\Delta t \leq \epsilon$, and consequently we have

$$\|z(t)\| \leq 2\epsilon, \quad t \in [t^* - \tau,\ t^*].$$

If ϵ was chosen such that $\Omega_{2\epsilon} = \{z \in \mathbb{R}^n :\ \|z\| \leq 2\epsilon\}$ lies inside Ω, we have a contradiction because this implies $V(t^*) \leq V_c$. Hence a trajectory can never cross the level set $2V_c$ and is bounded. ∎

The results of Theorems 4.1 and 4.2 are not global in the delay, even though the unperturbed system is delay-independent stable. Global results in the delay are generally not possible. We illustrate this fact with the following example, where it is impossible to bound the trajectories semiglobally in the state and globally in the delay, even if we make the size of the perturbation arbitrarily small with respect to the L_1 and L_∞ norms.

Example 4.3. Consider the following system:

$$\begin{cases} \dot{z_1} = -2z_1 + z_1(t-\tau), \\ \dot{z_2} = -\frac{(z_1-2)^2-1}{z_2^2+1} z_2 + z_2^3 \eta(t,a). \end{cases} \tag{4.3}$$

The unperturbed system, i.e., (4.3) with $\eta = 0$, is delay-independent stable. This is proven with the Lyapunov functional

$$V = z_1^2 + \int_{t-\tau}^{t} z_1^2 d\theta + \frac{1}{2} z_2^2.$$

Its time derivative,

$$\begin{aligned} \dot{V} &= [z_1\ z_1(t-\tau)] \begin{bmatrix} -3 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_1(t-\tau) \end{bmatrix} \\ &\quad -z_2^2 \tfrac{(z_1-2)^2-1}{z_2^2+1} \\ &\leq -2z_1^2 - z_2^2 \tfrac{(z_1-2)^2-1}{z_2^2+1}, \end{aligned}$$

is negative definite: when $z_1 \notin [1,\ 3]$, both terms are negative, and in the other case the second term is dominated, because it saturates in z_2. From this it follows that the conditions of Assumption 2.2 are satisfied.

With the perturbation

$$\eta(t,a) \begin{array}{ll} = (t-t_0)^2 e^{-a(t-t_0)} & t \geq t_0 = a^3 \\ = 0, & t \leq t_0, \end{array} \tag{4.4}$$

whereby increasing a leads to a reduction of both $\|\eta\|_1$ and $\|\eta\|_\infty$, we cannot bound the trajectories semiglobally in the state and globally in τ: for each value of a we can find a bounded initial condition (upper bound independent of a), leading to a diverging solution, provided τ is large enough. The first equation of (4.3) has a solution $z_1(t) = 2.5e^{-\alpha t}$, where $-\alpha$ is the real solution of

$$\lambda = -2 + e^{-\lambda\tau}.$$

Clearly, $\alpha \to 0$ as $\tau \to \infty$. Since $z_1(-\tau) = 2.5e^{\alpha\tau} \to 5$ as $\tau \to \infty$, uniform boundedness in τ of this solution over the interval $[-\tau,\ 0]$ (initial condition) is guaranteed. Furthermore, we have

$$z_1(t) \in [1.5,\ 2.5]$$

when $t \in [0,\ \frac{1}{\alpha}\log\frac{5}{3}]$, and thus

$$\dot{z_2} \geq \frac{z_2}{2(1+z_2^2)} + z_2^3\eta(t,a) \tag{4.5}$$

for any positive initial condition $z_2(0)$. A rather lengthy calculation shows that with $z_2(0) = 1$ and the perturbation (4.4), the solution of (4.5) always escapes to infinity in a finite time $t_f(a)$. Hence this also holds for the solution of the system (4.3)–(4.4) when the delay is large enough such that

$$\frac{1}{\alpha(\tau)}\log\frac{5}{3} > t_f(a).$$

This result is not in contradiction with the intuition that a perturbation with small L_1 norm can only cause a finite escape time when the initial condition is far away from the origin, as illustrated with example (3.1). In the system (4.3) with $\eta = 0$, z_2 is driven away from the origin as long as $z_1 \in [1,\ 3]$. By increasing the delay in the first equation, we can keep z_1 in this interval as long as desired. Thus the diverging transient of the *unperturbed* system is used to drive the state away from the origin, far enough to make the perturbation cause escape.

4.2. Controlling the shape of the perturbation. We now assume that the shape of a perturbation with a fixed L_1 norm can be controlled and consider the influence of a concentration of the perturbation in arbitrarily small time-intervals near $t = 0$. In the ODE-case this does not allow improvement of stability properties. This is illustrated with the first equation of example (3.1): instability occurs when $z(0) \geq \frac{1}{\int_0^t e^{-s}\eta(s)ds}$, and by concentrating the perturbation the stability domain may even shrink, because the beneficial influence of damping is reduced. In the DDE-case however, when the interconnection term is linear in the undelayed argument, it behaves linearly during one delay interval, preventing escape. Moreover, starting from a compact region of initial conditions, the reachable set after one delay interval can be bounded independently of the shape of the perturbation (because of the fixed L_1 norm). After one delay interval we are in the situation treated in Theorem 4.1. This

is expressed in the following theorem. As in Theorem 3.2, the polynomial growth condition prevents a destabilizing interaction between different components of the state vector $z(t)$.

THEOREM 4.4. *Consider*

$$\dot{z}(t) = f(z(t), z(t-\tau)) + \Psi(z(t), z(t-\tau))\eta(t,a), \tag{4.6}$$

and suppose that the unperturbed system is GAS with the Lyapunov functional $V(z_t) = k(z) + \int_{t-\tau}^{t} l(z(\theta))d\theta$ *satisfying Assumption* 2.2. *Let* $k(z)$ *satisfy the polynomial growth condition* $\|\frac{dk}{dz}\|\|z\| \le ck(z)$, *and assume that* Ψ *has linear growth in* $z(t)$, *i.e., there exist two class-*κ *functions* γ_1 *and* γ_2 *such that*

$$\|\Psi(z, z(t-\tau))\| \le \gamma_1(\|z(t-\tau)\|) + \gamma_2(\|z(t-\tau)\|)\|z\|.$$

Assume further that $\|\eta(t,a)\|_1 < M$ *for some constant* M *independent of* a, *and that* $\lim_{a\to\infty} \int_t^\infty |\eta(s,a)|ds = 0$ *for all* $t > 0$. *Then the solutions of* (4.6) *can be bounded semiglobally in* z *and for all* $\tau \in [\tau_1,\ \tau_2]$ *with* $0 < \tau_1 \le \tau_2 < \infty$.

Proof. Consider first a fixed value of $\tau \in [\tau_1,\ \tau_2]$. Let Ω be the desired stability domain in $\mathbb{R}^n$. Let $V_{\max} = \max\left(\sup_{z_t\in\Omega} V(z_t), 1\right)$. The time-derivative of the functional V along a trajectory starting in Ω satisfies

$$\begin{aligned}\dot{V} &\le \left\|\frac{dk}{dz}\right\| .\|\Psi(z, z(t-\tau))\|.|\eta(t,a)| \\ &\le \left\|\frac{dk}{dz}\right\| .\left(\gamma_1(\|z(t-\tau)\|) + \gamma_2(\|z(t-\tau)\|)\|z\|\right).|\eta(t,a)|.\end{aligned}$$

Consider the interval $[0, \tau]$. Then $z(t-\tau)$ belongs to the compact set Ω. When $\|z\| \ge 1$, there exists a constant c_1 such that

$$\begin{aligned}\dot{V} &\le ck(z).\left(\frac{\gamma_1(\|z(t-\tau)\|)}{\|z\|} + \gamma_2(\|z(t-\tau)\|)\right)|\eta(t,a)| \\ &\le cV|\eta(t,a)|\left(\frac{\gamma_1(\|z(t-\tau)\|)}{\|z\|} + \gamma_2(\|z(t-\tau)\|)\right) \\ &\le cc_1V|\eta(t,a)|.\end{aligned} \tag{4.7}$$

When $\|z\| \le 1$ and $V \ge 1$, we have

$$\dot{V} \le c_2|\eta(t,a)| \le c_2V|\eta(t,a)| \tag{4.8}$$

with $c_2 = \sup\left\{\|\frac{dk}{dz}\|\|\Psi(z, z(t-\tau))\| :\ \|z\| \le 1, z(t-\tau) \in \Omega\right\}$. From (4.7) and (4.8) we have, whenever $V \ge 1$ in $[0,\ \tau]$, $\dot{V} \le c_3V|\eta(t,a)|$ with $c_3 = \max(cc_1, c_2)$. Hence for all $t \in [0,\ \tau]$,

$$V \le V_{\max}e^{c_3\int_0^t |\eta(s,a)|ds} \le V_{\max}e^{c_3\|\eta(t,a)\|_1} \le V_{\max}e^{c_3M}.$$

As a consequence, $k(z)$ and $\|z(t)\|$ can also be uniformly bounded over the interval $[0,\ \tau]$, independently of a. Hence there exists a compact set $\Omega_2 \subset \mathbb{R}^n$ such that $z_\tau \in \Omega_2$, whatever the value of a and $z_0 \in \Omega$.

Now we can translate the original problem over one delay interval: at time τ the "initial conditions" belong to the compact set Ω_2, and with $t' = t - \tau$ we have

$$\|\eta(t', a)\|_1 = \int_\tau^\infty |\eta(s,a)|ds \to 0 \text{ as } a \to \infty.$$

Because of Theorem 4.1, we can increase a such that all solutions starting in Ω_2 are bounded.

Until now we assumed a fixed τ. But because $[\tau_1,\ \tau_2]$ is compact, we can take the largest threshold of a for bounded solutions over this delay interval. ∎

Remark 4.5. Whenever the perturbation in (1.1) is generated by a GAS and LES DDE, the boundedness results are strengthened to asymptotic stability results. This can be shown following the lines of the proof of Proposition 4.1 in [8]. Stability follows from a local version of Theorem 4.1, and attractivity follows from the application of a generalization to the time-delay case of the classical theorem proposed by LaSalle [2, Theorem V.3.1].

5. Stabilization of partially linear cascades. In the rest of the paper we consider the stabilization of the cascade (1.2) with the control law (1.3).

From the previous sections it is clear that the input y of the z-subsystem can act as a destabilizing disturbance. However, the control can drive the output of the linear system fast to zero. We will investigate under which conditions this is sufficient to stabilize the whole cascade. An important issue in this context is the so-called *fast peaking* phenomenon [10]. This is a structural property of the ξ-system whereby imposing faster convergence of the output to zero implies larger overshoots which can in turn destabilize the cascade and may form an obstacle to both global and semiglobal stabilizability. We start with a short description of the peaking phenomenon and then apply the results of previous sections to the stabilization of the cascade system (1.2).

Our presentation of the peaking phenomenon is inspired by [10], but, following [8], we place the phenomenon in an input-output framework rather than an input-state framework. We also emphasize the relation between a peaking system and the L_1 norm of its output.

5.1. The peaking phenomenon. When in the system

$$\begin{aligned} \dot{\xi} &= A\xi + Bu, \\ y &= C\xi, \end{aligned} \tag{5.1}$$

the pair (A, B) is controllable, one can always find state feedback laws $u = F\xi$ resulting in an exponential decay rate with exponent $-a$. Then the output of the closed loop system satisfies

$$\|y(t)\| \leq \gamma \|\xi(0)\| e^{-at}, \tag{5.2}$$

where γ depends on the choice of the feedback gain. We are interested in the lowest achievable value of γ among different feedback laws and its dependence upon a. This will be determined by the so-called peaking exponent, which we now define.

Denote by $\mathcal{F}(a)$ the collection of all stabilizing feedback laws $u : \xi \to F\xi$ with the additional property that all *observable*[1] eigenvalues λ of (C, A_F), with $A_F = A + BF$, satisfy $\mathrm{Re}(\lambda) < -a$. For a given a and $F \in \mathcal{F}(a)$, define the smallest value of γ in (5.2) as

$$\kappa_F(a) = \sup \left\{ \|y(t)\| e^{at} \right\},$$

where the supremum is taken over all $t \geq 0$ and all initial conditions satisfy $\|\xi(0)\| \leq 1$. Now denote $\kappa(a) = \inf_{F \in \mathcal{F}(a)} \kappa_F$. The output of system (5.1) is said to have peaking

[1] In [10], where the peaking phenomenon is rather studied in an input-state framework, one places all eigenvalues to the left of the line $\lambda = -a$.

exponent s when there exist constants α_1, α_2 such that

$$\alpha_1 a^s < \kappa(a) < \alpha_2 a^s \tag{5.3}$$

for large a. When $s = 0$ the output is said to be *nonpeaking*.

The peaking exponent s is a structural property related to the zero dynamics. When the system has relative degree r, it can be transformed (including a preliminary feedback transformation) into the normal form [3, 1]:

$$\begin{cases} \dot{\xi}_0 = A_0\xi_0 + B_0 y, \\ y^{(r)} = u, \end{cases} \tag{5.4}$$

which can be interpreted as an integrator chain linearly coupled with the zero-dynamics subsystem $\dot{\xi}_0 = A_0\xi_0$. Using state feedback the output of an integrator chain can be forced to zero rapidly without peaking [8]. Because of the linear interconnection term, asymptotic stability of the zero-dynamics subsystem then implies asymptotic stability of the whole cascade. On the contrary, when the zero dynamics are unstable, some amount of energy, expressed by $\int_0^\infty \|y(t)\|dt$, is needed for its stabilization, and therefore the output must peak. More precisely, we have the following theorem, proven in the appendix.

THEOREM 5.1. *The peaking exponent s equals the number of eigenvalues in the closed right half plane (RHP) of the zero-dynamics subsystem.*

The definition of the peaking exponent (5.3) is based on an upper bound of the exponentially weighted output, while its L_1 norm is important in most of the theorems of section 4. But because the overshoots related to peaking occur in a fast time-scale ($\sim at$), there is a connection. For instance, we have the following theorem, based on a result of Braslavsky and Middleton [7].

THEOREM 5.2. *When the output y of system* (5.1) *is peaking* ($s \geq 1$), *$\|y(t)\|_1$ cannot be reduced arbitrarily.*

Proof. Denote by z_0 an unstable eigenvalue of the zero dynamics of (5.1). When a feedback $u = F\xi$ asymptotically stabilizes the system, the relation between y and $w = u + F\xi$ in the Laplace domain is given by

$$\begin{aligned} Y(s) &= C(sI - \bar{A})^{-1}BW(s) + C(sI - \bar{A})^{-1}\xi(0) \\ &= H(s)W(s) + C(sI - \bar{A})^{-1}\xi(0) \end{aligned}$$

with $\bar{A} = A + BF$. The first term vanishes at z_0 because the eigenvalues of the zero dynamics appear as zeros in the corresponding transfer function $H(s)$, and since the feedback F is stabilizing, no unstable pole-zero cancellation occurs at z_0. Hence

$$\begin{aligned} \|y\|_1 &\geq \int_0^\infty |y(t)e^{-z_0 t}|dt \\ &\geq \left| \int_0^\infty y(t)e^{-z_0 t}dt \right| \\ &= |C(z_0 I - \bar{A})^{-1}\xi(0)|. \end{aligned} \tag{5.5}$$

∎

5.2. Nonpeaking cascades. When the ξ-subsystem is minimum-phase and thus nonpeaking, one can find state feedback laws $u = F_a\xi$ resulting in

$$|y(t)| \leq \alpha_2 e^{-at},$$

and the L_1 norm of the output can be made arbitrarily small. So by Theorem 4.1, the cascade (1.2) can be semiglobally asymptotically stabilized in both the state and the delay.

5.3. Peaking cascades. When the ξ-subsystem is nonminimum-phase, the peaking phenomenon forms an obstacle to semiglobal stabilizability because the L_1 norm of the output cannot be reduced (Theorem 5.2).

For ODE-cascades, we illustrate the peaking obstruction with the following example.

Example 5.3. In the cascade,

$$\begin{aligned}\dot{z} &= -z + z^2 y,\\ \dot{\xi}_1 &= \xi_1 + \xi_2,\\ \dot{\xi}_2 &= u, \quad y = -\xi_2.\end{aligned}$$

The peaking exponent of the ξ-subsystem is 1 (zero dynamics $\dot{\xi}_1 = \xi_1$). The cascade cannot be stabilized semiglobally since the explicit solution of the first equation is given by

$$z(t) = \frac{e^{-t}}{\frac{1}{z(0)} - \int_0^t e^{-s} y(s) ds},$$

whereby $\int_0^\infty e^{-s} y(s) ds = \xi_1(0)$. Hence the solution reaches infinity in a finite time when $0 < \frac{1}{\xi_1(0)} < z(0)$.

For DDE-cascades, we consider two cases.

Case 1. Peaking exponent=1. We can apply Theorem 4.4 and obtain semiglobal stabilizability in the state and in the delay when the interconnection term has linear growth in the undelayed argument: besides (5.5) the L_1 norm of y can also be bounded from above since there exist feedback laws $u = F_a \xi$ and a constant α_2 such that

$$\|y(t)\|_1 \leq \int_0^\infty \alpha_2 a e^{-as} ds = \alpha_2,$$

and because of the fast time-scale property, the energy can be concentrated since for all $t > 0$

$$\int_t^\infty |y(s)| ds \leq \int_t^\infty \alpha_2 a e^{-as} ds \to 0 \text{ as } a \to \infty.$$

Case 2. Peaking exponent > 1. In this case, we expect the L_1 norm of y to grow unbounded with a, as suggested by the following example.

Example 5.4. When ξ_k is considered as the output of the integrator chain,

$$\dot{\xi}_1 = \xi_2, \ \dot{\xi}_2 = \xi_3, \ldots, \dot{\xi}_n = u,$$

the peaking exponent is $k-1$ (Theorem 5.1), and $\|\xi_k(t)\|_1, \ k = 2, n$ cannot be reduced arbitrarily by achieving a faster exponential decay rate. In Proposition 4.32 of [8], it is shown that with the feedback law $u = K(a)\xi = -\sum_{k=1}^n a^{n-k+1} q_{k-1} \xi_k$, where all solutions of $\sum_{k=0}^{n-1} q_k \lambda^k + \lambda^n = 0$ satisfy $\mathrm{Re}(\lambda) < -1$, there exists a constant c independent of a such that

$$|\xi_k(t)| \leq c a^{k-1} e^{-at} \|\xi(0)\|;$$

hence the particular feedback $u = K(a)\xi$ is able to achieve an upper bound which corresponds to definition (5.3) for each choice of the output $y = \xi_k$. It is also shown in [8] that with the same feedback and with as initial condition $\xi_1(0) = 1, \ \xi_k(0) =$

$0,\ k = 2, n$, there exists a constant d such that $s_k = \sup_{t\geq 0} |\xi_k(t)| \geq da^{k-1}$. Define $t_k,\ k = 2, n$ such that $|\xi_k(t_k)| = s_k$. As a consequence,

$$\|\xi_k(t)\|_1 \geq \left| \int_0^{t_{k-1}} \xi_k(s)ds \right| = |\xi_{k-1}(t_{k-1})| \geq da^{k-2}, \quad k = 3, n,$$

while the peaking exponent of output $y = \xi_k$ is $k - 1$.

With the two following examples, we show that when the energy of an exponentially decaying input perturbation ($\sim e^{-at}$) grows unbounded with a, an interconnection term which is linear in the undelayed argument is not sufficient to bound the solutions semiglobally in the state. Because it is hard to deal in general with outputs generated by a linear system with peaking exponent $s > 1$, we use an artificial perturbation $a^s e^{-at}$, which has both the fast time-scale property and the suitable growth rate of the energy (a^{s-1}) with respect to a.

Example 5.5. The solutions of equation

$$\dot{z} = -bz + zz(t-\tau)^\alpha a^s e^{-at}, \quad \alpha > 0, \tag{5.6}$$

cannot be bounded semiglobally in z by increasing a for any $\tau > 0$ if the "peaking exponent" s is larger than one.

Proof. Equation (5.6) has an exponential solution $z_e(t)$:

$$z_e(t) = \left[\frac{(\frac{a}{\alpha} + b)e^{a\tau}}{a^s} \right]^{\frac{1}{\alpha}} e^{\frac{a}{\alpha}t}.$$

Consider the solution $z(t)$ with initial condition $z_0 \equiv L > 0$ on $[-\tau,\ 0]$. For $t \in [0,\ \tau]$, $z(t)$ satisfies

$$\dot{z} = -bz + zL^\alpha a^s e^{-at}$$

and consequently coincides on $[o,\ \tau]$ with

$$y(t) = Le^{L^\alpha a^{s-1}(1-e^{-at})-bt}. \tag{5.7}$$

For large a, expression (5.7) describes a lower bound for $z(t)$ on $[\tau,\ 2\tau]$. Furthermore, $y(t)$ decreases on this interval since it reaches its maximum in $t^*(a)$ with $t^* \to 0$ as $a \to \infty$. Thus imposing $y(2\tau) > z_e(2\tau)$ implies that $z(t) > z_e(t), t \in [\tau,\ 2\tau]$, and from this one can argue[2] that $z(t) \geq z_e(t),\ t \geq \tau$. Thus the trajectory starting with initial condition L on $[-\tau,\ 0]$ is unbounded when

$$Le^{L^\alpha a^{s-1}(1-e^{-2a\tau})-2b\tau} > x_e(2\tau) = \left[\frac{\frac{a}{\alpha} + b}{a^s} \right]^{\frac{1}{\alpha}} e^{\frac{3a}{\alpha}\tau}.$$

When $s > 2$, for each value of L, the solution is unstable for large a, and thus the attraction domain of the stable zero solution shrinks to zero. When $s = 2$, a solution with initial condition $L > \left[\frac{3\tau}{\alpha}\right]^{\frac{1}{\alpha}}$ is unstable for large a. ∎

Even when the interconnection term contains no terms in $z(t)$ but only delayed terms of z, semiglobal results are still not possible in general, as shown with the following example.

[2]From (5.6), intersection at t^* would imply $\dot{z}(t^*) > \dot{z_e}(t^*)$ and we have a contradiction.

Example 5.6. The solutions of the system

$$\dot{z} = -\mathrm{sat}(z) + e^{z(t-\tau)} a^s e^{-at} \tag{5.8}$$

with $\mathrm{sat}(z) = z$ when $|z| \leq 1$ and $\mathrm{sat}(z) = \mathrm{sign}(z)$ otherwise cannot be bounded semiglobally in z by increasing a for any $\tau > 0$ when the "peaking exponent" s is greater than one.

Proof. When $z \geq 1$, (5.8) reduces to

$$\dot{z} = -1 + e^{z(t-\tau)} a^s e^{-at},$$

which has the following explicit solution:

$$z_l(t) = at + b, \quad b = a\tau - \log\left(\frac{a^s}{a+1}\right).$$

When a is large enough such that $b \geq 1$, $z_l(t)$ is a solution of (5.8) for all $t \geq 0$.

When the initial condition of (5.8) is L on $[-\tau,\ 0]$ one can find a lower bound for the corresponding solution $z(t)$ on $[0,\ \tau]$ by integrating

$$\dot{z} = -z + e^L a^s e^{-at},\ z(0) = L,$$

yielding

$$z_u(t) = Le^{-t} + e^{-t} e^L \frac{a^s}{a-1}(1 - e^{-(a-1)t}).$$

Furthermore, for large a, $z_u(t)$ describes a lower bound for this solution in the interval $[0,\ 2\tau]$. When imposing $z_u(2\tau) > z_l(2\tau)$, we have for large a, $z_u(t) > z_l(t)$ for all $t \in [\tau,\ 2\tau]$. ($z_u(t)$ reaches its maximum in $t^*(a) \to 0$ as $a \to \infty$.) Hence $z(t) > z_l(t)$ for all $t \in [\tau,\ 2\tau]$ and consequently for all $t \geq \tau$. Thus the trajectory with initial condition L on $[-\tau,\ 0]$ is unbounded when

$$Le^{-2\tau} + e^{-2\tau} e^L \frac{a^s}{a-1}(1 - e^{-(a-1)2\tau}) > 3a\tau - \log\left(\frac{a^s}{a+1}\right). \qquad \square$$

5.4. Zero dynamics with eigenvalues on the imaginary axis. The situation where the zero dynamics possess eigenvalues on the imaginary axis but no eigenvalues in the open RHP deserves special attention. According to Theorem 5.1, the system is peaking; that is, the L_1 norm of the output cannot be reduced arbitrarily. However, this energy can be "spread out" over a long time-interval. It is indeed well known that a system with all its eigenvalues in the closed left half plane (LHP) can be stabilized with a low-gain feedback, as expressed by the following theorem taken from [8].

THEOREM 5.7. *If a system $\dot{\xi}_0 = A_0\xi_0 + B_0 y$ is stabilizable and the eigenvalues of A_0 are in the closed left half plane, then it can be stabilized with a low-gain control law $y = K_0(a)\xi_0$ which for large a satisfies*

$$|y(t)| \leq \frac{\gamma}{a}\|\xi_0(0)\|.$$

The infinity norm of such a low-gain control signal can be arbitrarily reduced, which results, by Theorem 4.2, in satisfactory stabilizability results when it also acts as an input disturbance of a nonlinear system. This suggests not to force the output

of (5.1) exponentially fast ($\sim e^{-at}$) to zero, which results in peaking, but to drive it rapidly without peaking to the manifold $y = K_0(a)\xi_0$, on which the dynamics are controlled by the low-gain control action. Mathematically, with $e = y - K_0(a)\xi_0$ and a feedback transformation $v = u + M\xi$, the normal form of the ξ-subsystem is transformed into

$$\begin{aligned}\dot{\xi}_0 &= A_0\xi_0 + B_0K_0(a)\xi + B_0e,\\ e^r &= v.\end{aligned}$$

Using a high-gain feedback driving $e(t)$ to zero without peaking, as proven in [8, Proposition 4.37], one can always force the output to satisfy the constraint

$$|y(t)| \le \gamma\left(e^{-at} + \frac{1}{a}\right)\|\xi(0)\| \tag{5.9}$$

with γ independent of a. A systematic treatment of such high-low-gain control laws can be found in [6].

For instance, the system

$$\begin{cases}\dot{\xi}_1 = \xi_2,\\ \dot{\xi}_2 = u, \quad y = \xi_2\end{cases} \tag{5.10}$$

is weakly minimum-phase (zero dynamics $\dot{\xi}_1 = 0$). With the high-low gain feedback $u = -\xi_1 - a\xi_2$ the explicit solution of (5.10) for large a can be approximated by

$$\begin{bmatrix}\xi_1\\ \xi_2 = y\end{bmatrix} \approx c_1e^{-at}\begin{bmatrix}\frac{1}{a}\\ -1\end{bmatrix} + c_2e^{-\frac{1}{a}t}\begin{bmatrix}1\\ -\frac{1}{a}\end{bmatrix}. \tag{5.11}$$

Perturbations satisfying constraint (5.9) can be decomposed in signals with vanishing L_1 or L_∞ norm. This suggests the combination of Theorems 4.1 and 4.2 as follows.

THEOREM 5.8. *Consider the interconnected system*

$$\begin{aligned}\dot{z} &= f(z, z(t-\tau)) + \Psi(z, z(t-\tau))y,\\ \dot{\xi} &= A\xi + Bu,\\ y &= C\xi.\end{aligned}$$

Suppose that the z-subsystem is GAS with the Lyapunov functional $V(z_t)$ satisfying Assumption 2.2 and the zeros of the ξ-subsystem are in the closed LHP. Then the interconnected system can be made semiglobally asymptotically stable in both $[z,\ \xi]$ and the delay, using only partial-state feedback.

Proof. As explained in Remark 4.5, the origin $(z, \xi) = (0, 0)$ is stable. Let Ω be the desired region of attraction in the $(z,\ \xi)$-space and choose R such that for all $(z_0, \xi) \in \Omega,\ \|\xi\| < R$. Because of the assumption on the ξ-subsystem, there exist partial-state feedback laws such that

$$\|y(t)\| \le \gamma\|\xi(0)\|\left(e^{-at} + \frac{1}{a}\right) \le \gamma R\left(e^{-at} + \frac{1}{a}\right)$$

with γ independent of a.

Consider the time-interval $[0,\ 1]$. Because

$$\int_0^1 \gamma R\left(e^{-at} + \frac{1}{a}\right) \to 0, \quad a \to \infty,$$

one can show, as in the proof of Theorem 4.1, that by taking a large, the increase of V can be limited arbitrarily. Hence for $t \leq 1$, the trajectories can be bounded inside a compact region Ω_2. We can now translate the original problem over one time-unit, and since

$$\sup_{t \geq 1} \gamma R \left(e^{-at} + \frac{1}{a} \right) \to 0$$

as $a \to \infty$, we can, by Theorem 4.2, increase a until the stability domain contains Ω_2. Then all trajectories starting in Ω are bounded and hence converge to the origin (Remark 4.5). □

6. Conclusions. In this paper, we first studied the effect of bounded input perturbations on the stability of nonlinear delay equations of the form (1.1).

Global stability results are generally not possible without structural assumptions on the unperturbed system and the interconnection term, because arbitrarily small perturbations can lead to unbounded trajectories, even when they are exponentially decaying. In the ODE-case this is caused by the fact that superlinear destabilizing terms can drive the state to infinity in a finite time. Superlinear delayed terms cannot cause a finite escape-time but can still make trajectories diverge faster than any exponential function.

We also considered semiglobal results when the size or shape of the perturbation can be controlled. We assumed that the unperturbed system is delay-independent stable. When the L_1 or the L_∞ norm of the perturbations is brought to zero, trajectories can be bounded semiglobally in both the state and the delay. By means of an example we explained why global results in the delay are generally not possible. Next we considered the effect of concentrating a perturbation with a fixed L_1 norm in arbitrarily small time-intervals. This leads to semiglobal stabilizability in both the state and the delay (compact delay-intervals not containing $\tau = 0$), when the interconnection term is linear in its undelayed arguments.

Using these boundedness results, we studied the stabilizability of the partially linear cascade (1.2) using partial state feedback. When the interconnection term is nonlinear, output peaking of the linear system can form an obstruction to semiglobal stabilizability because the L_1 norm of the output cannot be reduced by achieving a faster exponential decay rate. If we assume that the interconnection term has linear growth in the undelayed argument and the peaking exponent is one, we have semiglobal stabilizability results, because the L_1 norm of the output can be bounded from above while concentrating its energy. Even with this assumption on the interconnection term, higher peaking exponents may form an obstruction. When the zeros of the linear subsystem are in the closed left half plane, satisfactory stability results are obtained when using a high-low-gain feedback, whereby the output of the linear subsystem can be decomposed in two signals with vanishing L_1 and L_∞ norm, respectively.

The main contribution of this paper lies in generalizing classical cascade results to a class of functional differential equations. Instrumental to this generalization is the observation that the way a bounded input perturbation affects a nonlinear system mainly lies in the way its L_1 and L_∞ norm can be controlled.

Appendix. Proof of Theorem 5.1. We transform the system (5.1) into the normal form

$$\begin{aligned} \dot{\xi}_0 &= A_0\xi_0 + B_0 y, \\ \dot{y} &= y_2, \\ &\vdots \\ \dot{y_r} &= u, \end{aligned} \tag{A.1}$$

where $y = C[\xi_0^T\ Y^T]^T,\ Y = [y\ y_1 \cdots y_r]^T$, is the output, and $\dot{\xi}_0 = A_o\xi_0,\ A_0 \in \mathbb{R}^{m\times m}$, describes the zero dynamics. We consider two cases.

Case 1. All eigenvalues of A_0 lie in the closed RHP.

For the stabilization of the system (A.1), we use a state feedback

$$u = F_0\xi_0 + F_1 Y.$$

The closed loop matrix is

$$A_{cl} = \left[\begin{array}{c|cccc} A_0 & B_0 & & & \\ \hline & 0 & 1 & & \\ & \vdots & & \ddots & \\ & 0 & & \cdots\ 0 & 1 \\ F_0 & & & F_1 & \end{array}\right].$$

For asymptotic stability, the observability of (F_0, A_0) is required. In the other case (unstable) eigenvalues of A_0 will still be present in the closed loop system. Mathematically, when (F_0, A_0) would not be observable, one can perform a similarity transformation on ξ_0 leading to

$$\left[\begin{array}{c|c} A_0 & B_0 \\ \hline F_0 & \end{array}\right] \to \left[\begin{array}{cc|c} A_{\bar{o}} & A_{12} & B_{\bar{o}} \\ 0 & A_o & B_o \\ \hline 0 & F_o & \end{array}\right],$$

whereby $A_{\bar{o}}$ contains the unobservable modes of A_0. These unstable eigenvalues are still present in the closed loop matrix A_{cl}, which contradicts the stability assumption.

As a consequence, the whole system is observable since the observability matrix of (C, A_{cl}) is given by

$$\mathcal{O}_{cl} = \left[\begin{array}{c|c} 0 & \mathcal{O}_{1,2} \\ \hline \mathcal{O}_{2,1} & \mathcal{O}_{2,2} \end{array}\right],$$

whereby $\mathcal{O}_{1,2}$ is the unity matrix in $\mathbb{R}^r$ and

$$\mathcal{O}_{2,1} = \left[\begin{array}{cccc} 1 & & & \\ f_n & 1 & & \\ \vdots & & \ddots & \\ f_n^{m-1} & \cdots & f_n & 1 \end{array}\right] \cdot \left[\begin{array}{c} F_0 \\ F_0A_0 \\ \vdots \\ F_0A_0^{m-1} \end{array}\right]$$

with f_n the last component of F_1. From this it follows that the observability of (C, A_{cl}) is implied by the observability of (F_0, A_0).

Consequently, in order to achieve an exponential decay of the output ($\sim e^{-at}$), we need to place *all* eigenvalues to the left of the line $\lambda = -a$. But now we are in the situation considered by Sussmann and Kokotovic [10]. From Theorem 8.1 in [10], it follows that in this case the peaking exponent equals the dimension of the zero dynamics.

Case 2. A_0 has eigenvalues λ with $\mathrm{Re}(\lambda) < 0$.

With another similarity transformation we split off the asymptotically stable part A_{0s} of A_0. Equation (A.1) becomes

$$\begin{cases} \dot{\xi}_{0s} = A_{0s}\xi_{0s} + A_{0su}\xi_{0u} + B_{0s}y, \\ \dot{\xi}_{0u} = A_{0u}\xi_{0u} + B_{0u}y, \\ y^{(r)} = u. \end{cases}$$

Because ξ_{0s} is linearly coupled with the other states, it is sufficient to consider state-feedback laws for the (ξ_{0u}, Y) subsystem (which render the eigenvalues of A_{0s} unobservable). □

Acknowledgments. The authors thank W. Aernouts for fruitful discussions on the results presented in the paper. This paper presents research results of the Belgian program on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology, and Culture (IUAP P4/02). The scientific responsibility rests with its authors.

REFERENCES

[1] C. Byrnes and A. Isidori, *Asymptotic stability of minimum phase nonlinear systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 1122–1137.

[2] J. Hale and S. M. Verduyn Lunel, *Introduction to Functional Differential Equations*, Appl. Math. Sci. 99, Springer-Verlag, New York, 1993.

[3] A. Isidori, *Nonlinear Control Systems*, 3rd ed., Comm. Control Engrg. Ser., Springer-Verlag, Berlin, 1995.

[4] V. Kolmanovskii and A. Myshkis, *Introduction to the Theory and Application of Functional Differential Equations*, Math. Appl. 463, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999.

[5] V. Kolmanovskii and V. Nosov, *Stability of Functional Differential Equations*, Math. Sci. Engrg. 180, Academic Press, San Diego, CA, 1986.

[6] Z. Lin and A. Saberi, *Semi-global stabilization of partially linear composite systems via feedback of the state of the linear part*, Systems Control Lett., 20 (1993), pp. 199–207.

[7] R. Middleton, *Trade-offs in linear control system design*, Automatica J. IFAC, 27 (1991), pp. 281–292.

[8] R. Sepulchre, M. Janković, and P. Kokotović, *Constructive Nonlinear Control*, Comm. Control Engrg., Springer-Verlag, Berlin, 1997.

[9] E. Sontag, *On the input-to-state stability properties*, European J. Control, 1 (1995), pp. 24–36.

[10] H. Sussmann and P. Kokotovic, *The peaking phenomenon and the global stabilization of nonlinear systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 424–440.

[11] A. Teel, *A nonlinear small gain theorem for the analysis of control systems with saturation*, IEEE Trans. Automat. Control, 41 (1996), pp. 1256–1270.

[12] A. Teel and L. Praly, *Tools for semiglobal stabilization by partial state and output feedback*, SIAM J. Control Optim., 33 (1995), pp. 1443–1488.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 3, pp. 681–698

LEARNING ALGORITHMS FOR MARKOV DECISION PROCESSES WITH AVERAGE COST*

J. ABOUNADI†, D. BERTSEKAS†, AND V. S. BORKAR‡

Abstract. This paper gives the first rigorous convergence analysis of analogues of Watkins's Q-learning algorithm, applied to average cost control of finite-state Markov chains. We discuss two algorithms which may be viewed as stochastic approximation counterparts of two existing algorithms for recursively computing the value function of the average cost problem—the traditional relative value iteration (RVI) algorithm and a recent algorithm of Bertsekas based on the stochastic shortest path (SSP) formulation of the problem. Both synchronous and asynchronous implementations are considered and analyzed using the ODE method. This involves establishing asymptotic stability of associated ODE limits. The SSP algorithm also uses ideas from two-time-scale stochastic approximation.

Key words. simulation-based algorithms, Q-learning, controlled Markov chains, average cost control, stochastic approximation, dynamic programming

AMS subject classification. 93E20

PII. S0363012999361974

1. Introduction.

Q-learning algorithms are simulation-based reinforcement learning algorithms for learning the value function arising in the dynamic programming approach to Markov decision processes. They were first introduced for the discounted cost problem by Watkins [27] and analyzed partially in Watkins [27] and then in Watkins and Dayan [28]. A more comprehensive analysis was given by Tsitsiklis [25] (also reproduced in Bertsekas and Tsitsiklis [7]), which made the connection between Q-learning and stochastic approximation. (See also Jaakola, Jordan, and Singh [15] for a parallel treatment, which made the connection between TD(λ) and stochastic approximation.) In particular, Q-learning algorithms for discounted cost problems or stochastic shortest path (SSP) problems were viewed as stochastic approximation variants of well-known value iteration algorithms in dynamic programming.

These techniques, however, do not extend automatically to the average cost problem, which is harder to analyze even when the model (i.e., controlled transition probabilities) is readily available. Not surprisingly, the corresponding developments for the average cost problem have been slower. One of the first was the "R-learning" algorithm proposed by Schwartz [22]. This is a two-time-scale algorithm that carries out a value iteration-type step to update values of state-action pairs and updates concurrently an estimate of the optimal average cost using the immediate reward along with an adjustment factor. The idea is to obtain a good estimate for the average cost while searching for the optimal policy using a value iteration-type update. Although Schwartz presents some intuitive arguments to justify his algorithm along with some

*Received by the editors September 29, 1999; accepted for publication (in revised form) March 14, 2001; published electronically September 7, 2001.
http://www.siam.org/journals/sicon/40-3/36197.html

†Laboratory for Information and Decision Systems, M.I.T., 77 Massachusetts Avenue, Cambridge, MA 02139 (jinane@mit.edu, dimitrib@mit.edu). The work of these authors was supported by NSF under grant 9600494-DMI and grant ACI-9873339.

‡School of Technology and Computer Science, Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400005, India (borkar@tifr.res.in). The work of this author was supported in part by the US Army grant PAAL 03-92-G-0115 at M.I.T., in part by the Homi Bhabha Fellowship, and by the Government of India, Department of Science and Technology grant No. III 5(12)/96-ET.

numerical results, he does not provide any convergence analysis. Singh [23] presents two Q-learning algorithms for the average cost problem: one is a slight modification of Schwartz's algorithm which updates the estimate of optimal cost at every step. The other one updates the estimate of average cost in a fashion similar to Jalali and Ferguson's deterministic asynchronous algorithm for average cost problems [16]. He provides simulation results for medium-sized problems but no convergence analysis. Finally, Mahadevan [20] discusses average cost problems and the need to consider the average cost criterion, with an emphasis on the difference between gain-optimal and bias-optimal policies. He presents extensive numerical experiments, highlighting the problems the algorithm can run into. He does not, however, provide any convergence analysis. It is also noteworthy that none of these algorithms use the relative value iteration (RVI) algorithm for average cost problems (see, e.g., [4], [21], [24]) as a basis for the learning algorithms because the latter may not converge asynchronously, as shown in [3]. Nevertheless, a diminishing stepsize does work around this problem, as we show in this paper.

We propose and give for the first time a complete convergence analysis of two Q-learning algorithms for average cost. The first is a stochastic approximation analogue of (RVI). The second is a stochastic approximation analogue of a recent value iteration algorithm of Bertsekas based on the SSP formulation of the average cost problem. We consider both synchronous and asynchronous implementations. The analysis relies on the ODE method, based on establishing first the boundedness of iterates and then the asymptotic stability of limiting ODEs. The rest then follows as in the Kushner–Clark approach [18] (see also Kushner and Yin [19]) in the synchronous case and by Borkar's theorem [10] in the asynchronous case.

The paper is organized as follows. The next section describes the two algorithms in both synchronous and asynchronous modes and states the assumptions required in each case. Section 3 provides the convergence analysis of the RVI-based Q-learning algorithm. Section 4 does likewise for the SSP Q-learning algorithm. Section 5 concludes with some general remarks. An Appendix collects some key facts from the literature that we used here.

2. Average cost Q-learning algorithms.

2.1. Preliminaries. We consider a controlled Markov chain $\{X_n\}$ on a finite state space $S = \{1, 2, \ldots, d\}$ with a finite action space $A = \{a_1, \ldots, a_r\}$ and transition probabilities $p(i, a, j) =$ the probability of transition from i to j under action a for $i, j \in S, a \in A$. Associated with this transition is a "cost" $g(i, a, j)$ and the aim is to choose actions $\{Z_n\}$ nonanticipatively (i.e., conditionally independent of the future state trajectory given past states and actions) so as to minimize the "average cost"

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n-1} E[g(X_m, Z_m, X_{m+1})]. \tag{2.1}$$

This problem is extensively treated in [4], [21], and [24] among others, to which we refer the reader for details. We recall here the minimal necessary background material required to motivate our algorithms.

We shall be interested in "stationary policies" wherein $Z_n = v(X_n)$ for a map $v : S \to A$. It is known that an optimal one exists under the following "unichain" condition which we assume throughout.

ASSUMPTION 2.1. *Under any stationary policy, the chain has a single communicating class and a common state (say, s) that is recurrent.*

In particular, this implies that "limsup" in (2.1) is a limit under any stationary policy. It is known that one can then associate a "value function" $V : S \to R$ with the problem, given as the solution to the dynamic programming equations

$$V(i) = \min_a \left[\sum_j p(i,a,j)(g(i,a,j) + V(j)) - \beta \right], \qquad i \in S, \tag{2.2}$$

where β is the optimal cost. $V(\cdot)$ is the unique solution to (2.2) modulo an additive constant. Let $Q(i,a)$ denote the expression inside the square brackets on the right-hand-side (r.h.s.) of (2.2). Equation (2.2) is useful because of the following fact: A stationary policy $v : S \to A$ is optimal if and only if $v(i) \in \text{Argmin}\,(Q(i,\cdot))$ for all i that are recurrent under v. $Q(\cdot,\cdot)$ is called the "Q-factor," also defined uniquely only up to an additive constant. Thus $V(i) = \min_a Q(i,a)$ for all i, and $Q(\cdot,\cdot)$ satisfies

$$Q(i,a) = \sum_j p(i,a,j) \left(g(i,a,j) + \min_b Q(j,b) \right) - \beta, \qquad i \in S, \quad a \in A. \tag{2.3}$$

The aim of any Q-learning algorithm is to "learn" the Q-factors when $p(\cdot,\cdot,\cdot)$ is not known, but one has access to a simulation device that can generate an independent S-valued random variable (i.e., independent of other random variables that might have been generated up to that point in time) ξ_{ia} whose probability law is $p(i,a,\cdot), i \in S, a \in A$. Let $\xi = [\xi_{ia}]$.

2.2. RVI Q-learning. The RVI algorithm is given by (see, e.g., [4], [21], [24])

$$V^{n+1}(i) = \min_a \left[\sum_j p(i,a,j)(g(i,a,j) + V^n(j)) - V^n(i_0) \right], \qquad i \in S, \tag{2.4}$$

where $i_0 \in S$ is an arbitrary but fixed state. This can be shown, under some additional aperiodicity conditions (see [4, Chap. 4]), to converge to the unique $V(\cdot)$ that satisfies (2.2) with $V(i_0) = \beta$. The purpose of subtracting the scalar "offset" $V^n(i_0)$ from each component on the r.h.s. of (2.4) is to keep the iterations stable—recall that $V(\cdot)$ is specified anyway only up to an additive constant. It turns out that $V^n(i_0) \to \beta$. However, $V^n(i_0)$ is not the unique choice of an offset term that makes the algorithm work. More generally, one can replace it by $f(V)$ for an $f : R^d \to R$ satisfying suitable hypotheses. (See below.)

Algorithm (2.4) suggests the "relative Q-factor iteration"

$$Q^{n+1}(i,a) = \sum_j p(i,\alpha,j) \left(g(i,a,j) + \min_b Q^n(j,b) \right) - Q^n(i_0,a_0), \qquad i \in S, \qquad a \in A,$$

with $i_0 \in S, a_0 \in A$ prescribed. The idea behind RVI Q-learning is to replace the conditional average with respect to the transition probability $p(i,a,\cdot)$ by an actual evaluation at a random variable ξ_{ia} with law $p(i,a,\cdot)$ and then "see" the conditional average by means of the well-known averaging effect of the stochastic approximation algorithm. Thus the synchronous RVI Q-learning algorithm is

$$Q^{n+1}(i,a) = Q^n(i,a) + \gamma(n)\left(g(i,a,\xi^n_{ia}) + \min_b Q^n(\xi^n_{ia},b) - f(Q^n) - Q^n(i,a)\right), \qquad i \in S, \qquad a \in A, \tag{2.5}$$

where ξ^n_{ia} are independent with the law of ξ^n_{ia} being $p(i,a,\cdot)$ for all n. $\{\gamma(k)\} \in (0,\infty)$ is the usual diminishing stepsize schedule of stochastic approximation satisfying

$$\sum_k \gamma(k) = \infty, \quad \sum_k \gamma^2(k) < \infty. \tag{2.6}$$

The function $f : R^{d\times r} \to R$ satisfies the following assumption.

ASSUMPTION 2.2. *f is Lipschitz, and, furthermore, for e equal to the constant vector of all 1's in $R^{d\times r}$, $f(e) = 1$ and $f(x + ce) = f(x) + c$ for $c \in R$.*

Examples are $f(Q) = Q(i_0,b_0)$ for prescribed $i_0,b_0, f(Q) = \min_u Q(i_0,u)$ for prescribed $i_0, f(Q) = \frac{1}{dr}\sum_{i,a} Q(i,a)$, and so on.

For the asynchronous algorithm, we hypothesize a set-valued process $\{Y^n\}$ taking values in the set of nonempty subsets of $S \times A$ with the interpretation: $Y^n = \{(i,a) :$ (i,a)th component of Q was updated at time $n\}$. (Thus $Y^n \equiv S\times A$ is the synchronous case.)

Remarks. As argued in [10], we may take $Y^n = \{\phi^n\}$ for some $\phi^n \in S \times A$, i.e., a singleton. This can be achieved by unfolding a single iteration that updates k components into k iterations that update one component each. While this introduces "delays" in the formulation of the algorithm below, that does not affect the results of [10] that we use here. Alternatively, we may use the results of [17, section 4], which work with the Y^n's directly. The only difference is that the resultant ODE is a time-scaled version of the one arising in the former approach with a nonautonomous time-scaling which, however, does not affect its qualitative behavior.

Define

$$\nu(n,i,a) = \sum_{k=0}^{n} I\{(i,a) \in Y^k\},$$

where $I\{\ldots\}$ is the indicator function. Thus $\nu(n,i,a) =$ the number of times $Q^m(i,a)$ was updated up to time n.

The asynchronous RVI Q-learning algorithm then is

$$Q^{n+1}(i,a) = Q^n(i,a) + \gamma(\nu(n,i,a))\left(g(i,a,\xi^n_{ia}) + \min_u Q^n(\xi^n_{ia},u) - f(Q^n) - Q^n(i,a)\right) I\{(i,a) \in Y^n\} \tag{2.7}$$

for $(i,a) \in S \times A$. For the asynchronous case, we need the following additional assumptions.

ASSUMPTION 2.3. *In addition to* (2.6), *$\{\gamma(n)\}$ satisfy the following: If $[\ldots]$ stands for "the integer part of ...," then for $x \in (0,1)$,*

$$\sup_k \gamma([xk])/\gamma(k) < \infty,$$

and

$$\frac{\sum_{m=0}^{[yk]} \gamma(m)}{\sum_{m=0}^{k} \gamma(m)} \to 1 \quad \textit{uniformly in } y \in [x,1].$$

Examples of stepsizes satisfying Assumption 2.3 are $\gamma(n) = \frac{1}{n}, \frac{1}{n \log n}, \frac{\log n}{n}$, etc., for $n \geq 2$.

ASSUMPTION 2.4. *There exists* $\Delta > 0$ *such that*

$$\liminf_{n \to \infty} \frac{\nu(n,i,a)}{n+1} \geq \Delta \text{ a.s.}, \qquad (i,a) \in S \times A.$$

Furthermore, for all $x > 0$ *and*

$$N(n,x) = \min \left\{ m \geq n : \sum_{k=n}^{m} \gamma(k) \geq x \right\},$$

the limit

$$\lim_{n \to \infty} \frac{\sum_{k=\nu(n,i,a)}^{\nu(N(n,x),i,a)} \gamma(k)}{\sum_{k=\nu(n,j,u)}^{\nu(N(n,x),j,u)} \gamma(k)}$$

exists a.s. for all i, j, a, u.

That is, all components are updated comparably often in an evenly distributed manner.

2.3. SSP Q-learning. SSP Q-learning is based on the observation that the average cost under any stationary policy is simply the ratio of expected total cost and expected time between two successive visits to the reference state s. This connection was exploited by Bertsekas in [5] to give a new algorithm for computing $V(\cdot)$, which we describe below.

Define a parametrized family of SSP problems parametrized by a scalar λ as follows.

(i) The state space is $S' = S \cup \{s'\}$, where s' is an artificially added terminal state (i.e., zero-cost and absorbing).

(ii) The action set is A for all states.

(iii) The transition probabilities are

$$p'(i,a,j) = \begin{cases} p(i,a,j) & \text{if} \quad j \neq s, s', \\ 0 & \text{if} \quad j = s, \\ p(i,a,s) & \text{if} \quad j = s'. \end{cases}$$

(iv) The costs are defined by

$$g'(i,a,j) = \begin{cases} g(i,a,j) - \lambda & \text{if} \quad j \neq s, s', \\ 0 & \text{if} \quad j = s, \\ g(i,a,s) - \lambda & \text{if} \quad j = s'. \end{cases}$$

By Assumption 2.1, s' is reached from every state with probability 1. Thus all policies are proper (as defined in [4]). Let $V_\lambda(\cdot)$ denote the value function given as the unique solution to the dynamic programming equations

$$V_\lambda(i) = \min_a \left[\sum_{j=1}^{d} p(i,a,j) \left(g(i,a,j) + \sum_{j \neq s} p(i,a,j) V_\lambda(j) \right) - \lambda \right], \quad 1 \leq i \leq d,$$

$$V_\lambda(s') = 0.$$

For each fixed policy, the cost is linear in λ with negative slope. Thus $V_\lambda(\cdot)$, which by standard dynamic programming arguments is the lower envelope thereof, is piecewise linear with finitely many linear pieces and concave decreasing in λ for each component. If $\lambda = \beta$, we "recover" (2.2), which can be shown to happen when $V_\lambda(s) = 0$. This suggests the coupled iterations

$$V^{k+1}(i) = \min_a \left[\sum_{j=1}^{d} p(i,a,j) \left(g(i,a,j) + \sum_{j \neq s} p(i,a,j) V^k(j) \right) - \lambda^k \right], \qquad i \in S,$$
$$\lambda^{k+1} = \lambda^k + b(k) V^k(s),$$

where $\{b(k)\} \subset (0,\infty)$ with $\sum_k b(k) = \infty$ and $\sum_k b^2(k) < \infty$. This is the algorithm of [5], wherein the first "fast" iteration sees λ^k as quasi-static ($b(k)$'s are "small") and thus tracks $V_{\lambda^k}(\cdot)$, while the second "slow" iteration gradually guides λ^k to the desired value.

This suggests the SSP Q-learning algorithm (synchronous) as

(2.8a)
$$Q^{n+1}(i,a) = Q^n(i,a) + \gamma(n) \left[(g(i,a,\xi^n_{ia}) + \min_u Q^n(\xi^n_{ia},u)) I\{\xi^n_{ia} \neq s\} - \lambda^n - Q^n(i,a) \right],$$

(2.8b)
$$\lambda^{n+1} = \lambda^n + b(n) \min_u Q^n(s,u),$$

where $b(n) = o(\gamma(n))$. Unfortunately, it appears hard to ensure boundedness of $\{\lambda^n\}$. So we propose replacing (2.8b) by

(2.8b′)
$$\lambda^{n+1} = \Gamma \left(\lambda^n + b(n) \min_u Q^n(s,u) \right),$$

where $\Gamma(\cdot)$ is the projection onto an interval $[-K,K]$ with K chosen so that $\beta \in (-K,K)$. (This assumes prior knowledge of a bound on β, but this can be obtained from a bound on $g(\cdot,\cdot,\cdot)$.)

As in the case of RVI Q-learning, we impose Assumptions 2.3 and 2.4 for the asynchronous SSP Q-learning, which is

$$Q^{n+1}(i,a) = Q^n(i,a) + \gamma(\nu(n,i,a)) \left[\left(g(i,a,\xi^n_{ia}) + \min_u Q^n(\xi^n_{ia},u) I\{\xi^n_{ia} \neq s\} \right) \right.$$

(2.9a)
$$\left. - \lambda^n - Q^n(i,a) \right] I\{(i,a) \in Y^n\},$$

(2.9b)
$$\lambda^{n+1} = \Gamma \left(\lambda^n + b(n) \min_u Q^n(s,u) \right).$$

3. Convergence of RVI Q-learning.

3.1. ODE analysis. We can rewrite the synchronous RVI Q-learning algorithm (2.5) as

(3.1)
$$Q^{n+1} = Q^n + \gamma(n)(T(Q^n) - f(Q^n)e - Q^n + M^{n+1}),$$

where Q^n stands for $Q^n(i,a)$, $T : R^{d \times r} \to R^{d \times r}$ is the map defined by

$$(TQ)(i,a) = \sum_j p(i,a,j) \left(g(i,a,j) + \min_u Q(j,u) \right),$$

and, for $n \geq 0$,

$$M^{n+1}(i,a) = g(i,a,\xi_{ia}^n) + \min_u Q^n(\xi_{ia}^n, u) - (TQ^n)(i,a).$$

Letting $\mathcal{F}_n = \sigma(Q^m, M^m, m \leq n), n \geq 0$, we note that, for all n,

$$E[M^{n+1} \mid \mathcal{F}_n] = 0, \tag{3.2}$$

$$E[||M^{n+1}||^2 \mid \mathcal{F}_n] \leq C_1(1 + ||Q^n||^2) \tag{3.3}$$

for a suitable constant $C_1 > 0$.

Define $\hat{T} : R^{d\times r} \to R^{d\times r}, T' : R^{d\times r} \to R^{d\times r}$ by

$$\hat{T}(Q) = T(Q) - \beta e,$$
$$T'(Q) = T(Q) - f(Q)e = \hat{T}(Q) + (\beta - f(Q))e,$$

where, as before, $e \in R^{d\times r}$ is the constant vector of all 1's.

Let $||x||_\infty = \max_{i,a} |x_{ia}|, ||x||_s = \max_{i,a} x_{ia} - \min_{i,a} x_{ia}$ for $x \in R^{d\times r}$. These are, respectively, the max-norm and the span seminorm, the latter having the property that $||x||_s = 0$ if and only if x is a scalar multiple of e. The following "nonexpansivity" properties are then easily verified:

$$||T(Q) - T(Q')||_\infty \leq ||Q - Q'||_\infty,$$

and likewise for $\hat{T}(\cdot)$. Also,

$$||T(Q) - T(Q')||_s \leq ||Q - Q'||_s,$$

and likewise for $\hat{T}(\cdot), T'(\cdot)$. In fact, $||T(Q)||_s = ||T'(Q)||_s = ||\hat{T}(Q)||_s$ since $||e||_s = 0$.

Algorithm (3.1) is in the form of a standard stochastic approximation algorithm with the martingale difference sequence $\{M^{n+1}\}$ serving as the "noise." The ODE approach to analyzing the convergence of such algorithms (described in [2], [13], [18], and [19], among others) is based on the stability of the associated ODE

$$\dot{Q}(t) = T'(Q(t)) - Q(t). \tag{3.4}$$

This subsection is devoted to studying the stability properties of this ODE. We do so through a succession of lemmas. The analysis is inspired by a similar analysis in [9] in the context of value iteration. (See also [17].)

We shall also consider the related ODE

$$\dot{Q}(t) = \hat{T}(Q(t)) - Q(t). \tag{3.5}$$

Note that by the properties of $T(\cdot), f(\cdot)$, both (3.4) and (3.5) have Lipschitz r.h.s.'s and thus are well-posed.

The set G of equilibrium points of (3.5) is precisely the set of fixed points of $\hat{T}(\cdot)$, i.e., the solutions of (2.3) which are unique up to an additive constant. Thus $G = \{Q : Q = Q^* + ce, c \in R\}$, where Q^* is the solution to (2.3) satisfying $f(Q^*) = \beta$. (That there will indeed be one such solution follows from the fact that $f(x + ce) = f(x) + c$ for $c \in R$.)

The next lemma is a special case of Theorem 4.1 of [14] (see the appendix).

LEMMA 3.1. *Let $y(\cdot)$ and z be a solution and an equilibrium point of* (3.5)*, respectively. Then $||y(t) - z||_\infty$ is nonincreasing, and $y(t) \to y^*$ for some equilibrium point y^* of* (3.5) *that may depend on $y(0)$.*

We use this to analyze (3.4). But first note the following.

LEMMA 3.2. *Equation* (3.4) *has a unique equilibrium point at Q^*.*

Proof. Since $f(Q^*) = \beta$, it follows that $T'(Q^*) = \hat{T}(Q^*) = Q^*$; thus Q^* is an equilibrium point for (3.4). Conversely, if $T'(Q) = Q$, then $Q = \hat{T}(Q) + (\beta - f(Q))e$. But the Bellman equation

$$Q = \hat{T}(Q) + ce$$

has a solution if and only if $c = 0$. (This can be deduced from the corresponding statement for (2.2), which is well known, and the relation $V(i) = \min_u Q(i, u)$ modulo an additive constant.) Thus $f(Q) = \beta$, implying $Q = Q^*$. □

LEMMA 3.3. *Let $x(\cdot), y(\cdot)$ satisfy* (3.4) *and* (3.5)*, respectively, with $x(0) = y(0) = x_0$. Then $x(t) = y(t) + r(t)e$, where $r(\cdot)$ satisfies the ODE*

$$\dot{r}(t) = -r(t) + (\beta - f(y(t))).$$

Proof. By the variation of constants formula,

$$x(t) = x_0 e^{-t} + \int_0^t e^{-(t-s)} \hat{T}(x(s)) ds + \left[\int_0^t e^{-(t-s)} (\beta - f(x(s))) ds \right] e,$$
$$y(t) = x_0 e^{-t} + \int_0^t e^{-(t-s)} \hat{T}(y(s)).$$

Therefore, with $\hat{T}_i(\cdot)$ denoting the ith component of $\hat{T}(\cdot)$,

$$\max_i (x_i(t) - y_i(t)) \leq \int_0^t e^{-(t-s)} \max_i (\hat{T}_i(x(s)) - \hat{T}_i(y(s))) ds + \int_0^t e^{-(t-s)} (\beta - f(x(s))) ds,$$
$$\min_i (x_i(t) - y_i(t)) \geq \int_0^t e^{-(t-s)} \min_i (\hat{T}_i(x(s)) - \hat{T}_i(y(s))) ds + \int_0^t e^{-(t-s)} (\beta - f(x(s))) ds.$$

Subtracting, we have

$$\begin{aligned} ||x(t) - y(t)||_s &\leq \int_0^t e^{-(t-s)} ||\hat{T}(x(s)) - \hat{T}(y(s))||_s ds \\ &\leq \int_0^t e^{-(t-s)} ||x(s) - y(s)||_s ds. \end{aligned}$$

By Gronwall's inequality, $||x(t) - y(t)||_s = 0$ for all $t \geq 0$. Since $||x||_s = 0$ if and only if $x = ce$ for some $c \in R$, we have $x(t) = y(t) + r(t)e, t \geq 0$. Since $x(0) = y(0), r(0) = 0$. Since

$$\hat{T}(x + ce) = \hat{T}(x) + ce,$$
$$f(x + ce) = f(x) + c,$$

for $r \in R$ we have

$$\begin{aligned} \dot{r}(t)e &= \dot{x}(t) - \dot{y}(t) \\ &= (\hat{T}(x(t)) - x(t) + \beta - f(x(t))e) - (\hat{T}(y(t)) - y(t)) \\ &= (-r(t) + \beta - f(y(t)))e. \end{aligned}$$ □

THEOREM 3.4. *Q^* is the globally asymptotically stable equilibrium point for* (3.4).

Proof. By the variation of constants formula, in the foregoing,

$$r(t) = \int_0^t e^{-(t-s)}(\beta - f(y(s)))ds. \tag{3.6}$$

Let $y(t) \to y^* \in G$. Then $r(t) \to \beta - f(y^*)$ so that $x(t) \to y^* + (\beta - f(y^*))e$, which must coincide with Q^*, since that is the only equilibrium point for (3.4). To claim asymptotic stability, we also need to prove Liapunov stability. (That is, we need to show that given any $\epsilon > 0$, we can find a $\delta > 0$ such that $\|x(0) - Q^*\|_\infty < \delta$ implies $\|x(t) - Q^*\|_\infty < \epsilon$ for $t \geq 0$.) Now

$$\begin{aligned} ||x(t) - Q^*||_\infty &\leq ||y(t) - Q^*||_\infty + |r(t)| \\ &\leq ||y(0) - Q^*||_\infty + \int_0^t e^{-(t-s)}|\beta - f(y(s))|ds \\ &\leq ||x(0) - Q^*||_\infty + \int_0^t e^{-(t-s)}|f(Q^*) - f(y(s))|ds. \end{aligned} \tag{3.7}$$

Since $f(\cdot)$ is Lipschitz,

$$\begin{aligned} |f(Q^*) - f(y(s))| &\leq L||y(s) - Q^*||_\infty \\ &\leq L||y(0) - Q^*||_\infty \\ &= L||x(0) - Q^*||_\infty \end{aligned} \tag{3.8}$$

for a suitable $L > 0$. Thus

$$||x(t) - Q^*||_\infty \leq (1 + L)||x(0) - Q^*||_\infty.$$

Liapunov stability follows, completing the proof. □

3.2. Boundedness and convergence. The ODE method described variously in [2], [13], [18], [19], etc. immediately yields the following.

THEOREM 3.5. *In both the synchronous and the asynchronous Q-learning iterations (cf.* (2.5), (2.7)*), if $\{Q^n\}$ remain bounded a.s., then $Q^n \to Q^*$ a.s.*

Proof. The synchronous case follows from the standard ODE approach in view of Theorem 3.4. The asynchronous case follows likewise from the results of [10]. (In either case, given our prior assumptions, the only things left to verify are the a.s. boundedness of the iterates, which we simply *assumed* for the time being, and the global asymptotic stability of the associated ODE, which we just proved in the previous subsection.) □

The problem of proving a.s. boundedness remains. We shall indicate two proof approaches. The first, which works only for the synchronous case, is based on Lemma 2.2 of [1] (see Appendix). Note that by Theorem 3.4 and the converse Liapunov theorem [29], there exists a C^1 Liapunov function $V : R^{d\times r} \to R^+$ with $\lim_{||x||\to\infty} V(x) = \infty$ and

$$\langle \nabla V(x), T'(x) - x\rangle < 0, \quad x \neq Q^*.$$

Let B be an open neighborhood of Q^*, and let $C = \{x : V(x) \leq c\}$, where $c > 0$ is chosen sufficiently large so as to ensure that $B \subset$ interior (C). Note that C is compact. Define $\pi : R^{d\times r} \to C$ by

$$\pi(x) = x \qquad \text{if} \quad x \in C,$$

$$= Q^* + \eta(x)(x - Q^*) \quad \text{if} \quad x \notin C,$$

where $\eta(x) = \max\{a > 0 : Q^* + a(x - Q^*) \in B\}$. Consider the "scaled" version of (2.5) given by

$$\begin{aligned} \overline{Q}^{n+1}(i,a) = \tilde{Q}^n(i,a) + \gamma(n) \Big(& g(i,a,\xi^n_{ia}) + \min_u \tilde{Q}^n(\xi^n_{ia},u) \\ & -f(\tilde{Q}^n) - \tilde{Q}^n(i,a) \Big), \qquad i \in S, \quad a \in A, \end{aligned} \tag{3.9}$$

where $\tilde{Q}^n = \pi(\overline{Q}^n)$. The iterates (3.9) remain bounded a.s. by construction. To use Lemma 2.2 of [1], we need the following:

(i) The maps $x \to (1-\gamma(n))x + \gamma(n)T'(x)$ are nonexpansive with respect to $||\cdot||_s$ (where without any loss of generality we take $\gamma(n) < 1$). Note that they are so if $T'(\cdot)$ is, which it indeed is, as already observed.

(ii) The iterates $\{\overline{Q}^n\}$ converge to Q^* a.s., which, in view of Theorem 3.4, is proved exactly as in [1, section 3].

We shall also need the following additional assumption on $f(\cdot)$.

ASSUMPTION 3.6. $|f(Q)| \leq ||Q||_\infty$ for all $Q \in R^{d\times r}$.

Note that this is satisfied by the examples of $f(\cdot)$ that follow Assumption 2.2.

LEMMA 3.7. *Under the additional Assumption* 3.6, $\{Q^n\}$ *given by the synchronous Q-learning iteration* 2.5 *is bounded a.s.*

Proof. In view of above remarks and Lemma 2.2 of [1], $||\overline{Q}^n - Q^n||_s$ remains bounded a.s. Note that

$$\sup_n ||Q^n||_s \leq \sup_n ||\overline{Q}^n||_s + \sup_n ||Q^n - \overline{Q}^n||_s \triangleq K < \infty.$$

Let $D = \max(||Q^0||, \max_{i,a,j} |g(i,a,j)| + K)$. Then by Assumptions 2.2 and 3.6,

$$\begin{aligned} \left|\min_u Q^n(\xi^n_{ia},u) - f(Q^n)\right| &= \left|f\left(Q^n - \left(\min_u Q^n(\xi^n_{ia},u)\right)e\right)\right| \\ &\leq \left|\left|Q^n - \left(\min_u Q^n(\xi^n_{ia},u)\right)e\right|\right|_\infty \\ &\leq ||Q^n||_s \leq K. \end{aligned}$$

Then

$$\begin{aligned} |Q^{n+1}(i,a)| &\leq (1-\gamma(n))||Q^n||_\infty + \gamma(n)\left(\max_{i,a,j} g(i,a,j) + K\right) \\ &\leq (1-\gamma(n))||Q^n||_\infty + \gamma(n)D. \end{aligned}$$

A simple induction shows that $||Q^n||_\infty \leq D$ for all n. ∎

The boundedness argument above does not work for the asynchronous iteration (2.7). The reason is as follows: The term $f(Q^n)e$ (resp., $f(\overline{Q}^n)e$) being subtracted from the r.h.s. of (2.5) (resp., (3.7)) implies that exactly the same "offset" is being subtracted from all the components. These terms, being scalar multiples of e, contribute nothing to the span seminorm, a fact that is crucial in the analysis of [1] used above. In the asynchronous case, there is no way of achieving this without artificial restrictions.

The second technique is that of [13], which applies to both synchronous and asynchronous cases. We need the following assumption.

ASSUMPTION 3.6′. $f(cQ) = cf(Q)$ for all $c \in R, Q \in R^{d\times r}$.

Once again, this is satisfied by all the examples of $f(\cdot)$ given in the preceding section. Define $T_0 : R^{d\times r} \to R^{d\times r}$ by

$$(T_0(x))_{ia} = \sum_j p(i,a,j) \min_b x_{jb}, \quad x = [[x_{ia}]] \in R^{d\times r}.$$

The technique of [13] requires that we look at

$$\begin{aligned} h(x) &\stackrel{\Delta}{=} \lim_{c\to\infty} (T'(cx) - cx)/c \\ &= T_0(x) - x - f(x)e \end{aligned}$$

(in view of Assumption 3.6′) and requires that the origin be the globally asymptotically stable equilibrium point of the ODE $\dot{x}(t) = h(x(t))$. But this is merely a special case of Theorem 3.4, corresponding to $g(\cdot,\cdot,\cdot)$ being identically zero. Thus Theorem 2.2 of [13] applies, implying that $\{Q^n\}$ remains bounded a.s. for both the synchronous iteration (2.5), and its asynchronous version (2.7). (For the latter, see section 4 of [13].) We state this conclusion as a lemma.

LEMMA 3.8. *Under the additional Assumption* 3.6′, $\{Q^n\}$ *given by the synchronous Q-learning iteration* (2.5) *and its asynchronous version* (2.7) *is bounded a.s.*

4. Convergence of SSP Q-learning.

4.1. ODE analysis. Redefine T, T', f as follows. $T : R^{d\times r} \to R^{d\times r}, T' : R^{d\times r\times 1} \to R^{d\times r}, f : R^{d\times r} \to R$ are given by

$$\begin{aligned} (TQ)(i,a) &= \sum_{j=1}^{d} p(i,a,j) \left(g(i,a,j) + \sum_{j\neq s} p(i,a,j) \min_u Q(j,u) \right), \\ (T'(Q,\lambda))(i,a) &= (TQ)(i,a) - \lambda, \\ f(Q) &= \min_u Q(s,u). \end{aligned}$$

Then the synchronous iteration (2.8a)–(2.8b′) can be rewritten as

$$Q^{n+1} = Q^n + \gamma(n)[T'(Q^n,\lambda^n) - Q^n + M^{n+1}], \tag{4.1}$$

$$\lambda^{n+1} = \Gamma(\lambda^n + b(n) f(Q^n)), \tag{4.2}$$

where $M^{n+1} = [M^{n+1}(i,a)]$ with

$$\begin{aligned} M^{n+1}(i,a) = &\left[g(i,a,\xi_{ia}^n) + \min_u Q^n(\xi_{ia}^n, u) \right] I\{\xi_{ia}^n \neq s\} \\ &- \lambda^n - T'(Q^n,\lambda^n). \end{aligned}$$

In this new setup one verifies (3.2), (3.3) as before. Note that (4.2) can be rewritten as

$$\lambda^{n+1} = \lambda^n + e(n),$$

where $e(n) = O(b(n)) = o(\gamma(n))$. Thus the limiting ODE associated with (4.1)–(4.2) is

$$\dot{Q}(t) = T'(Q(t),\lambda(t)) - Q(t),$$

$$\dot{\lambda}(t) = 0.$$

Thus it suffices to consider

$$\dot{Q}(t) = T'(Q(t), \lambda) - Q(t) \tag{4.3}$$

for a fixed λ. As observed in [25], the map $T(\cdot)$, and therefore the map $T'(\cdot, \lambda)$ for fixed λ, is a contraction on $R^{d\times r}$ with respect to a certain weighted max-norm

$$||x||_w \stackrel{\Delta}{=} \max |w_i x_i|, \quad x \in R^{d\times r},$$

for an appropriate weight vector $w = [w_1, \ldots, w_{rd}], w_i > 0$ for all i. In particular, $T'(\cdot, \lambda)$ has a unique fixed point $Q(\lambda)$. A straightforward adaptation of the arguments of [14] then shows the following.

LEMMA 4.1. *$Q(\lambda)$ is the globally asymptotically stable equilibrium for* (4.1). *In fact, $||Q(t) - Q(\lambda)||_w$ decreases monotonically to zero.*

4.2. Boundedness and convergence. Once again we present two alternative schemes for proving the a.s. boundedness of $\{Q^n\}$. (Note that $\{\lambda^n\}$ are bounded anyway, as they are constrained to remain in $[-K, K]$.) The first approach is based on [25].

LEMMA 4.2. *For both synchronous and asynchronous SSP Q-learning algorithms, $\{Q^n\}$ remain bounded a.s.*

Proof. Since $T(\cdot)$ is a contraction with respect to $||\cdot||_w$, we have

$$||T(Q)||_w \leq \alpha||Q||_w + D$$

for some $\alpha \in (0, 1), D > 0$. Thus

$$||T'(Q, \lambda)||_w \leq \alpha||Q||_w + D'$$

with $D' = D + K$. Since the r.h.s. does not involve λ, one can mimick the arguments of [25] to conclude. □

An alternative proof of Lemma 4.2 is to directly quote the results of [13]. For this, consider $T^0 : R^{d\times r} \to R^{d\times r}$ defined by

$$(T^0 Q)(i, a) = \sum_{j\neq s} p(i, a, j) \min_u Q(j, u).$$

Then

$$\lim_{a\to\infty} \frac{T'(cQ, \lambda)}{c} = T^0(Q),$$

and the ODE

$$\dot{Q}(t) = T^0(Q) - Q$$

has the origin as the globally asymptotically stable equilibrium. (This is just a special case of Lemma 4.1 with $g(\cdot) \equiv 0$.) Thus the results of [13] apply, allowing us to conclude Lemma 4.2.

Given the a.s. boundedness of iterates, one proves a.s. convergence for the synchronous case as follows.

LEMMA 4.3. *$||Q^n - Q(\lambda^n)|| \to 0$ a.s.*

Proof. Note that $Q(\lambda)$ is simply the Q-factor associated with the SSP problem described in section 2.3 with the prescribed λ, that is,

$$(Q(\lambda))(i,a) = \sum_{j=1}^{d} p(i,a,j)\left(g(i,a,j) + \sum_{j\neq s} p(i,a,j)V_\lambda(j) - \lambda\right), \qquad i \in S, \qquad a \in A.$$

Since the map $\lambda \to V_\lambda$ is concave, it is continuous, and therefore so is the map $\lambda \to Q(\lambda)$. In view of Lemmas 4.1 and 4.2, the claim now follows as in Corollary 2.1 of [8]. □

We shall also need the following lemma.

LEMMA 4.4.

$$\Pi_{i=0}^{n}(1-b(i)) \to 0, \quad \limsup_{n\to\infty} \sum_{i=0}^{n} \Pi_{j=i+1}^{n}(1-b(j))b(i) < \infty,$$

and for any sequence $\{a_n\}$ *with* $a_n \to 0$,

$$\sum_{i=0}^{n}\left(\prod_{j=i+1}^{n}(1-b(j))\right)b(i)a_i \to 0.$$

Proof. Since $\sum_i b(i) = \infty$ and $1 - x \le e^{-x}$ for all x,

$$\prod_{i=0}^{n}(1-b(i)) \le e^{-\sum_{i=0}^{n} b(i)} \to 0$$

as $n \to \infty$. Let $t(0) = 0, t(n) = \sum_{i=0}^{n-1} b(i), n \ge 1$. Then

$$\begin{aligned}\sum_{i=0}^{n}\prod_{j=i+1}^{n}(1-b(j))b(i) &\le \sum_{i=0}^{n} b(i)\exp\left(-\sum_{j=i+1}^{n} b(j)\right)\\ &= \sum_{i=0}^{n} e^{-(t(n+1)-t(i))}(t(i+1)-t(i))\\ &\le \int_0^{t(n+1)} e^{-(t(n+1)-s)}ds \to 1\end{aligned}$$

as $n \to \infty$. Define $h(t), t \ge 0$, by $h(t) = a_n$ for $t \in [t(n), t(n+1)), n \ge 0$. Then a similar argument shows that

$$\sum_{i=0}^{n}\prod_{j=i+1}^{n}(1-b(j))b(i)a_i \le \int_0^{t(n+1)} e^{-(t(n+1)-s)}h(s)ds.$$

Since $h(t) \to 0$ as $t \to \infty$, the r.h.s. $\to 0$ as $n \to \infty$. □

THEOREM 4.5. *For the synchronous SSP Q-learning algorithm* (2.8a)–(2.8b′), $(Q^n, \lambda^n) \to (Q^*, \beta)$ *a.s.*

Proof. Define $\Delta^n = \lambda^n - \beta, r_n = f(Q^n) - f(Q(\lambda^n)), n \ge 0$. By Lemma 4.3, $r_n \to 0$. Also,

$$\begin{aligned}\Delta^{n+1} &= \Gamma(\Delta^n + \beta + b(n)f(Q^n)) - \beta \\ &= \Gamma(\Delta^n + \beta + b(n)f(Q(\lambda^n)) + b(n)r_n) - \beta. \end{aligned} \tag{4.4}$$

Since the map $\lambda \to V_\lambda$, and therefore also the map $\lambda \to f(Q(\lambda)) = \min_u(Q(\lambda))(s,u)$, is concave and piecewise linear with finitely many linear pieces, it follows that there exist $0 < L_1 \le L_2$ such that

$$-L_2(\lambda_1 - \lambda_2) \le f(Q(\lambda_1)) - f(Q(\lambda_2)) \le -L_1(\lambda_1 - \lambda_2)$$

for all $\lambda_1, \lambda_2 \in R$. (Basically, these are upper and lower bounds on the slope of the map $\lambda \to f(Q(\lambda))$.) With $\lambda_1 = \lambda^n, \lambda_2 = \beta$, and using the fact that $f(Q(\beta)) = f(Q^*) = 0$, this reduces to

$$-L_2\Delta^n \le f(Q(\lambda^n)) \le -L_1\Delta^n.$$

Using this, (4.4), and the fact that $\Gamma(\cdot)$ is nondecreasing, we have

$$\begin{aligned}&\Gamma((1 - L_2 b(n))\Delta^n + b(n)r_n + \beta) - \beta \le \Delta^{n+1} \\ &\quad \le \Gamma((1 - L_1 b(n))\Delta^n + b(n)r_n + \beta) - \beta.\end{aligned}$$

Note that for $i = 1, 2$,

$$\begin{aligned}(1 - L_i b(n))\Delta^n + b(n)r_n + \beta &= \lambda^n + b(n)(r_n - L_i\Delta^n) \\ &= \lambda^n + O(b(n)).\end{aligned}$$

Since $\lambda^n \in [-K, K]$ for all n and $b(n) \to 0$, it follows from the definition of $\Gamma(\cdot)$ that for any $\epsilon > 0$, there is an $N \ge 1$ sufficiently large so that for $n \ge N$,

$$\begin{aligned}&(1 - L_2 b(n))\Delta^n + (b(n)r_n - \epsilon) + \beta \\ &\quad \le \Gamma((1 - L_2 b(n))\Delta^n + b(n)r_n + \beta) \\ &\quad \le \Gamma((1 - L_1 b(n))\Delta^n + b(n)r_n + \beta) \\ &\quad \le (1 - L_i b(n))\Delta^n + (b(n)r_n + \epsilon) + \beta.\end{aligned}$$

Therefore,

$$(1 - L_2 b(n))\Delta^n + b(n)r_n - \epsilon \le \Delta^{n+1} \le (1 - L_1 b(n))\Delta^n + b(n)r_n + \epsilon.$$

Iterating the inequalities, we have for $n > N$

$$\begin{aligned}&\prod_{i=N}^{n+1}(1 - L_2 b(i))\Delta^N + \sum_{i=N}^{n}\prod_{j=i+1}^{n}(1 - L_2 b(j))(b(i)r_i - \epsilon) \le \Delta^{n+1} \\ &\qquad \le \prod_{i=N}^{n+1}(1 - L_1 b(i))\Delta^N + \sum_{i=N}^{n}\prod_{j=i+1}^{n}(1 - L_1 b(j))b(i)(r_i - \epsilon).\end{aligned}$$

Letting $n \to \infty$ and using Lemma 4.4, we have $\Delta^n \to [-C\epsilon, C\epsilon]$ for a suitable constant $C > 0$. Since $\epsilon > 0$ was arbitrary, $\Delta^n \to 0$, i.e., $\lambda^n \to \beta$. Since $\lambda \to Q(\lambda)$ is continuous, $Q(\lambda^n) \to Q(\beta) = Q^*$, and, by Lemma 4.3, $Q^n \to Q^*$. ∎

Remark. If we consider instead the SSP Q-learning algorithm (2.8a)–(2.8b) that does not use the projection $\Gamma(\cdot)$, it is possible to argue as above to conclude that *if* the iterates $\{\lambda^n\}$ remain bounded a.s., then Theorem 4.5 holds.

Finally, we have the following theorem.

THEOREM 4.6. *For the asynchronous SSP Q-learning algorithm* (2.9a)–(2.9b), $(Q^n, \lambda^n) \to (Q^*, \lambda^*)$ *a.s.*

Proof. The analysis of [10], [17] applies, implying, in particular, that Lemma 4.3 holds exactly as before. The only difference is that now the interpolated algorithm would track a time-scaled version of (4.1). The rest is as before because the iteration scheme for $\{\lambda^n\}$ is unchanged. ∎

5. Conclusions. We have presented two Q-learning algorithms for average cost control of finite Markov chains—one based on RVI and another on an SSP formulation of the average cost problem. We have rigorously established their stability and convergence to the desired limits with probability one. As already remarked in the introduction, this is the first rigorous analysis of any Q-learning algorithms for average cost problems. Nevertheless, this is only a first step toward a better understanding of these algorithms. In conclusion, we mention three important directions for future work in this area:

(i) Typically, the state space can be very large. This calls for approximations, such as state aggregation or considering a parametrized family of candidate Q-factor functions with a low dimensional parameter space. (See, e.g., [7], [26].) The algorithms presented need to be interlaced with such approximation architectures and analyzed as such. A popular architecture is a linear combination of suitable basis functions, the weights in question being the parameters that are tuned [7], [26]. A good choice of basis functions is crucial, and it has been suggested that they be based upon sample simulation runs [6]. Yet another technique for reducing computation is to update not at every sample but at an appropriately chosen subsequence of samples. This can be combined with "kernel" methods, where one updates in a neighborhood of the sample in a weighted fashion. While there is an enormous amount of empirical work on such ideas in recent years, the theory has been lacking. Finally, it is worthwhile exploring the use of acceleration techniques in traditional Monte Carlo methods (such as importance sampling) to reinforcement learning.

(ii) Simulation-based algorithms are slow. An analysis of rate of convergence and good speed-up procedures are needed. To some extent, the rate of convergence statements for general stochastic approximation algorithms, based on associated limit theorems or asymptotics for moments, will also bear upon these algorithms. However, they have enough special structure that one should be able to say more. A recent work [12] takes a step in this direction by estimating the number of steps required by Q-learning for discounted cost problems to attain a prescribed level of accuracy with a prescribed probability.

(iii) Extension to the case where the state space is not finite is an open issue. See, however, [11] for the discounted cost problem.

Appendix. We briefly recall the key results from the literature that have been used here in a crucial manner. To begin with, let $F(\cdot,\cdot) = [F_1(\cdot,\cdot),\ldots,F_d(\cdot,\cdot)]^T : R^d \times R^m \to R^d$ be Lipschitz in the first argument uniformly with respect to the second, i.e., for some scalar K, we have

$$\|F(x,y) - F(y,u)\| \le K\|x-y\| \qquad \forall\ x,y,u.$$

Consider the stochastic approximation algorithm of the form

$$x^{k=1} = x^k + \gamma(k)F(x^k,\xi^k), \qquad k \ge 0,$$

for $x^k = [x_1^k,\ldots,x_d^k]$, with $\{\xi^k\}$ i.i.d., R^n-valued random variables. Let $h(x) = E[F(x,\xi)]$. The ODE approach views the above recursion as

$$x^{k=1} = x^k + \gamma(k)[h(x^k) + M^{k+1}]$$

with

$$M^{k+1} = F(x^k,\xi^k) - h(x^k), \qquad k \ge 0,$$

a "martingale difference" sequence. The term in square brackets is viewed as a noisy measurement of $h(x^k)$ with M^{k+1} as the "noise." The iteration can then be viewed as a noisy discretization of the ODE $\dot{x}(t) = h(x(t))$ with diminishing time steps. We assume that this ODE has a globally asymptotically stable equilibrium x^*. If $\{x^k\}$ remains bounded and the martingale $\sum \gamma(k) M^{k+1}$ converges with probability one, both the discretization error and the error due to noise in the above "approximation" of the ODE become asymptotically negligible, and therefore the iterates track the ODE. In particular, $x^k \to x^*$ a.s.

The asynchronous version of this algorithm is

$$x_i^{k+1} = x_i^k + \nu(k,i) I(i \in Y^k) F_i(x_1^{k-\tau_{1i}}(k), \ldots, x_d^{k-\tau_{di}}(k), \xi^k), \qquad 1 \le i \le d,$$

for $k \ge 0$, where (1) $\{Y^k\}$ is a set-valued random process taking values in the subsets of $\{1, \ldots, d\}$, representing the components that do get updated at time k, (2) $\{\tau_{ij}(k)\}$, $1 \le i, j \le d$, $k \ge 0\}$, are bounded random delays, and (3) $\nu(k,i) = \sum_{m=0}^{k} I(i \in Y^m)$ denotes the number of times component i gets updated up to time k. Under the kind of assumptions on $\{\gamma(k)\}$ and $\{\nu(k,i)\}$ we have used here, Borkar [10] shows that the asynchronous iterations track the ODE $\dot{x}(t) = \frac{1}{d} h(x(t))$, which is a time-scaled version of $\dot{x}(t) = h(x(t))$ and has the same trajectories.

The two-time-scale stochastic approximation algorithm of Borkar [8] considers the following iteration:

$$x^{k+1} = x^k + \gamma(k) F(x^k, y^k, \xi^k),$$

$$y^{k=1} = y^k + \beta(k) G(x^k, y^k, \zeta^k),$$

where $\{\xi^k\}$, $\{\zeta^k\}$ are i.i.d., and $\{\beta(k)\}$ satisfy

$$\sum_{k=0}^{\infty} \beta(k) = \infty, \qquad \sum_{k=0}^{\infty} \beta(k)^2 < \infty, \qquad \beta(k) = o(\gamma(k)).$$

Thus $\{y^k\}$ (resp., $\{x^k\}$) is the slow (resp., fast) component of the iteration. One can analyze $\{x^k\}$ viewing $\{y^k\}$ as quasi-static, and then analyze $\{y^k\}$, viewing $\{x^k\}$ as essentially equilibrated. In other words, consider the ODE $\dot{x}(t) = h(x(t), \overline{y})$, where $h(x,y) = E[F(x,y,\xi)]$ and $\overline{y}$ is treated as a fixed parameter. Suppose it has a globally asymptotically stable equilibrium $\lambda(\overline{y})$, where $\lambda(\cdot)$ is a Lipschitz function. Then $\{x^k\}$ tracks $\{\lambda(y^k)\}$. In turn, $\{y^k\}$ tracks the ODE $\dot{y}(t) = g(\lambda(y(t)), y(t))$, where $g(x,y) = E[G(x,y,\zeta)]$. If the latter ODE has a globally asymptotically stable equilibrium y^*, one can show that $(x^k, y^k) \to (\lambda(y^*), y^*)$ a.s.

In dynamic programming applications, an important special class of ODEs arises, wherein $h(x) = f(x) - x$ for an $f : R^d \to R^d$ satisfying the "nonexpansivity property"

$$\|f(x) - f(y)\|_\infty \le \|x - y\|_\infty.$$

The set $B = \{x : f(x) = x\}$ of fixed points of $f(\cdot)$, assumed to be nonempty, is precisely the set of equilibria for the ODE $\dot{x}(t) = f(x(t)) - x(t)$. It is shown in Borkar and Soumyanath [14] that $x(\cdot)$ converges to a point in B and, furthermore, for any $x^* \in B$, $\|x(t) - x^*\|_\infty$ is nonincreasing in t.

Finally, we recall Lemma 2.2 of [1], which has been used here. Let $D \subset R^d$ be an open bounded set, and let $C \subset R^d$ be a set containing D. Define $\Pi_{D,C} : R^d \to \overline{D}$ by

$$\prod_{D,C}(x) = \alpha_{D,C}(x) \cdot x,$$

where

$$\alpha_{D,C}(x) = \begin{cases} 1 & \text{if } x \in C, \\ \max\{\beta > 0 \,:\, \beta x \in \overline{D}\} & \text{if } x \notin C. \end{cases}$$

Let $\|x\|_s = \max_i x_i - \min_i x_i$ define the span seminorm on R^d. Consider an iteration

$$x^{k+1} = G^k(x^k, \xi^k),$$

where $\{\xi^k\}$ is a random process and $\{G^k\}$ satisfy

$$\|G^k(x,\xi) - G^k(y,\xi)\|_s \le \|x - y\|_s \qquad \forall\, x, y, \xi.$$

Suppose the sequence $\{\tilde{x}^k\}$ generated by the "scaled" iteration

$$\tilde{x}^{k+1} = G^k\left(\prod_{D,C}(\tilde{x}^k), \xi^k\right)$$

converges a.s. Lemma 2.1 of [1] then says that $\{\|x^k\|_s\}$ remains bounded with probability one.

REFERENCES

[1] J. Abounadi, D. P. Bertsekas, and V. S. Borkar, *Stochastic Approximation for Nonexpansive Maps: Application to Q-Learning*, Report LIDS-P-2433, Laboratory for Information and Decision systems, MIT, Cambridge, MA, 1998.

[2] A. Benveniste, M. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, Berlin, Heidelberg, 1990.

[3] D. P. Bertsekas, *Distributed dynamic programming*, IEEE Trans. Automat. Control, 27 (1982), pp. 610–616.

[4] D. P. Bertsekas, *Dynamic Programming and Optimal Control, Vol.* 2, Athena Scientific, Belmont, MA, 1995.

[5] D. P. Bertsekas, *A new value iteration method for the average cost dynamic programming problem*, SIAM J. Control Optim., 36 (1998), pp. 742–759.

[6] D. P. Bertsekas and D. A. Castanon, *Rollout algorithms for stochastic scheduling problems*, J. Heuristics, 5 (1999), pp. 89–108.

[7] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, Belmont, MA, 1996.

[8] V. S. Borkar, *Stochastic approximation with two time scales*, Systems Control Lett., 29 (1996), pp. 291–294.

[9] V. S. Borkar, *Recursive self-tuning control of finite Markov chains*, Appl. Math. (Warsaw), 24 (1996), pp. 169–188.

[10] V. S. Borkar, *Asynchronous stochastic approximations*, SIAM J. Control Optim., 36 (1998), pp. 840–851.

[11] V. S. Borkar, *A learning algorithm for discrete time stochastic control*, Probab. Engrg. Inform. Sci., 14 (2000), pp. 243–248.

[12] V. S. Borkar, *On the number of samples required for Q-learning*, in Proceedings of the 38th Allerton Conference, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL, 2000.

[13] V. S. Borkar and S. P. Meyn, *The ODE method for convergence of stochastic approximation and reinforcement learning*, SIAM J. Control Optim., 38 (2000), pp. 447–469.

[14] V. S. Borkar and K. Soumyanath, *A new analog parallel scheme for fixed point computation, part* I: *Theory*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 44 (1997), pp. 351–355.

[15] T. Jaakola, M. I. Jordan, and S. P. Singh, *On the convergence of stochastic iterative dynamic programming algorithms*, Neural Computation, 6 (1994), pp. 1185–1201.

[16] A. Jalali and M. Ferguson, *Adaptive control of Markov chains with local updates*, Systems Control Lett., 14 (1990), pp. 209–218.

[17] V. R. Konda and V. S. Borkar, *Actor-critic–type learning algorithms for Markov decision processes*, SIAM J. Control Optim., 38 (2000), pp. 94–123.
[18] H. J. Kushner and D. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, New York, 1978.
[19] H. J. Kushner and G. Yin, *Stochastic Approximation Algorithms and Applications*, Springer-Verlag, New York, 2000.
[20] S. Mahadevan, *Average reward reinforcement learning: Foundations, algorithms and empirical results*, Machine Learning, 22 (1996), pp. 1–38.
[21] M. L. Puterman, *Markov Decision Processes*, John Wiley and Sons, New York, 1994.
[22] A. Schwartz, *A reinforcement learning method for maximizing undiscounted rewards*, in Proceedings of the 10th International Conference on Machine Learning, Morgan Kaufmann, San Mateo, 1993, pp. 298–305.
[23] S. P. Singh, *Reinforcement learning algorithms for average payoff Markovian decision processes*, in Proceedings of the 12th National Conference on Artificial Intelligence, MIT Press, Cambridge, MA,1994, pp. 202–207.
[24] H. C. Tijms, *Stochastic Modeling and Analysis: A Computational Approach*, John Wiley and Sons, New York, 1986.
[25] J. N. Tsitsiklis, *Asynchronous stochastic approximation and Q-learning*, Machine Learning, 16 (1994), pp. 185–202.
[26] J. N. Tsitsiklis and B. van Roy, *Feature-based methods for large scale dynamic programming*, Machine Learning, 22 (1996), pp. 59–94.
[27] C. Watkins, *Learning from Delayed Rewards*, Ph.D. thesis, Cambridge University, Cambridge, U.K., 1989.
[28] C. Watkins and P. Dayan, *Q-learning*, Machine Learning, 8 (1992), pp. 279–292.
[29] F. W. Wilson, *Smoothing derivatives of functions and applications*, Trans. Amer. Math. Soc., 139 (1967), pp. 413–428.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 3, pp. 699–723

SENSITIVITY ANALYSIS OF THE VALUE FUNCTION FOR OPTIMIZATION PROBLEMS WITH VARIATIONAL INEQUALITY CONSTRAINTS*

YVES LUCET† AND JANE J. YE‡

Abstract. In this paper we perform sensitivity analysis for optimization problems with variational inequality constraints (OPVICs). We provide upper estimates for the limiting subdifferential (singular limiting subdifferential) of the value function in terms of the set of normal (abnormal) coderivative (CD) multipliers for OPVICs. For the case of optimization problems with complementarity constraints (OPCCs), we provide upper estimates for the limiting subdifferentials in terms of various multipliers. An example shows that the other multipliers may not provide useful information on the subdifferentials of the value function, while the CD multipliers may provide tighter bounds. Applications to sensitivity analysis of bilevel programming problems are also given.

Key words. sensitivity analysis, optimization problems, variational inequality constraints, complementarity constraints, limiting subdifferentials, value functions, bilevel programming problems

AMS subject classification. 49K40

PII. S0363012999361718

1. Introduction.

In this paper, we consider the sensitivity analysis for the following optimization problem with variational inequality constraints (OPVIC):

$$\begin{aligned} \text{(OPVIC)} \qquad & \text{minimize} && f(x,y) \\ & \text{subject to} && \Psi(x,y) \le 0, H(x,y) = 0, (x,y) \in C, \\ & && y \in \Omega, \langle F(x,y), y - z\rangle \le 0 \quad \forall z \in \Omega, \end{aligned} \tag{1}$$

where the following basic assumptions are satisfied:

(BA) The functions $f : R^{n+m} \to R$, $\Psi : R^{n+m} \to R^d$, $H : R^{n+m} \to R^l$, and $F : R^{n+m} \to R^m$ are Lipschitz near any given point of C; C is a closed subset of R^{n+m}, and Ω is a closed convex subset of R^m. Note that the OPVIC is also called the mathematical program with equilibrium constraints (MPEC).

By definition of a normal cone in the sense of convex analysis, the variational inequality (1) is equivalent to saying that $y \in \Omega$ and the vector $-F(x,y)$ is in the normal cone of the convex set Ω at y. Hence the OPVIC can be rewritten as an optimization problem with a generalized equation constraint:

$$\begin{aligned} \text{(GP)} \qquad & \text{minimize } f(x,y) \\ & \text{subject to } \Psi(x,y) \le 0, H(x,y) = 0, (x,y) \in C, \\ & 0 \in F(x,y) + N_\Omega(y), \end{aligned} \tag{2}$$

where

$$N_\Omega(y) := \begin{cases} \text{the normal cone of } \Omega \text{ if } y \in \Omega, \\ \emptyset \text{ otherwise} \end{cases}$$

*Received by the editors September 15, 1999; accepted for publication (in revised form) April 2, 2001; published electronically September 7, 2001. This work was partly supported by the Pacific Institute for the Mathematical Sciences and by an NSERC research grant.
http://www.siam.org/journals/sicon/40-3/36171.html

†Centre for Experimental and Constructive Mathematics, Simon Fraser University, 8888 University Dr., Burnaby, British Columbia, Canada V5A 1S6 (lucet@cecm.sfu.ca).

‡Department of Mathematics and Statistics, University of Victoria, Victoria, BC, Canada V8W 3P4 (janeye@Math.UVic.CA).

is the normal cone operator.

Let $(\bar{x}, \bar{y})$ be an optimal solution of the OPVIC. If $N_\Omega(y)$ is single-valued and smooth, then the generalized equation constraint (2) would reduce to an ordinary equation $0 = F(x, y) + N_\Omega(y)$. Moreover, if all problem data are smooth and there is no abstract constraint, then the Fritz John necessary optimality condition can be stated as follows. There exist scalar $\lambda \geq 0$ and the vectors (γ, β, η) not all zero such that

$$\begin{cases} 0 = \lambda \nabla f(\bar{x}, \bar{y}) + \nabla \Psi(\bar{x}, \bar{y})^\top \gamma + \nabla H(\bar{x}, \bar{y})^\top \beta + \nabla F(\bar{x}, \bar{y})^\top \eta + \{0\} \times \nabla N_\Omega(\bar{y})^\top \eta, \\ \gamma \geq 0, \text{ and } \langle \Psi(\bar{x}, \bar{y}), \gamma \rangle = 0, \end{cases}$$

where ∇ denotes the usual gradient and $A^\top$ denotes the transpose of a matrix A. In general, however, the map $y \rightrightarrows N_\Omega(y)$ is a set-valued map. Naturally, the usual gradient $\nabla N_\Omega(\bar{y})$ has to be replaced by some kinds of derivatives of set-valued maps.

The Kuhn–Tucker-type necessary conditions with the transpose of the usual gradient ∇N_Ω replaced by the Mordukhovich coderivative $D^* N_\Omega$ were first derived in Ye and Ye [24] under the so-called pseudo-upper-Lipschitz condition for the case of no inequality, no equality constraints, and an abstract constraint in x only. They were further studied under the strong regularity condition in the sense of Robinson and the generalized Mangasarian–Fromovitz constraint qualifications by Outrata in [14] in the case of complementarity constraints and constraints in x only. The first order theory including the necessary optimality conditions involving the Mordukhovich coderivative, various constraint qualifications and their relationships for the general setting of this paper was given in Ye [23]. (Although the equality constraint $H(x, y) = 0$ was not considered explicitly there, the general results under the presence of an equality constraint still hold without any difficulty.) In Ye [22] the Kuhn–Tucker-type necessary conditions with the proximal coderivative for the case of optimization problems with complementarity constraints (OPCCs) were also studied . For recent developments and references on other optimality conditions and computational algorithms, the reader is referred to recent monographs of Luo, Pang, and Ralph [8] and Outrata, Kočvara, and Zowe [15].

In this paper we continue the study by considering the value function $V(p, q, r)$ associated with the right-hand side perturbations

$$\begin{array}{lll} \mathrm{GP}(p,q,r) & \text{minimize} & f(x,y) \\ & \text{subject to} & \Psi(x,y) \leq p, H(x,y) = q, (x,y) \in C, \\ & & r \in F(x,y) + N_\Omega(y), \end{array} \tag{3}$$

i.e.,

$$\begin{aligned} V(p,q,r) := \inf\{ f(x,y) : \quad & \Psi(x,y) \leq p, H(x,y) = q, (x,y) \in C \\ & r \in F(x,y) + N_\Omega(y)\}, \end{aligned}$$

where by convention $\inf \emptyset := +\infty$.

Our main result shows that as in sensitivity analysis for ordinary nonlinear programming (NLP) problems, under certain growth hypotheses, the value function V is lower semicontinuous near 0, and the limiting subdifferentials of the value functions are contained in the negative of the multiplier sets, i.e.,

$$\partial V(0) \subseteq -M^1(\Sigma), \tag{4}$$

$$\partial^\infty V(0) \subseteq -M^0(\Sigma), \tag{5}$$

where Σ is the set of solutions of GP and $M^\lambda(\Sigma)$ is the set of index λ CD multipliers for problem GP, which is the set of vectors (γ, β, η) satisfying the Fritz John necessary condition stated above with the transpose of the usual gradient ∇N_Ω replaced by the Mordukhovich coderivative $D^* N_\Omega$ in the case of smooth problem data and no abstract constraints.

In the case of $M^0(\Sigma) = \{0\}$, (5) implies that the singular limiting subgradient $\partial^\infty V(0)$ contains only the zero vector, and hence the value function is Lipschitz continuous near 0. Moreover, if the optimal solution is unique, if the set of abnormal multipliers $M^0(\Sigma)$ contains only the zero vector, and if the set of Kuhn–Tucker multipliers $M^1(\Sigma)$ is a singleton ζ, then inclusion (4) implies that the value function is smooth and $\nabla V(0) = -\zeta$.

In the case where $\Omega = R^m_+, C = R^{n+m}$, OPVIC reduces to the following OPCC.

$$\begin{array}{lll} \text{(OPCC)} & \text{minimize} & f(x,y), \\ & \text{subject to} & \Psi(x,y) \le 0, H(x,y) = 0, \\ & & y \ge 0,\ F(x,y) \ge 0, \langle y, F(x,y)\rangle = 0. \end{array}$$

In this case (when all functions involved are smooth), an index λ CD multiplier set corresponding to a feasible solution $(\bar{x}, \bar{y})$ denoted by $M^\lambda_{CD}(\bar{x}, \bar{y})$ consists of $(\gamma, \beta, \eta) \in R^d \times R^l \times R^m$ such that

$$0 = \lambda \nabla f(\bar{x}, \bar{y}) + \nabla \Psi(\bar{x}, \bar{y})^\top \gamma + \nabla H(\bar{x}, \bar{y})^\top \beta + \nabla F(\bar{x}, \bar{y})^\top \eta + (0, \xi), \tag{6}$$

$$\gamma \ge 0 \text{ and } \langle \Psi(\bar{x}, \bar{y}), \gamma \rangle = 0, \tag{7}$$

$$\xi_i = 0 \qquad \text{if } \bar{y}_i > 0 \text{ and } F_i(\bar{x}, \bar{y}) = 0, \tag{8}$$

$$\eta_i = 0 \qquad \text{if } \bar{y}_i = 0 \text{ and } F_i(\bar{x}, \bar{y}) > 0, \tag{9}$$

and

$$\text{either } \xi_i < 0,\ \eta_i < 0, \text{or } \xi_i \eta_i = 0 \qquad \text{if } \bar{y}_i = 0 \text{ and } F_i(\bar{x}, \bar{y}) = 0.$$

We call vectors $(\gamma, \beta, \eta) \in R^d \times R^l \times R^m$ satisfying (6)–(9) and

$$\xi_i \eta_i \ge 0 \qquad \text{if } \bar{y}_i = 0 \text{ and } F_i(\bar{x}, \bar{y}) = 0$$

an index λ C-multiplier set and denote it by $M^\lambda_C(\bar{x}, \bar{y})$, and we call those satisfying (6)–(9) and

$$\xi_i \le 0, \eta_i \le 0 \qquad \text{if } \bar{y}_i = 0 \text{ and } F_i(\bar{x}, \bar{y}) = 0$$

an index λ S-multiplier set and denote it by $M^\lambda_S(\bar{x}, \bar{y})$.

Under certain growth hypotheses, we show that the value function

$$\begin{aligned} V(p,q,r) := \{ f(x,y) : & \Psi(x,y) \le p, H(x,y) = q, \\ & y \ge 0, F(x,y) - r \ge 0, \langle y, F(x,y) - r \rangle = 0 \} \end{aligned}$$

is lower semicontinuous near 0 and

$$\partial V(0) \subseteq -M^1 \quad \partial^\infty V(0) \subseteq -M^0,$$

where

$$\begin{aligned} M^1 &= M^1_{CD}(\Sigma), M^1_C(\Sigma), M^1_S(\Sigma), \text{ or } \{(\gamma, \beta, \mu\bar{y} - r^F) : (\gamma, \beta, r^F, r^y, \mu) \in M^1_{NLP}(\Sigma)\}, \\ M^0 &= M^0_{CD}(\Sigma), M^0_C(\Sigma), M^0_S(\Sigma), \text{ or } \{(\gamma, \beta, \mu\bar{y} - r^F) : (\gamma, \beta, r^F, r^y, \mu) \in M^0_{NLP}(\Sigma)\}, \end{aligned}$$

where $M^{\lambda}_{NLP}(\bar{x},\bar{y})$ is the set of index λ ordinary NLP multipliers when the OPCC is treated as an ordinary NLP problem.

Moreover, we show that the above multiplier sets can be ordered as follows:

$$\{(\gamma,\beta,\mu\bar{y}-r^F) : (\gamma,\beta,r^F,r^y,\mu)\in M^{\lambda}_{NLP}(\Sigma)\}\subseteq M^{\lambda}_{S}(\Sigma)\subseteq M^{\lambda}_{CD}(\Sigma)\subseteq M^{\lambda}_{C}(\Sigma).$$

It is obvious that one should use the smallest multiplier sets as possible. However, the smaller multiplier sets may be empty and hence may not provide any information on the properties of the value function. We show that under reasonable constraint qualifications such as the generalized Mangasarian–Fromovitz constraint qualification and the strongly regular constraint qualification, the abnormal CD multiplier set contains only the zero vector, and the set of normal CD multipliers is nonempty. An example is given to show that in sensitivity analysis the CD multipliers may provide more useful information than the other multipliers. In this example, the value function is Lipschitz, and the limiting subdifferentials of the value function coincide with the set of negative CD multipliers, while the limiting subdifferentials are contained strictly in the set of negative C multipliers and the set of P multipliers, NLP multipliers, and S multipliers are empty. Applications to the bilevel programming problem are also given.

In this paper we deal only with the sensitivity analysis of the optimal values. For the sensitivity analysis of the optimal solutions, the reader is referred to Scheel and Scholtes [19].

The following notations are used throughout the paper: B denotes the open unit ball; $B(\bar{z};\delta)$ denotes the open ball centered at $\bar{z}$ with radius $\delta>0$. For a set E, co$E$ denotes the convex hull of E, and int$E$ and cl$E$ denote the interior and the closure of E, respectively. The notation $\langle a,b\rangle$ denotes the inner products of vectors a and b. For a differentiable function f, $\nabla f(\bar{x})$ denotes the gradient of f at $\bar{x}$. For a vector $a\in R^n$, a_i denotes the ith component of a. For an m by n matrix A and index sets $I\subseteq\{1,2,\dots,m\}, J\subseteq\{1,2,\dots,n\}$, A_I and $A_{I,J}$ denote the submatrix of A with rows specified by I and the submatrix of A with rows and columns specified by I and J, respectively. $A^{\top}$ denotes the transpose of a matrix A. For a vector $d\in R^m$, d_I is the subvector composed from the components $d_i, i\in I$.

2. Preliminaries. The purpose of this section is to provide the background material on nonsmooth analysis which will be used later. We give only concise definitions and facts that will be needed in the paper. For more detailed information on the subject, our references are Clarke [3], Loewen [7], Rockafellar and Wets [18], and Mordukhovich [10, 12, 13].

First we give some definitions for various subdifferentials and normal cones.

DEFINITION 2.1. *Let $f: R^n\to R\cup\{+\infty\}$ be lower semicontinuous and finite at $\bar{x}\in R^n$. The proximal subdifferential of f at $\bar{x}$ is the set defined by*

$$\begin{aligned}\partial^{\pi} f(\bar{x}) = \{v\in R^n :&\exists M>0,\delta>0 \text{ s.t.}\\ &f(x)\ge f(\bar{x})+\langle v,x-\bar{x}\rangle + M\|x-\bar{x}\|^2 \ \ \forall x\in\bar{x}+\delta B\},\end{aligned}$$

the limiting subdifferential of f at $\bar{x}$ is the set defined by

$$\partial f(\bar{x}) := \left\{v\in R^n : v=\lim_{\nu\to\infty} v^{\nu} \text{ with } v^{\nu}\in\partial^{\pi} f(x^{\nu}) \text{ and } x^{\nu}\to\bar{x}\right\},$$

the singular limiting subdifferential of f at $\bar{x}$ is the set defined by

$$\partial^{\infty} f(\bar{x}) := \left\{v\in R^n : v=\lim_{\nu\to\infty} \lambda^{\nu}v^{\nu} \text{ with } v^{\nu}\in\partial^{\pi} f(x^{\nu}) \text{ and } \lambda^{\nu}\downarrow 0, x^{\nu}\to\bar{x}\right\}.$$

Let $f : R^n \to R$ be Lipschitz near $\bar{x} \in R^n$. The Clarke generalized gradient of f at $\bar{x}$ is the set

$$\partial_C f(\bar{x}) := clco\partial f(\bar{x}).$$

For set-valued maps, the definition for a limiting normal cone leads to the definition of the coderivative of a set-valued map introduced by Mordukhovich in [9].

DEFINITION 2.2. *For a closed set $C \subset R^n$ and $\bar{x} \in C$, the proximal normal cone to C at $\bar{x}$ is defined by*

$$N_C^\pi(\bar{x}) := \{v \in R^n : \exists M > 0 \text{ s.t. } \langle v, x - \bar{x}\rangle \le M\|x - \bar{x}\|^2 \quad \forall x \in C\},$$

and the limiting normal cone to C at $\bar{x}$ is defined by

$$N_C(\bar{x}) := \left\{ \lim_{\nu\to\infty} v^\nu : v^\nu \in N_C^\pi(x^\nu), x^\nu \to \bar{x} \right\}.$$

DEFINITION 2.3. *Let $\Phi : R^n \Rightarrow R^q$ be a set-valued map. Let $(\bar{x}, \bar{p}) \in clGph\Phi$, where $Gph\Phi := \{(x, p) \ : \ p \in \Phi(x)\}$ is the graph of the set-valued map Φ. The set-valued map $D^*\Phi(\bar{x}, \bar{p})$ from R^q into R^n, defined by*

$$D^*\Phi(\bar{x}, \bar{p})(\eta) := \{\xi \in R^n \ : \ (\xi, -\eta) \in N_{Gph\Phi}(\bar{x}, \bar{p})\},$$

is called the Mordukhovich coderivative of Φ at $(\bar{x}, \bar{p})$.

In general, we have the following inclusions, which may be strict:

$$\partial^\pi f(\bar{x}) \subseteq \partial f(\bar{x}) \subseteq \partial_C f(\bar{x}).$$

In the case where f is a convex function, all subdifferentials coincide with the subdifferentials in the sense of convex analysis, i.e.,

$$\partial^\pi f(\bar{x}) = \partial f(\bar{x}) = \partial_C f(\bar{x}) = \{\zeta : f(x) - f(\bar{x}) \ge \langle\, \zeta, x - \bar{x}\rangle \quad \forall x\}.$$

In the case where f is strictly differentiable (see the definition, e.g., in Clarke [2]), we have

$$\partial f(\bar{x}) = \partial_C f(\bar{x}) = \{\nabla f(\bar{x})\}.$$

The following facts about the subdifferentials are well known.

PROPOSITION 2.4.

(i) *A function $f : R^n \to R$ is Lipschitz near $\bar{x}$ and $\partial f(\bar{x}) = \{\zeta\}$ if and only if f is strictly differentiable at $\bar{x}$ and the gradient of f at $\bar{x}$ equals ζ.*

(ii) *A function $f : R^n \to R$ is Lipschitz near $\bar{x}$ if and only if $\partial^\infty f(\bar{x}) = \{0\}$.*

(iii) *If a function $f : R^n \to R$ is Lipschitz near $\bar{x}$ with positive constant L_f, then $\partial f(\bar{x}) \subseteq L_f clB$.*

The following calculus rules will be useful and can be found in the references given in the beginning of this section.

PROPOSITION 2.5 (see, e.g., [7, Proposition 5A.4]). *Let $f : R^n \to R$ be Lipschitz near $\bar{x}$, and let $g : R^n \to R \cup \{+\infty\}$ be lower semicontinuous and finite at $\bar{x}$. Then*

$$\partial(f + g)(\bar{x}) \subseteq \partial f(\bar{x}) + \partial g(\bar{x}),$$
$$\partial^\infty(f + g)(\bar{x}) \subseteq \partial^\infty g(\bar{x}).$$

PROPOSITION 2.6 (see, e.g., [7, Lemma 5A.3]). *Let $f : R^n \times R^m \to R \cup \{+\infty\}$ be lower semicontinuous and finite at $(\bar{x}, \bar{y})$. If $(\zeta, 0) \in \partial^\infty f(\bar{x}, \bar{y})$ implies that $\zeta = 0$, then*

$$\partial_y f(\bar{x}, \bar{y}) \subseteq \{\eta : (\zeta, \eta) \in \partial f(\bar{x}, \bar{y}) \quad \text{for some } \zeta\},$$
$$\partial_y^\infty f(\bar{x}, \bar{y}) \subseteq \{\eta : (\zeta, \eta) \in \partial^\infty f(\bar{x}, \bar{y}) \quad \text{for some } \zeta\}.$$

PROPOSITION 2.7 (see, e.g., [13, Theorem 7.6]). *Let the minimum function be*

$$(\wedge f_j)(x) := \min\{f_j(x) | j = 1, 2, \ldots, m\},$$

where $f_j : R^n \to R \cup \{+\infty\}$. Assume that f_j are lower semicontinuous around $\bar{x}$ for $j \in J(\bar{x})$ and lower semicontinuous at $\bar{x}$ for $j \notin J(\bar{x})$, where

$$J(x) := \{j | f_j(x) = \wedge f_j(x)\}.$$

Then the minimum function $\wedge f_j(x)$ is lower semicontinuous around $\bar{x}$ and

$$\partial(\wedge f_j)(\bar{x}) \subseteq \bigcup\{\partial f_j(\bar{x}) | j \in J(\bar{x})\},$$
$$\partial^\infty(\wedge f_j)(\bar{x}) \subseteq \bigcup\{\partial^\infty f_j(\bar{x}) | j \in J(\bar{x})\}.$$

Classical results on the value function can be found in [2, 4, 7, 11, 18], while the results we quote are from [7].

PROPOSITION 2.8 (see [7, (b) and (d) of Theorem 5A.2]). *Let $g : R^n \times R^m \to R \cup \{+\infty\}$ be lower semicontinuous everywhere and finite at $(\bar{z}, \bar{\alpha})$. Suppose g is bounded below on some set $E \times O$, where E is a compact neighborhood of $\bar{z}$ and O is an open set containing $\bar{\alpha}$. Define the value function $V : R^m \to R \cup \{+\infty\}$ and the set of minimizers Σ as follows:*

$$V(\alpha) := \inf\{g(z, \alpha) : z \in E\},$$
$$\Sigma(\alpha) := \{z \in E : g(z, \alpha) = V(\alpha)\}.$$

If $\Sigma(\bar{\alpha}) \subseteq \text{int}E$, then the value function V is lower semicontinuous on O, and the subdifferentials of V satisfy these estimates:

$$\partial V(\bar{\alpha}) \subseteq \{\eta \in R^m : (0, \eta) \in \partial g(z, \bar{\alpha}) \quad \text{for some } z \in \Sigma(\bar{\alpha})\},$$
$$\partial^\infty V(\bar{\alpha}) \subseteq \{\eta \in R^m : (0, \eta) \in \partial^\infty g(z, \bar{\alpha}) \quad \text{for some } z \in \Sigma(\bar{\alpha})\}.$$

Our results are stated using the limiting subdifferentials. Alternatively, they could be derived by using the Fréchet subdifferentials instead of the proximal subdifferentials. (Both lead to the same limiting subdifferentials in finite dimensional spaces.) In [18] arguments are given in favor of the former (called there the regular subdifferentials). In the present paper we use the proximal subdifferentials to provide the same framework as in [23].

3. Main results. Let $(\bar{x}, \bar{y})$ be a feasible solution of the OPVIC and let λ be a nonnegative number. We define $M^\lambda(\bar{x}, \bar{y})$, the index λ CD multiplier set corresponding to $(\bar{x}, \bar{y})$, to be the set of vectors (γ, β, η) in $R^d \times R^l \times R^m$ satisfying the Fritz John-type necessary optimality condition involving the Mordukhovich coderivatives for GP, that is, the vectors (γ, β, η) such that

$$\begin{cases} 0 \in \lambda \partial f(\bar{x}, \bar{y}) + \partial\langle \Psi, \gamma\rangle(\bar{x}, \bar{y}) + \partial\langle H, \beta\rangle(\bar{x}, \bar{y}) + \partial\langle F, \eta\rangle(\bar{x}, \bar{y}) \\ \qquad + \{0\} \times D^* N_\Omega(\bar{y}, -F(\bar{x}, \bar{y}))(\eta) + N_C(\bar{x}, \bar{y}), \\ \gamma \geq 0, \text{ and } \langle \Psi(\bar{x}, \bar{y}), \gamma\rangle = 0. \end{cases}$$

Then by Ye [23, Theorem 3.1], the Fritz John-type necessary optimality condition involving the Mordukhovich coderivatives can be rephrased as follows.

PROPOSITION 3.1. *Under the basic assumption* (BA), *if* $(\bar{x}, \bar{y})$ *is a local solution of* OPVIC, *then either the set of normal* CD *multipliers is nonempty or there is a nonzero abnormal* CD *multiplier, i.e.,*

$$M^1(\bar{x}, \bar{y}) \cup (M^0(\bar{x}, \bar{y}) \setminus \{0\}) \neq \emptyset.$$

Note that by the definition of the Mordukhovich coderivative,

$$\xi \in D^* N_\Omega(\bar{y}, -F(\bar{x}, \bar{y}))(\eta) \text{ if and only if } (\xi, -\eta) \in N_{GphN_\Omega}(\bar{y}, -F(\bar{x}, \bar{y})).$$

In the case where $\Omega = \{0\}$, OPVIC reduces to an ordinary mathematical programming problem with equality, inequality, and abstract constraints. The term $D^* N_\Omega(\bar{y}, -F(\bar{x}, \bar{y}))(\eta)$ vanishes, and the above Fritz John condition can be considered as a limiting subdifferential version of the generalized Lagrange multiplier rule as found in Clarke [2, Theorem 6.1.1] and was obtained by Mordukhovich [9, Theorem 1(b)].

In the case where $\Omega = R^m_+$, (1) reduces to a complementarity constraint,

$$y \geq 0,\ F(x, y) \geq 0,\ \langle F(x, y), y \rangle = 0,$$

and the coderivative of the normal cone to the set R^m_+ can be calculated using the following lemma whose proof follows from [22, Proposition 2.7] and the definition of the limiting normal cones.

LEMMA 3.2. *For any* $(\bar{u}, -\bar{v}) \in GphN_{R^m_+}$,

$$\begin{aligned} N_{GphN_{R^m_+}}(\bar{u}, -\bar{v}) = \{(\xi, -\eta) \in R^{2m} :& \xi_i = 0 \text{ if } \bar{u}_i > 0, \bar{v}_i = 0, \\ & \eta_i = 0 \text{ if } \bar{u}_i = 0, \bar{v}_i > 0, \\ & \text{either } \xi_i \eta_i = 0 \text{ or } \xi_i < 0 \text{ and } \eta_i < 0 \text{ if } \bar{u}_i = 0, \bar{v}_i = 0\}. \end{aligned}$$

In the case where Ω is a polyhedral convex set, one can calculate the Mordukhovich coderivative of the normal cone to the set Ω by using the formula of the limiting normal cone to the graph of the normal cone to the set Ω, which was first given in the proof of Dontchev and Rockafellar [5, Theorem 2] and stated in Poliquin and Rockafellar [16, Proposition 4.4].

We first consider the following additively (right-hand side) perturbed GP:

$$\begin{array}{lll} \text{GP}(p, q, r) & \text{minimize} & f(x, y) \\ & \text{subject to} & \Psi(x, y) \leq p, H(x, y) = q, (x, y) \in C, \\ & & r \in F(x, y) + N_\Omega(y), \end{array}$$

with the solution set denoted by $\Sigma(p, q, r)$.

In order to obtain useful information on the subdifferentials of the value function at $(\bar{p}, \bar{q}, \bar{r})$, some hypotheses are usually made for $\text{GP}(p, q, r)$, where (p, q, r) are sufficiently close to the point of interest $(\bar{p}, \bar{q}, \bar{r})$ (see, for example, [4, Growth Hypothesis 3.1.1], [2, Hypothesis 6.5.1], [18, Definition 1.8]). In this paper, we make the following growth hypothesis [7, Theorem 5A.2]:

(GH) at $(\bar{p}, \bar{q}, \bar{r})$: There exists $\delta > 0$ such that the set

$$\begin{aligned} \{(x, y) \in C\ :\ & \Psi(x, y) \leq p, H(x, y) = q, r \in F(x, y) + N_\Omega(y), f(x, y) \leq M, \\ & (p, q, r) \in B(\bar{p}, \bar{q}, \bar{r}; \delta)\} \end{aligned}$$

is bounded for each M.

In order to apply Proposition 2.8, we rewrite the value function in the following form:

$$V(p,q,r) = \inf g(x,y,p,q,r),$$

where g is the extended-value function defined by

$$g(x,y,p,q,r) := f(x,y) + I_{(Gph\Phi)\cap(C\times R^{d+l+m})}(x,y,p,q,r)$$

with I_E being the indicator function of a set E defined by

$$I_E(x) := \begin{cases} 0 \text{ if } x \in E, \\ \infty \text{ if } x \notin E \end{cases}$$

and Φ being the set-valued map defined by

$$\Phi(x,y) = (\Psi(x,y), H(x,y), F(x,y)) + R^d_+ \times \{0\} \times N_\Omega(y).$$

The growth hypothesis (GH) amounts to saying the function g is level-bounded in (x,y) uniformly for any $(p,q,r) \in B(\bar p,\bar q,\bar r;\delta)$. Hence by virtue of [18, Theorem 1.9], $\bigcup_{(p,q,r)\in B(\bar p,\bar q,\bar r;\delta)} \Sigma(p,q,r)$ is a compact set and for all $(p,q,r) \in B(\bar p,\bar q,\bar r;\delta)$,

$$V(p,q,r) = \inf\{g(x,y,p,q,r) : (x,y) \in E\},$$

where E is a compact set with interior containing $\bigcup_{(p,q,r)\in B(\bar p,\bar q,\bar r;\delta)} \Sigma(p,q,r)$. It is clear that g is lower semicontinuous everywhere and finite at any $(x,y,p,q,r) \in (Gph\Phi) \cap (C \times R^{d+l+m})$. Since f is Lipschitz on E, g is bounded below on $E \times B(\bar x,\bar y;\epsilon)$. The following result then follows immediately by applying Proposition 2.8.

PROPOSITION 3.3. *Under the basic assumption* (BA) *and the growth hypothesis* (GH) *at* $(\bar p,\bar q,\bar r)$ *the value function* V *is lower semicontinuous on* $B(\bar p,\bar q,\bar r;\delta)$ *and*

$$\partial V(\bar p,\bar q,\bar r) \subseteq \bigcup_{(\bar x,\bar y)\ \in\Sigma(\bar p,\bar q,\bar r)} \{(u,v,w)\ :\ (0,0,u,v,w) \in \partial g(\bar x,\bar y,\bar p,\bar q,\bar r)\}, \tag{10}$$

$$\partial^\infty V(\bar p,\bar q,\bar r) \subseteq \bigcup_{(\bar x,\bar y)\in\Sigma(\bar p,\bar q,\bar r)} \{(u,v,w)\ :\ (0,0,u,v,w) \in \partial^\infty g(\bar x,\bar y,\bar p,\bar q,\bar r)\}. \tag{11}$$

We now prove that the set in the right-hand side of (10) (respectively, (11)) is included in the normal multiplier set M^1 (respectively, the abnormal multiplier set M^0).

By the sum rule (see Proposition 2.5) and the fact that for any closed set E with $\bar z \in E$

$$\partial I_E(\bar z) = \partial^\infty I_E(\bar z) = N_E(\bar z),$$

we have

$$\begin{aligned} \partial g(\bar x,\bar y,\bar p,\bar q,\bar r) &\subset \partial f(\bar x,\bar y) \times \{(0,0)\} + N_{(Gph\Phi)\cap(C\times R^{d+l+m})}(\bar x,\bar y,\bar p,\bar q,\bar r), \\ \partial^\infty g(\bar x,\bar y,\bar p,\bar q,\bar r) &\subset N_{(Gph\Phi)\cap(C\times R^{d+l+m})}(\bar x,\bar y,\bar p,\bar q,\bar r). \end{aligned}$$

Hence we need only to compute the normal cone.

LEMMA 3.4. *If* $(s_x,s_y,s_p,s_q,s_r) \in N_{(Gph\Phi)\cap(C\times R^{d+l+m})}(\bar x,\bar y,\bar p,\bar q,\bar r)$, *then*

$$\begin{cases} (s_x,s_y) \in \partial\langle\Psi,-s_p\rangle(\bar x,\bar y) + \partial\langle H,-s_q\rangle(\bar x,\bar y) + \partial\langle F,-s_r\rangle(\bar x,\bar y) + N_C(\bar x,\bar y) \\ \qquad\qquad +\{0\} \times D^*N_\Omega(\bar y,\bar r - F(\bar x,\bar y))(-s_r), \\ s_p \ge 0, \ \textit{and}\ \langle\Psi(\bar x,\bar y) - \bar p, s_p\rangle = 0. \end{cases}$$

Proof. Step 1. Let $(\tilde{x},\tilde{y},\tilde{p},\tilde{q},\tilde{r})$ be any point in a neighborhood of $(\bar{x},\bar{y},\bar{p},\bar{q},\bar{r})$ on which Ψ, H, and F are Lipschitz continuous and

$$(s_x,s_y,s_p,s_q,s_r)\in N^{\pi}_{(Gph\Phi)\cap(C\times R^{d+l+m})}(\tilde{x},\tilde{y},\tilde{p},\tilde{q},\tilde{r}).$$

By definition of the proximal normal cone, there is $M>0$ such that for all $(x,y,p,q,r)\in (Gph\Phi)\cap(C\times R^{d+l+m})$,

$$\langle (s_x,s_y,s_p,s_q,s_r),(x,y,p,q,r)-(\tilde{x},\tilde{y},\tilde{p},\tilde{q},\tilde{r})\rangle \le M\|(x,y,p,q,r)-(\tilde{x},\tilde{y},\tilde{p},\tilde{q},\tilde{r})\|^2.$$

In other words, $(\tilde{x},\tilde{y},\tilde{p},\tilde{q},\tilde{r})$ is a solution to the optimization problem

$$\begin{array}{ll}\text{minimize} & \langle -(s_x,s_y,s_p,s_q,s_r),(x,y,p,q,r)\rangle + M\|(x,y,p,q,r)-(\tilde{x},\tilde{y},\tilde{p},\tilde{q},\tilde{r})\|^2\\ \text{subject to} & \Psi(x,y)\le p, H(x,y)=q,(x,y)\in C,\\ & r\in F(x,y)+N_\Omega(y).\end{array}$$

We now prove that the only abnormal CD multiplier for the above problem is the zero vector. Indeed, the set of abnormal CD multipliers at $(\tilde{x},\tilde{y},\tilde{p},\tilde{q},\tilde{r})$ for the above problem are the vectors (γ,β,η) satisfying

$$\begin{cases}0\in \partial\langle\Psi,\gamma\rangle(\tilde{x},\tilde{y})\times\{(-\gamma,0,0)\}+\partial\langle H,\beta\rangle(\tilde{x},\tilde{y})\times\{(0,-\beta,0)\}+\partial\langle F,\eta\rangle(\tilde{x},\tilde{y})\\ \times\{(0,0,-\eta)\}+\{0\}\times D^*N_\Omega(\tilde{y},\tilde{r}-F(\tilde{x},\tilde{y}))(\eta)\times\{(0,0,0)\}+N_C(\tilde{x},\tilde{y})\\ \times\{(0,0,0)\},\ \gamma\ge 0,\ \text{and}\ \langle\Psi(\tilde{x},\tilde{y})-\tilde{p},\gamma\rangle=0,\end{cases}$$

which obviously coincides with the set $\{(0,0,0)\}$. Applying Proposition 3.1, we conclude that the set of normal CD multipliers for the above problem must be nonempty. That is, there are vectors $\eta\in R^m,\beta\in R^l$, and $\gamma\in R^d$ such that

$$\begin{cases}0\in -\{(s_x,s_y,s_p,s_q,s_r)\}+\partial\langle\Psi,\gamma\rangle(\tilde{x},\tilde{y})\times\{(-\gamma,0,0)\}+\partial\langle H,\beta\rangle(\tilde{x},\tilde{y})\times\{(0,-\beta,0)\}\\ \quad+\partial\langle F,\eta\rangle(\tilde{x},\tilde{y})\times\{(0,0,-\eta)\}+\{0\}\times D^*N_\Omega(\tilde{y},\tilde{r}-F(\tilde{x},\tilde{y}))(\eta)\times\{(0,0,0)\}\\ \qquad\qquad +N_C(\tilde{x},\tilde{y})\times\{(0,0)\},\\ \gamma\ge 0,\ \text{and}\ \langle\Psi(\tilde{x},\tilde{y})-\tilde{p},\gamma\rangle=0.\end{cases}$$

That is,

$$\begin{cases}(s_x,s_y)\in \partial\langle\Psi,-s_p\rangle(\tilde{x},\tilde{y})+\partial\langle H,-s_q\rangle(\tilde{x},\tilde{y})\\ \qquad +\partial\langle F,-s_r\rangle(\tilde{x},\tilde{y})+\{0\}\times D^*N_\Omega(\tilde{y},\tilde{r}-F(\tilde{x},\tilde{y}))(-s_q)+N_C(\tilde{x},\tilde{y}),\\ s_p\ge 0,\ \text{and}\ \langle\Psi(\tilde{x},\tilde{y})-\tilde{p},s_p\rangle=0.\end{cases}$$

Step 2. Now take any $(s_x,s_y,s_p,s_q,s_r)\in N_{(Gph\Phi)\cap(C\times R^{d+l+m})}(\bar{x},\bar{y},\bar{p},\bar{q},\bar{r})$. Then by definition of limiting normal cones, there are sequences $(x^\nu,y^\nu,p^\nu,q^\nu,r^\nu)\to(\bar{x},\bar{y},\bar{p},\bar{q},\bar{r})$ and $(s_x^\nu,s_y^\nu,s_p^\nu,s_q^\nu,s_r^\nu)\to(s_x,s_y,s_p,s_q,s_r)$ with

$$(s_x^\nu,s_y^\nu,s_p^\nu,s_q^\nu,s_r^\nu)\in N^{\pi}_{(Gph\Phi)\cap(C\times R^{d+l+m})}(x^\nu,y^\nu,p^\nu,q^\nu,r^\nu).$$

By virtue of step 1,

$$\begin{cases}(s_x^\nu,s_y^\nu)\in \partial\langle\Psi,-s_p^\nu\rangle(x^\nu,y^\nu)+\partial\langle H,-s_q^\nu\rangle(x^\nu,y^\nu)+\partial\langle F,-s_r^\nu\rangle(x^\nu,y^\nu)\\ \qquad +\{0\}\times D^*N_\Omega(y^\nu,r^\nu-F(x^\nu,y^\nu))(-s_r^\nu)+N_C(x^\nu,y^\nu),\\ s_p^\nu\ge 0,\ \text{and}\ \langle\Psi(x^\nu,y^\nu)-p^\nu,s_p^\nu\rangle=0.\end{cases}$$

Since Ψ is Lipschitz near $(\bar{x}, \bar{y})$, we have

$$\begin{aligned}\partial\langle\Psi, -s_p^\nu\rangle(x^\nu, y^\nu) &\subseteq \partial\langle\Psi, -s_p\rangle(x^\nu, y^\nu) + \partial\langle\Psi, s_p - s_p^\nu\rangle(x^\nu, y^\nu) \text{ by Proposition 2.5}\\ &\subseteq \partial\langle\Psi, -s_p\rangle(x^\nu, y^\nu) + \|s_p^\nu - s_p\| L_\Psi clB \text{ by Proposition 2.4,}\end{aligned}$$

where L_Ψ is the Lipschitz constant of Ψ. Similarly,

$$\begin{aligned}\partial\langle H, -s_q^\nu\rangle(x^\nu, y^\nu) &\subseteq \partial\langle H, -s_q\rangle(x^\nu, y^\nu) + \|s_q^\nu - s_q\| L_H clB,\\ \partial\langle F, -s_r^\nu\rangle(x^\nu, y^\nu) &\subseteq \partial\langle F, -s_r\rangle(x^\nu, y^\nu) + \|s_r^\nu - s_r\| L_F clB,\end{aligned}$$

where L_H, L_F are the Lipschitz constants of F and H. Hence we have

$$\begin{cases}(s_x^\nu, s_y^\nu) \in \partial\langle\Psi, -s_p\rangle(x^\nu, y^\nu) + \partial\langle H, -s_q\rangle(x^\nu, y^\nu) + \partial\langle F, -s_r\rangle(x^\nu, y^\nu)\\ \qquad + (\|s_p^\nu - s_p\| + \|s_q^\nu - s_q\| + \|s_r^\nu - s_r\|)(L_\Psi + L_H + L_F) clB\\ \qquad + \{0\} \times D^* N_\Omega(y^\nu, r^\nu - F(x^\nu, y^\nu))(-s_r^\nu) + N_C(x^\nu, y^\nu),\\ s_p^\nu \geq 0, \text{ and } \langle\Psi(x^\nu, y^\nu) - p^\nu, s_p^\nu\rangle = 0.\end{cases}$$

Taking limits as $\nu \to \infty$ and using the definitions of the limiting normal cone and the limiting subdifferentials completes the proof. □

Remark. As is pointed out by referee 1, alternatively, Lemma 3.4 can also be proved by formulating the constraints in the form of [12, equation (6.19)] and applying [12, Theorem 6.10].

All in all, we proved the following result.

THEOREM 3.5. *Assume* (GH) *and* (BA) *hold. Then the value function V is lower semicontinuous on $B(\bar{p}, \bar{q}, \bar{r}; \delta)$ and*

$$\partial V(\bar{p}, \bar{q}, \bar{r}) \subset \bigcup_{(\bar{x}, \bar{y}) \in \Sigma(\bar{p}, \bar{q}, \bar{r})} -M^1(\bar{x}, \bar{y}) \text{ and } \partial^\infty V(\bar{p}, \bar{q}, \bar{r}) \subset \bigcup_{(\bar{x}, \bar{y}) \in \Sigma(\bar{p}, \bar{q}, \bar{r})} -M^0(\bar{x}, \bar{y}).$$

We now consider the value function $V(\alpha)$ associated with the following perturbed GP:

$$\begin{array}{lll}\text{GP}(\alpha) & \text{minimize} & f(x, y, \alpha)\\ & \text{subject to} & \Psi(x, y, \alpha) \leq 0, H(x, y, \alpha) = 0, (x, y) \in C,\\ & & 0 \in F(x, y, \alpha) + N_\Omega(y),\end{array}$$

i.e.,

$$\begin{aligned}V(\alpha) := \inf\{f(x, y, \alpha) : \quad &\Psi(x, y, \alpha) \leq 0, H(x, y, \alpha) = 0, (x, y) \in C,\\ &0 \in F(x, y, \alpha) + N_\Omega(y)\},\end{aligned}$$

where the following basic assumptions are satisfied:

(BH) The functions $f : R^{n+m+c} \to R, \Psi : R^{n+m+c} \to R^d, H : R^{n+m+c} \to R^l$, and $F : R^{n+m+c} \to R^m$ are locally Lipschitz near any points in $C \times R^c$; C is a closed subset of R^{n+m}; and Ω is a closed convex subset of R^m.

It is easy to see that we can turn the nonadditive perturbations into additive perturbations by adding an auxiliary variable:

$$\begin{array}{lll}\text{GP}(\alpha) & \text{minimize} & f(x, y, z)\\ & \text{subject to} & \Psi(x, y, z) \leq 0, H(x, y, z) = 0, (x, y, z) \in C \times R^c,\\ & & 0 \in F(x, y, z) + N_\Omega(y),\\ & & z = \alpha,\end{array}$$

which is the partially perturbed problem of the fully perturbed problem

$$\begin{array}{lll} \mathrm{GP}(p,q,r,\alpha) & \text{minimize} & f(x,y,z) \\ & \text{subject to} & \Psi(x,y,z)\le p, H(x,y,z)=q, (x,y,z)\in C\times R^c, \\ & & r\in F(x,y,z)+N_\Omega(y), \\ & & z=\alpha. \end{array}$$

By Theorem 3.5, if the fully perturbed problem $\mathrm{GP}(p,q,r,\alpha)$ satisfies the growth hypothesis (GH) at $(0,0,0,\bar\alpha)$, then the value function $\widetilde V(p,q,r,\alpha)$ defined by

$$\begin{aligned}\widetilde V(p,q,r,\alpha) := \inf\{f(x,y,z) :&\Psi(x,y,z)\le p, H(x,y,z)=q, (x,y,z)\in C\times R^c,\\ &r\in F(x,y,z)+N_\Omega(y), z=\alpha\}\end{aligned}$$

is lower semicontinuous on $B(0,0,0,\bar\alpha;\delta)$ and

$$\begin{aligned}\partial\widetilde V(0,0,0,\bar\alpha)&\subseteq \bigcup_{(\bar x,\bar y,\bar\alpha)\in\Sigma(0,0,0,\bar\alpha)} -M^1(\bar x,\bar y,\bar\alpha),\\ \partial^\infty\widetilde V(0,0,0,\bar\alpha)&\subseteq \bigcup_{(\bar x,\bar y,\bar\alpha)\in\Sigma(0,0,0,\bar\alpha)} -M^0(\bar x,\bar y,\bar\alpha).\end{aligned}$$

For any $(0,0,0,\zeta)\in\partial^\infty\widetilde V(0,0,0,\bar\alpha)$, we have $(0,0,0,\zeta)\in -M^0(\bar x,\bar y,\bar\alpha)$ for some point $(\bar x,\bar y,\bar\alpha)\in\Sigma(0,0,0,\bar\alpha)$. Therefore,

$$(0,0,\zeta)\in N_C(\bar x,\bar y)\times\{0\},$$

which implies that $\zeta=0$. By Proposition 2.6, we have

$$\begin{aligned}\partial_\alpha\widetilde V(0,0,0,\bar\alpha)&\subseteq\{-\zeta : -(\gamma,\beta,\eta,\zeta)\in\partial\widetilde V(0,0,0,\bar\alpha)\text{ for some }(\gamma,\beta,\eta)\},\\ \partial^\infty_\alpha\widetilde V(0,0,0,\bar\alpha)&\subseteq\{-\zeta : -(\gamma,\beta,\eta,\zeta)\in\partial^\infty\widetilde V(0,0,0,\bar\alpha)\text{ for some }(\gamma,\beta,\eta)\}.\end{aligned}$$

Moreover, since all functions involved are continuous, it suffices to fix α at $\bar\alpha$ in the growth hypothesis (GH) at $(0,0,0,\bar\alpha)$ for the fully perturbed problem $\mathrm{GP}(p,q,r,\alpha)$. Consequently, noticing that $V(\alpha)=\widetilde V(0,0,0,\alpha)$, we have proved the following theorem.

THEOREM 3.6. *In addition to the basic assumption* (BH)*, assume that there exists $\delta>0$ such that the set*

$$\begin{aligned}\{(x,y)\in C\ :\ &\Psi(x,y,\bar\alpha)\le p, H(x,y,\bar\alpha)=q, r\in F(x,y,\bar\alpha)+N_\Omega(y), f(x,y,\bar\alpha)\le M,\\ &(p,q,r)\in B(0;\delta)\}\end{aligned}$$

is bounded for each M. Then the value function $V(\alpha)$ is lower semicontinuous near $\bar\alpha$ and

$$\begin{aligned}\partial V(\bar\alpha)&\subseteq \bigcup_{(\bar x,\bar y)\in\Sigma(\bar\alpha)}\{-\zeta : (\gamma,\beta,\eta,\zeta)\in M^1(\bar x,\bar y,\bar\alpha)\},\\ \partial^\infty V(\bar\alpha)&\subseteq \bigcup_{(\bar x,\bar y)\in\Sigma(\bar\alpha)}\{-\zeta : (\gamma,\beta,\eta,\zeta)\in M^0(\bar x,\bar y,\bar\alpha)\},\end{aligned}$$

where $M^\lambda(\bar x,\bar y,\bar\alpha)$ is the set of index λ multipliers for problem $\mathrm{GP}(p,q,r,\alpha)$ *at $(0,0,0,\bar\alpha)$, i.e., vectors $(\gamma,\beta,\eta,\zeta)$ in $R^d\times R^l\times R^m\times R$ satisfying*

$$\begin{cases} 0\in\lambda\partial f(\bar x,\bar y,\bar\alpha)+\partial\langle\Psi,\gamma\rangle(\bar x,\bar y,\bar\alpha)+\partial\langle H,\beta\rangle(\bar x,\bar y,\bar\alpha)+\partial\langle F,\eta\rangle(\bar x,\bar y,\bar\alpha)\\ \qquad +\{0\}\times D^*N_\Omega(\bar y,-F(\bar x,\bar y,\bar\alpha))(\eta)\times\{0\}+\{(0,0,\zeta)\}+N_C(\bar x,\bar y)\times\{0\},\\ \gamma\ge 0,\ \text{and}\ \langle\Psi(\bar x,\bar y,\bar\alpha),\gamma\rangle=0,\end{cases}$$

and $\Sigma(\bar{\alpha})$ is the set of solutions of problem GP($\bar{\alpha}$).

The above estimates may not be useful in the case where $\partial V(\bar{\alpha})$ is empty. The following consequence of Theorem 3.6 and Proposition 2.4 provides conditions which rule out this possibility.

COROLLARY 3.7. *Under the assumption of Theorem* 3.6*, if the set of ζ components of the abnormal* CD *multiplier set contains only the zero vector, i.e.,*

$$\bigcup_{(\bar{x},\bar{y})\in\Sigma(\bar{\alpha})} \{-\zeta : (\gamma,\beta,\eta,\zeta) \in M^0(\bar{x},\bar{y},\bar{\alpha})\} = \{0\},$$

then $V(\bar{\alpha})$ is finite and Lipschitz near $\bar{\alpha}$ with

$$\emptyset \neq \partial V(\bar{\alpha}) \subset \bigcup_{(\bar{x},\bar{y})\in\Sigma(\bar{\alpha})} \{-\zeta : (\gamma,\beta,\eta,\zeta) \in M^1(\bar{x},\bar{y},\bar{\alpha})\}.$$

In addition to the above assumptions, if the ζ components of the normal CD multiplier set are unique, i.e.,

$$\bigcup_{(\bar{x},\bar{y})\in\Sigma(\bar{\alpha})} \{-\zeta : (\gamma,\beta,\eta,\zeta) \in M^1(\bar{x},\bar{y},\bar{\alpha})\} = \{-\zeta\},$$

then V is strictly differentiable at $\bar{\alpha}$ and $\nabla V(\bar{\alpha}) = -\zeta$.

In the case where all functions are smooth, the estimates have the following simple expression.

COROLLARY 3.8. *In addition to the assumptions in Theorem* 3.6*, assume that f, Ψ, H, F are C^1 at each $(\bar{x},\bar{y},\bar{\alpha})$, where $(\bar{x},\bar{y}) \in \Sigma(\bar{\alpha})$; then the value function V is lower semicontinuous near $\bar{\alpha}$, and*

$$\begin{aligned}
\partial V(\bar{\alpha}) \subseteq \bigcup_{(\bar{x},\bar{y})\in\Sigma(\bar{\alpha})} \{&\nabla_\alpha f(\bar{x},\bar{y},\bar{\alpha}) + \nabla_\alpha \Psi(\bar{x},\bar{y},\bar{\alpha})^\top \gamma + \nabla_\alpha H(\bar{x},\bar{y},\bar{\alpha})^\top \beta \\
&+\nabla_\alpha F(\bar{x},\bar{y},\bar{\alpha})^\top \eta : \quad (\gamma,\beta,\eta) \in M^1(\bar{x},\bar{y})\}, \\
\partial^\infty V(\bar{\alpha}) \subseteq \bigcup_{(\bar{x},\bar{y},\bar{z})\in\Sigma(\bar{\alpha})} \{&\nabla_\alpha \Psi(\bar{x},\bar{y},\bar{\alpha})^\top \gamma + \nabla_\alpha H(\bar{x},\bar{y},\bar{\alpha})^\top \beta \\
&+\nabla_\alpha F(\bar{x},\bar{y},\bar{\alpha})^\top \eta : \quad (\gamma,\beta,\eta) \in M^0(\bar{x},\bar{y})\},
\end{aligned}$$

where $M^\lambda(\bar{x},\bar{y})$ is the set of index λ CD *multipliers for problem* GP($\bar{\alpha}$).

Note that in the case where there are no variational inequality constraints, the CD multipliers are the ordinary NLP multipliers, and the above results recover the well-known results in the sensitivity analysis of NLP.

4. Applications to OPCCs. In this section, we apply our main results to the following perturbed OPCC:

$$\begin{array}{lll}
\text{(OPCC)}(\alpha) & \text{minimize} & f(x,y,\alpha), \\
& \text{subject to} & \Psi(x,y,\alpha) \le 0, H(x,y,\alpha) = 0, (x,y) \in C, \\
& & y \ge 0,\ F(x,y,\alpha) \ge 0, \\
& & \langle y, F(x,y,\alpha)\rangle = 0,
\end{array}$$

which is GP(α) with $\Omega = R^m_+$.

For easier exposition, we assume in this section that all problem data f, Ψ, H, F are C^1. We denote by $\nabla f(x,y,\alpha)$ the gradient of function f with respect to (x,y).

For $(\bar{x},\bar{y})$, a feasible solution of (OPCC)$(\bar{\alpha})$, we define the index sets

$$\begin{aligned} L := L(\bar{x},\bar{y}) &:= \{1 \le i \le m : \bar{y}_i > 0, F_i(\bar{x},\bar{y},\bar{\alpha}) = 0\},\\ I_+ := I_+(\bar{x},\bar{y}) &:= \{1 \le i \le m : \bar{y}_i = 0, F_i(\bar{x},\bar{y},\bar{\alpha}) > 0\},\\ I_0 := I_0(\bar{x},\bar{y}) &:= \{1 \le i \le m : \bar{y}_i = 0, F_i(\bar{x},\bar{y},\bar{\alpha}) = 0\}. \end{aligned}$$

4.1. Sensitivity analysis of the value function via NLP multipliers. Let $(\bar{x},\bar{y})$ be a local optimal solution for (OPCC)$(\bar{\alpha})$. Treating (OPCC)$(\bar{\alpha})$ as an ordinary NLP problem with inequality constraints

$$\Psi(x,y,\bar{\alpha}) \le 0, y \ge 0, F(x,y,\bar{\alpha}) \ge 0,$$

equality constraints

$$H(x,y,\bar{\alpha}) = 0, \langle y, F(x,y,\bar{\alpha}\rangle = 0,$$

and the abstract constraint $(x,y) \in C$, it is easy to see that the Fritz John optimality condition implies the existence of $\lambda \ge 0, \gamma \in R^d, \beta \in R^l, r^F \in R^m, r^y \in R^m, \mu \in R$, not all zero, such that

$$\begin{aligned} &0 \in \lambda\nabla f(\bar{x},\bar{y},\bar{\alpha}) + \nabla\Psi(\bar{x},\bar{y},\bar{\alpha})^\top\gamma + \nabla H(\bar{x},\bar{y},\bar{\alpha})^\top\beta\\ &\quad -\nabla F(\bar{x},\bar{y},\bar{\alpha})^\top r^F - \{(0,r^y)\} + \mu\nabla\langle y,F\rangle(\bar{x},\bar{y},\bar{\alpha}) + N_C(\bar{x},\bar{y}),\\ &\gamma \ge 0, \langle\gamma, \Psi(\bar{x},\bar{y},\bar{\alpha})\rangle = 0,\\ &r^F \ge 0, r^y \ge 0, \langle r^F, F(\bar{x},\bar{y},\bar{\alpha})\rangle = 0, \langle r^y,\bar{y}\rangle = 0. \end{aligned}$$

Using the sum and product rules, we have

$$\nabla\langle y,F\rangle(\bar{x},\bar{y},\bar{\alpha}) = \{(0,F(\bar{x},\bar{y},\bar{\alpha}))\} + \nabla F(\bar{x},\bar{y},\bar{\alpha})^\top\bar{y}.$$

Therefore, the Fritz John necessary condition becomes

$$\begin{aligned} &0 \in \lambda\nabla f(\bar{x},\bar{y},\bar{\alpha}) + \nabla\Psi(\bar{x},\bar{y},\bar{\alpha})^\top\gamma + \nabla H(\bar{x},\bar{y},\bar{\alpha})^\top\beta\\ &\quad +\nabla F(\bar{x},\bar{y},\bar{\alpha})^\top(\mu\bar{y} - r^F) + \{(0,\mu F(\bar{x},\bar{y},\bar{\alpha}) - r^y)\} + N_C(\bar{x},\bar{y}),\\ &\gamma \ge 0, \langle\gamma,\Psi(\bar{x},\bar{y},\bar{\alpha})\rangle = 0,\\ &r^F \ge 0, r^y \ge 0, \text{ and } \langle r^F, F(\bar{x},\bar{y},\bar{\alpha})\rangle = 0, \langle r^y,\bar{y}\rangle = 0. \end{aligned}$$

Definition 4.1 (NLP multipliers). *We call all vectors $(\gamma,\beta,r^F,r^y,\mu) \in R^d \times R^l \times R^m \times R^m \times R$ satisfying the above Fritz John necessary condition for any $\lambda \ge 0$ the index λ NLP multipliers for OPCC$(\bar{\alpha})$ and denote the set by $M^\lambda_{NLP}(\bar{x},\bar{y})$.*

Since we treat OPCC(α) as an ordinary NLP problem, $\Omega = \{0\}$ in the corresponding problem GP(α). Hence the CD multipliers for the corresponding $GP(\alpha)$ are the NLP multipliers defined above. Applying Corollary 3.8 and Proposition 2.4, we derive the following upper estimates of the limiting subdifferentials of the value function in terms of the NLP multipliers.

Theorem 4.2. *Assume that there exists $\delta > 0$ such that the set*

$$\begin{aligned} \{(x,y) \in C :&(p,q,p_y,p_F,q_\mu) \in B(0;\delta), \Psi(x,y,\bar{\alpha}) \le p, H(x,y,\bar{\alpha}) = q,\\ &y \ge p_y,\ F(x,y,\bar{\alpha}) \le p_F, \langle y, F(x,y,\bar{\alpha})\rangle = q_\mu, f(x,y,\bar{\alpha}) \le M\} \end{aligned}$$

is bounded for each M. Then the value function V is lower semicontinuous near $\bar{\alpha}$ and

$$\partial V(\bar{\alpha}) \subseteq \bigcup_{(\bar{x},\bar{y})\in\Sigma(\bar{\alpha})} \{\nabla_\alpha f(\bar{x},\bar{y},\bar{\alpha}) + \nabla_\alpha \Psi(\bar{x},\bar{y},\bar{\alpha})^\top \gamma + \nabla_\alpha H(\bar{x},\bar{y},\bar{\alpha})^\top \beta$$

$$+\nabla_\alpha F(\bar{x},\bar{y},\bar{\alpha})^\top(\mu\bar{y} - r^F) : (\gamma,\beta,r^F,r^y,\mu) \in M^1_{NLP}(\bar{x},\bar{y})\}, \tag{12}$$

$$\partial^\infty V(\bar{\alpha}) \subseteq \bigcup_{(\bar{x},\bar{y})\in\Sigma(\bar{\alpha})} \{\nabla_\alpha \Psi(\bar{x},\bar{y},\bar{\alpha})^\top \gamma + \nabla_\alpha H(\bar{x},\bar{y},\bar{\alpha})^\top \beta$$

$$+\nabla_\alpha F(\bar{x},\bar{y},\bar{\alpha})^\top(\mu\bar{y} - r^F) : (\gamma,\beta,r^F,r^y,\mu) \in M^0_{NLP}(\bar{x},\bar{y})\}. \tag{13}$$

If the set in the right-hand side of inclusion (13) *contains only the zero vector, then the value function V is Lipschitz near $\bar{\alpha}$. If the set in the right-hand side of inclusion* (13) *contains only the zero vector and the set in the right-hand side of inclusion* (12) *is a singleton, then the value function is strictly differentiable at $\bar{\alpha}$.*

4.2. Sensitivity analysis of the value function via CD multipliers. Since OPCC($\bar{\alpha}$) is OPVIC($\bar{\alpha}$) with $\Omega = R^m_+$, the following expression of CD multiplers follows immediately from Lemma 3.2.

PROPOSITION 4.3. *For* OPCC($\bar{\alpha}$)*, an index* λ CD *multiplier corresponding to a feasible solution $(\bar{x},\bar{y})$ is a vector $(\gamma,\beta,\eta) \in R^d \times R^l \times R^m$ such that*

$$0 \in \lambda\nabla f(\bar{x},\bar{y},\bar{\alpha}) + \nabla\Psi(\bar{x},\bar{y},\bar{\alpha})^\top\gamma + \nabla H(\bar{x},\bar{y},\bar{\alpha})^\top\beta$$

$$+\nabla F(\bar{x},\bar{y},\bar{\alpha})^\top\eta + (0,0,\xi) + N_C(\bar{x},\bar{y}), \tag{14}$$

$$\gamma \geq 0 \text{ and } \langle\Psi(\bar{x},\bar{y},\bar{\alpha}),\gamma\rangle = 0, \tag{15}$$

$$\xi_i = 0 \qquad \text{if } \bar{y}_i > 0 \text{ and } F_i(\bar{x},\bar{y},\bar{\alpha}) = 0, \tag{16}$$

$$\eta_i = 0 \qquad \text{if } \bar{y}_i = 0 \text{ and } F_i(\bar{x},\bar{y},\bar{\alpha}) > 0, \tag{17}$$

$$\text{either } \xi_i < 0,\ \eta_i < 0,\ \text{or } \xi_i\eta_i = 0 \qquad \text{if } \bar{y}_i = 0 \text{ and } F_i(\bar{x},\bar{y}) = 0. \tag{18}$$

Corollary 3.8 and Proposition 2.4 now lead to the following result.

THEOREM 4.4. *Assume that there exists $\delta > 0$ such that the set*

$$\{(x,y) \in C : (p,q,r) \in B(0;\delta), \Psi(x,y,\bar{\alpha}) \leq p, H(x,y,\bar{\alpha}) = q,$$
$$y \geq 0,\ F(x,y,\bar{\alpha}) \geq r, \langle y, F(x,y,\bar{\alpha}) - r\rangle = 0, f(x,y,\bar{\alpha}) \leq M\}$$

is bounded for each M. Then the value function V is lower semicontinuous near $\bar{\alpha}$ and

$$\partial V(\bar{\alpha}) \subseteq \bigcup_{(\bar{x},\bar{y})\in\Sigma(\bar{\alpha})} \{\nabla_\alpha f(\bar{x},\bar{y},\bar{\alpha}) + \nabla_\alpha \Psi(\bar{x},\bar{y},\bar{\alpha})^\top \gamma + \nabla_\alpha H(\bar{x},\bar{y},\bar{\alpha})^\top \beta$$

$$+\nabla_\alpha F(\bar{x},\bar{y},\bar{\alpha})^\top\eta : (\gamma,\beta,\eta) \in M^1_{CD}(\bar{x},\bar{y})\}, \tag{19}$$

$$\partial^\infty V(\bar{\alpha}) \subseteq \bigcup_{(\bar{x},\bar{y})\in\Sigma(\bar{\alpha})} \{\nabla_\alpha \Psi(\bar{x},\bar{y},\bar{\alpha})^\top \gamma + \nabla_\alpha H(\bar{x},\bar{y},\bar{\alpha})^\top \beta$$

$$+\nabla F(\bar{x},\bar{y},\bar{\alpha})^\top\eta : (\gamma,\beta,\eta) \in M^0_{CD}(\bar{x},\bar{y})\}. \tag{20}$$

If the set in the right-hand side of inclusion (20) *contains only the zero vector, then the value function V is Lipschitz near $\bar{\alpha}$. If the set in the right-hand side of inclusion* (20) *contains only the zero vector and the set in the right-hand side of inclusion* (19) *is a singleton, then the value function is strictly differentiable at $\bar{\alpha}$.*

We say that the generalized Mangasarian–Fromovitz constraint qualification for OPCC($\bar{\alpha}$) is satisfied at $(\bar{x},\bar{y})$ if $C = D \times R^m$ and

(i) for every partition of I_0 into sets P, Q, R with $R \neq \emptyset$, there exist vectors $k \in \text{int}T_C(\bar{x}, D), h \in R^m$ such that $h_{I_+} = 0, h_Q = 0, h_R \geq 0$,

$$\begin{aligned}
&\nabla_x \Psi_{I(\Psi)}(\bar{x}, \bar{y}, \bar{\alpha})k + \nabla_y \Psi_{I(\Psi)}(\bar{x}, \bar{y}, \bar{\alpha})h \leq 0,\\
&\nabla_x H(\bar{x}, \bar{y}, \bar{\alpha})k + \nabla_y H(\bar{x}, \bar{y}, \bar{\alpha})h = 0,\\
&\nabla_x F_{L\cup P}(\bar{x}, \bar{y}, \bar{\alpha})k + \nabla_y F_{L\cup P}(\bar{x}, \bar{y}, \bar{\alpha})h = 0,\\
&\nabla_x F_R(\bar{x}, \bar{y}, \bar{\alpha})k + \nabla_y F_R(\bar{x}, \bar{y}, \bar{\alpha})h \geq 0,
\end{aligned}$$

and either $h_i > 0$ or

$$\nabla_x F_i(\bar{x}, \bar{y}, \bar{\alpha})k + \nabla_y F_i(\bar{x}, \bar{y}, \bar{\alpha})h > 0 \text{ for some } i \in R;$$

(ii) for every partition of I_0 into the sets P, Q, the matrix

$$\begin{bmatrix} \nabla_x H(\bar{x}, \bar{y}, \bar{\alpha}) & \nabla_y H_{A,L\cup P}(\bar{x}, \bar{y}, \bar{\alpha}) \\ \nabla_x F_{L\cup P}(\bar{x}, \bar{y}, \bar{\alpha}) & \nabla_y F_{L\cup P, L\cup P}(\bar{x}, \bar{y}, \bar{\alpha}) \end{bmatrix}$$

has full row rank and there exist vectors $k \in \text{int}T_C(\bar{x}, D), h \in R^m$ such that

$$\begin{aligned}
&h_{I_+} = 0, h_Q = 0,\\
&\nabla_x \Psi_{I(\Psi)}(\bar{x}, \bar{y}, \bar{\alpha})k + \nabla_y \Psi_{I(\Psi)}(\bar{x}, \bar{y}, \bar{\alpha})h < 0,\\
&\nabla_x H(\bar{x}, \bar{y}, \bar{\alpha})k + \nabla_y H(\bar{x}, \bar{y}, \bar{\alpha})h = 0,\\
&\nabla_x F_{L\cup P}(\bar{x}, \bar{y}, \bar{\alpha})k + \nabla_y F_{L\cup P}(\bar{x}, \bar{y}, \bar{\alpha})h = 0,
\end{aligned}$$

where $A := \{1, \ldots, l\}$, $T_C(\bar{x}, D)$ denotes the Clarke tangent cone of D at $\bar{x}$, and $I(\Psi) := \{i : \Psi_i(\bar{x}, \bar{y}) = 0\}$ is the index set of the binding inequality constraints.

In [23, Proposition 4.5] it was proved that the generalized Mangasarian–Fromovitz constraint qualification implies that the only abnormal CD multiplier is the zero vector. Hence Theorem 4.4 has the following consequence.

COROLLARY 4.5. *In addition to the assumptions of Theorem* 4.4*, if the generalized Mangasarian–Fromovitz constraint qualification as defined above is satisfied for* OPCC($\bar{\alpha}$)*, then $V(\alpha)$ is finite and Lipschitz near $\bar{\alpha}$.*

Another sufficient condition for $M^0_{CD}(\Sigma(\bar{\alpha})) = \{0\}$ is the strong regularity condition in the sense of Robinson [17]. For OPCC ($\bar{\alpha}$), the strong regularity condition has the following form according to [17, Theorem 3.1].

COROLLARY 4.6. *In addition to the assumptions of Theorem* 4.4*, assume that $C = D \times R^m$ for some $D \subseteq R^n$, that there are no inequality constraints, and that the following conditions are satisfied:*

(i) *the matrix*

$$\begin{bmatrix} \nabla_y H_{A,L}(\bar{x}, \bar{y}, \bar{\alpha}) \\ \nabla_y F_{L,L}(\bar{x}, \bar{y}, \bar{\alpha}) \end{bmatrix}$$

is nonsingular, where $A := \{1, \ldots, l\}$;

(ii) *the Schur complement of the above matrix in the matrix*

$$\begin{bmatrix} \nabla_y H_{A,L}(\bar{x}, \bar{y}, \bar{\alpha}) & \nabla_y H_{A,I_0}(\bar{x}, \bar{y}, \bar{\alpha}) \\ \nabla_y F_{L,L}(\bar{x}, \bar{y}, \bar{\alpha}) & \nabla_y F_{L,I_0}(\bar{x}, \bar{y}, \bar{\alpha}) \\ \nabla_y F_{I_0,L}(\bar{x}, \bar{y}, \bar{\alpha}) & \nabla_y F_{I_0,I_0}(\bar{x}, \bar{y}, \bar{\alpha}) \end{bmatrix}$$

has positive principle minors;

then $V(\alpha)$ is finite and Lipschitz near $\bar{\alpha}$.

4.3. Sensitivity analysis of the value function via C multipliers. It is easy to see that OPCC $(\bar{\alpha})$ can be formulated as the following optimization problem with a nonsmooth equation:

$$\begin{array}{lll} \text{OPCC}(\bar{\alpha}) & \text{minimize} & f(x,y,\alpha) \\ & \text{subject to} & \Psi(x,y,\alpha) \le 0, H(x,y,\alpha) = 0, (x,y) \in C, \\ & & \min\{y_i, F_i\}(x,y,\alpha) = 0, \qquad i = 1,2,\ldots,m. \end{array} \tag{21}$$

It can be shown as in Scheel and Scholtes [19, Lemma 1] that a solution of the OPCC is C stationary defined as follows.

DEFINITION 4.7 (C multipliers). *Let $(\bar{x},\bar{y})$ be a feasible point of the* OPCC. *The point $(\bar{x},\bar{y})$ is C stationary if there exist vectors $(\gamma,\beta,\eta,\xi) \in R^d \times R^l \times R^m \times R^m$ satisfying* (14)–(17) *and*

$$\xi_i \eta_i \ge 0 \qquad \text{if } \bar{y}_i = 0 \text{ and } F_i(\bar{x},\bar{y},\bar{\alpha}) = 0.$$

The set of vectors (γ,β,η) satisfying the above condition for some ξ is called the index λ C multiplier set and is denoted by $M_C^{\lambda}(\bar{x},\bar{y})$.

THEOREM 4.8. *Assume that there exists $\delta > 0$ such that the set*

$$\begin{aligned} \{(x,y) \in C :& (p,q,q^m) \in B(0;\delta), \Psi(x,y,\bar{\alpha}) \le p, H(x,y,\bar{\alpha}) = q, \\ & \min\{y_i, F_i(x,y,\bar{\alpha})\} = q_i^m, i = 1,\ldots,m, f(x,y,\bar{\alpha}) \le M\} \end{aligned}$$

is bounded for each M. Then the value function V is lower semicontinuous near $\bar{\alpha}$ and

$$\begin{aligned} \partial V(\bar{\alpha}) \subseteq \bigcup_{(\bar{x},\bar{y}) \in \Sigma(\bar{\alpha})} & \{\nabla_\alpha f(\bar{x},\bar{y},\bar{\alpha}) + \nabla_\alpha \Psi(\bar{x},\bar{y},\bar{\alpha})^\top \gamma + \nabla_\alpha H(\bar{x},\bar{y},\bar{\alpha})^\top \beta \\ & + \nabla_\alpha F(\bar{x},\bar{y},\bar{\alpha})^\top \eta : (\gamma,\beta,\eta) \in M_C^1(\bar{x},\bar{y})\}, \end{aligned} \tag{22}$$

$$\begin{aligned} \partial^\infty V(\bar{\alpha}) \subseteq \bigcup_{(\bar{x},\bar{y}) \in \Sigma(\bar{\alpha})} & \{\nabla_\alpha \Psi(\bar{x},\bar{y},\bar{\alpha})^\top \gamma + \nabla_\alpha H(\bar{x},\bar{y},\bar{\alpha})^\top \beta \\ & + F(\bar{x},\bar{y},\bar{\alpha})^\top \eta : (\gamma,\beta,\eta) \in M_C^0(\bar{x},\bar{y})\}. \end{aligned} \tag{23}$$

If the set in the right-hand side of inclusion (23) *contains only the zero vector, then the value function V is Lipschitz near $\bar{\alpha}$. If the set in the right-hand side of inclusion* (23) *contains only the zero vector and the set in the right-hand side of inclusion* (22) *is a singleton, then the value function is strictly differentiable at $\bar{\alpha}$.*

Proof. By Theorem 3.6, since the growth assumption is satisfied, the value function is lower semicontinuous near $\bar{\alpha}$ and

$$\begin{aligned} \partial V(\bar{\alpha}) &\subseteq \bigcup_{(\bar{x},\bar{y}) \in \Sigma(\bar{\alpha})} \{-\zeta : (\gamma,\eta,\zeta) \in M^1(\bar{x},\bar{y},\bar{\alpha})\}, \\ \partial^\infty V(\bar{\alpha}) &\subseteq \bigcup_{(\bar{x},\bar{y}) \in \Sigma(\bar{\alpha})} \{-\zeta : (\gamma,\eta,\zeta) \in M^0(\bar{x},\bar{y},\bar{\alpha})\}, \end{aligned}$$

where $M^\lambda(\bar{x},\bar{y},\bar{\alpha})$ is the set of vectors $(\gamma,\beta,r,\zeta) \in R^{d+l+m+c}$ such that

$$\begin{aligned} 0 \in{}& \lambda \nabla f(\bar{x},\bar{y},\bar{\alpha}) + \nabla \Psi(\bar{x},\bar{y},\bar{\alpha})^\top \gamma + \nabla H(\bar{x},\bar{y},\bar{\alpha})^\top \beta \\ & + \partial \sum_{i=1}^m r_i \min\{y_i, F_i\}(\bar{x},\bar{y},\bar{\alpha}) + \{(0,0,\zeta)\} + N_C(\bar{x},\bar{y}) \times \{0\}, \\ \gamma \ge 0, & \langle \gamma, \Psi \rangle (\bar{x},\bar{y},\bar{\alpha}) = 0. \end{aligned}$$

Note that, in the above, ∇f denotes the gradient of a function f with respect to (x, y, α). Since

$$\partial \sum_{i=1}^{m} r_i \min\{y_i, F_i\}(\bar{x}, \bar{y}, \bar{\alpha}) \subseteq \sum_{i=1}^{m} r_i \partial_C \min\{y_i, F_i\}(\bar{x}, \bar{y}, \bar{\alpha})$$

and

$$\partial_C \min\{y_i, F_i\}(\bar{x}, \bar{y}, \bar{\alpha}) = \begin{cases} (0, e_i, 0) & \forall i \in I_+, \\ \nabla F_i(\bar{x}, \bar{y}, \bar{\alpha}) & \forall i \in L, \\ \{t(0, , e_i, 0) + (1-t)\nabla F_i(\bar{x}, \bar{y}, \bar{\alpha}) : t \in [0,1]\} & \forall i \in I_0, \end{cases}$$

where e_i is the unit vector whose ith component is 1 and those other components are zero, there exist γ, β, η such that

$$\zeta = \lambda \nabla_\alpha f(\bar{x}, \bar{y}, \bar{\alpha}) + \nabla_\alpha \Psi(\bar{x}, \bar{y}, \bar{\alpha})^\top \gamma + \nabla_\alpha H(\bar{x}, \bar{y}, \bar{\alpha})^\top \beta + \nabla_\alpha F(\bar{x}, \bar{y}, \bar{\alpha})^\top \eta$$

and

$$\begin{aligned} 0 \in\ & \lambda \nabla f(\bar{x}, \bar{y}, \bar{\alpha}) + \nabla \Psi(\bar{x}, \bar{y}, \bar{\alpha})^\top \gamma + \nabla H(\bar{x}, \bar{y}, \bar{\alpha})^\top \beta + \nabla F(\bar{x}, \bar{y}, \bar{\alpha})^\top \eta \\ & + (0, \xi) + N_C(\bar{x}, \bar{y}), \\ & \gamma \geq 0, \langle \Psi, \gamma \rangle (\bar{x}, \bar{y}, \bar{\alpha}) = 0, \end{aligned}$$

where

$$\begin{array}{ll} \eta_i = 0 & \forall i \in I_+, \\ \xi_i = 0 & \forall i \in L, \\ \eta_i = r_i(1 - \bar{t}_i), \xi_i = r_i \bar{t}_i \text{ for some } \bar{t}_i \in [0,1], & \forall i \in I_0. \end{array}$$

It is then easy to see that

$$\forall i \in I_0, \eta_i \xi_i \geq 0.$$

Hence (γ, β, η) is a C multiplier, and the proof of the theorem is complete. □

4.4. Sensitivity analysis via P multipliers and S multipliers. Taking the "piecewise programming" approach, for any given index set $\nu \subseteq I := \{1, \ldots, m\}$, we consider the subproblem associated with ν:

$$\begin{array}{lll} \text{OPCC}(\alpha)_\nu & \text{minimize} & f(x, y, \alpha) \\ & \text{subject to} & \Psi(x, y, \alpha) \leq 0, H(x, y, \alpha) = 0, (x, y) \in C, \\ & & y_i \geq 0,\ F_i(x, y, \alpha) = 0 \quad \forall i \in \nu \\ & & y_i = 0,\ F_i(x, y, \alpha) \geq 0 \quad \forall i \in I \backslash \nu. \end{array}$$

As suggested by referee 2, since the value function is the minimum of the value functions for the subproblems, i.e.,

$$V(\alpha) = \min_{\nu \subset I} V_\nu(\alpha)$$

and

$$V(\bar{\alpha}) = V_\nu(\bar{\alpha})\ \forall \nu = L(\bar{x}, \bar{y}) \cup \sigma, \sigma \subseteq I_0(\bar{x}, \bar{y}), (\bar{x}, \bar{y}) \in \Sigma(\bar{\alpha}),$$

applying the calculus for the minimum functions in Proposition 2.7, we conclude that the value function V is lower semicontinuous if each $V_\nu(\alpha), \nu = L(\bar{x},\bar{y}) \cup \sigma, \sigma \subseteq I_0(\bar{x},\bar{y}), (\bar{x},\bar{y}) \in \Sigma(\bar{\alpha})$, is lower semicontinuous and the following inclusion holds:

$$\partial^\infty V(\bar{\alpha}) \subseteq \{\partial^\infty V_\nu(\bar{\alpha}) : \nu = L(\bar{x},\bar{y}) \cup \sigma, \sigma \subseteq I_0(\bar{x},\bar{y}), (\bar{x},\bar{y}) \in \Sigma(\bar{\alpha})\}, \tag{24}$$

$$\partial V(\bar{\alpha}) \subseteq \{\partial V_\nu(\bar{\alpha}) : \nu = L(\bar{x},\bar{y}) \cup \sigma, \sigma \subseteq I_0(\bar{x},\bar{y}), (\bar{x},\bar{y}) \in \Sigma(\bar{\alpha})\}. \tag{25}$$

The Fritz John condition for the subproblem $\mathrm{OPCC}(\bar{\alpha})_\nu$ with

$$\nu = L(\bar{x},\bar{y}) \cup \sigma, \sigma \subseteq I_0(\bar{x},\bar{y}), (\bar{x},\bar{y}) \in \Sigma(\bar{\alpha})$$

implies the existence of vectors $(\gamma,\beta,\eta,\xi) \in R^d \times R^a \times R^b \times R^b$ satisfying (14)–(17) and

$$\xi_\sigma \le 0, \eta_{I_0\setminus\sigma} \le 0. \tag{26}$$

Definition 4.9 (P multipliers). *The set of all vectors (γ,β,η) satisfying the above Fritz John condition at $(\bar{x},\bar{y})$ is denoted by $M^\lambda_\sigma(\bar{x},\bar{y})$, and $\bigcup_{\sigma\subseteq I_0} M^\lambda_\sigma(\bar{x},\bar{y})$ is called the set of P multipliers.*

Applying Corollary 3.8, we have the following result.

Proposition 4.10. *For any $(\bar{x},\bar{y}) \in \Sigma(\bar{\alpha})$ and any given index set $\sigma \subseteq I_0(\bar{x},\bar{y})$, assume that there exists $\delta > 0$ such that the set*

$$\begin{aligned}\{(x,y) \in C :&(p,q,q^y,q^F,) \in B(0;\delta), \Psi(x,y,\bar{\alpha}) \le p, H(x,y,\bar{\alpha}) = q,\\ &y_i \ge q^y_i,\ F_i(x,y,\bar{\alpha}) = q^F_i \quad \forall i \in \nu := \sigma \cup L(\bar{x},\bar{y}),\\ &y_i = q^y_i,\ F_i(x,y,\bar{\alpha}) \ge q^F_i \quad \forall i \in I\setminus\nu, f(x,y,\bar{\alpha}) \le M\}\end{aligned}$$

is bounded for each M. Then the value function for subproblem $\mathrm{OPCC}(\bar{\alpha})_\nu$ with $\nu = L(\bar{x},\bar{y}) \cup \sigma$ is lower semicontinuous near $\bar{\alpha}$ and

$$\begin{aligned}\partial V_\nu(\bar{\alpha}) \subseteq \bigcup_{(\bar{x},\bar{y})\in\Sigma_\nu(\bar{\alpha})} &\{\nabla_\alpha f(\bar{x},\bar{y},\bar{\alpha}) + \nabla_\alpha \Psi(\bar{x},\bar{y},\bar{\alpha})^\top \gamma + \nabla_\alpha H(\bar{x},\bar{y},\bar{\alpha})^\top \beta\\ &+\nabla_\alpha F(\bar{x},\bar{y},\bar{\alpha})^\top \eta : (\gamma,\beta,\eta) \in M^1_\sigma(\bar{x},\bar{y})\},\\ \partial^\infty V_\nu(\bar{\alpha}) \subseteq \bigcup_{(\bar{x},\bar{y})\in\Sigma_\nu(\bar{\alpha})} &\{\nabla_\alpha \Psi(\bar{x},\bar{y},\bar{\alpha})^\top \gamma + \nabla_\alpha H(\bar{x},\bar{y},\bar{\alpha})^\top \beta\\ &+\nabla_\alpha F(\bar{x},\bar{y},\bar{\alpha})^\top \eta : (\gamma,\beta,\eta) \in M^0_\sigma(\bar{x},\bar{y})\},\end{aligned}$$

where $\Sigma_\nu(\bar{\alpha})$ denotes the set of solutions for the subproblem $\mathrm{OPCC}\ (\alpha)_\nu$.

We have the following estimates for the value function in terms of P multipliers.

Theorem 4.11. *Assume that there exists $\delta > 0$ such that for $(\bar{x},\bar{y}) \in \Sigma(\bar{\alpha})$ and each index set $\sigma \subseteq I_0(\bar{x},\bar{y})$, the set in Proposition 4.10 is bounded for each M. Then the value function V is lower semicontinuous near $\bar{\alpha}$ and*

$$\begin{aligned}\partial V(\bar{\alpha}) \subseteq \bigcup_{(\bar{x},\bar{y})\in\Sigma(\bar{\alpha})} &\{\nabla_\alpha f(\bar{x},\bar{y},\bar{\alpha}) + \nabla_\alpha \Psi(\bar{x},\bar{y},\bar{\alpha})^\top \gamma + \nabla_\alpha H(\bar{x},\bar{y},\bar{\alpha})^\top \beta\\ &+\nabla_\alpha F(\bar{x},\bar{y},\bar{\alpha})^\top \eta : (\gamma,\beta,\eta) \in \cup_{\sigma\subseteq I_0} M^1_\sigma(\bar{x},\bar{y})\},\end{aligned} \tag{27}$$

$$\begin{aligned}\partial^\infty V(\bar{\alpha}) \subseteq \bigcup_{(\bar{x},\bar{y})\in\Sigma(\bar{\alpha})} &\{\nabla_\alpha \Psi(\bar{x},\bar{y},\bar{\alpha})^\top \gamma + \nabla_\alpha H(\bar{x},\bar{y},\bar{\alpha})^\top \beta\\ &+\nabla_\alpha F(\bar{x},\bar{y},\bar{\alpha})^\top \eta : (\gamma,\beta,\eta) \in \cup_{\sigma\subseteq I_0} M^0_\sigma(\bar{x},\bar{y})\}.\end{aligned} \tag{28}$$

If the set in the right-hand side of inclusion (28) *contains only the zero vector, then the value function V is Lipschitz near $\bar{\alpha}$. If the set in the right-hand side of inclusion* (28) *contains only the zero vector and the set in the right-hand side of inclusion* (27) *is a singleton, then the value function is strictly differentiable at $\bar{\alpha}$.*

DEFINITION 4.12 (S multipliers). *The set of index λ S multipliers, denoted by $M_S^{\lambda}(\bar{x},\bar{y})$, is the set of all vectors $(\gamma,\beta,\eta) \in R^d \times R^a \times R^b$ satisfying* (14)–(17) *and*

$$\xi_i \le 0, \eta_i \le 0 \qquad \text{if } \bar{y}_i = 0 \text{ and } F_i(\bar{x},\bar{y},\bar{\alpha}) = 0.$$

In the following theorem, we give a condition under which the set of P multipliers and S multipliers coincide, and so we have the estimates in terms of the S multipliers.

THEOREM 4.13. *In addition to the assumptions of Theorem* 4.11*, assume that $C = R^n \times R^a \times R^b$ and for all $(\bar{x},\bar{z},\bar{u}) \in \Sigma(\bar{\alpha})$, the partial MPEC linear independence constraint qualification is satisfied, i.e.,*

$$\begin{cases} 0 = \nabla\Psi(\bar{x},\bar{y},\bar{\alpha})^{\top}\gamma + \nabla H(\bar{x},\bar{y},\bar{\alpha})^{\top}\beta + \nabla F(\bar{x},\bar{y},\bar{\alpha})^{\top}\eta + (0,0,\xi), \\ \gamma_{J(\Psi)} = 0,\ \eta_{I_+} = 0,\ \xi_L = 0, \end{cases}$$

implies that $\eta_{I_0} = 0, \xi_{I_0} = 0$, where $J(\Psi) := \{i : \Psi_i(\bar{x},\bar{y},\bar{\alpha}) < 0\}$. Then the value function V is lower semicontinuous near $\bar{\alpha}$ and

$$\begin{aligned} \partial V(\bar{\alpha}) \subseteq \bigcup_{(\bar{x},\bar{y})\in\Sigma(\bar{\alpha})} & \{\nabla_\alpha f(\bar{x},\bar{y},\bar{\alpha}) + \nabla_\alpha\Psi(\bar{x},\bar{y},\bar{\alpha})^{\top}\gamma + \nabla_\alpha H(\bar{x},\bar{y},\bar{\alpha})^{\top}\beta \\ & + \nabla_\alpha F(\bar{x},\bar{y},\bar{\alpha})^{\top}\eta : (\gamma,\beta,\eta) \in M_S^1(\bar{x},\bar{y})\}, \\ \partial^\infty V(\bar{\alpha}) \subseteq \bigcup_{(\bar{x},\bar{y})\in\Sigma(\bar{\alpha})} & \{\nabla_\alpha\Psi(\bar{x},\bar{y},\bar{\alpha})^{\top}\gamma + \nabla_\alpha H(\bar{x},\bar{y},\bar{\alpha})^{\top}\beta \\ & + \nabla_\alpha F(\bar{x},\bar{y},\bar{\alpha})^{\top}\eta : (\gamma,\beta,\eta) \in M_S^0(\bar{x},\bar{y})\}. \end{aligned}$$

Remark. As in the proof of [22, Theorem 3.2], it is easy to see that under the partial MPEC linear independence constraint qualification, all multipliers including the S multiplier, the CD multiplier, the C multiplier, and the P multiplier coincide.

Recently, the MPEC linear independence constraint qualifications have received a lot of attention. It is known that under the MPEC linear independence constraint qualification, the computation of the OPCC is much easier and more efficient (see, e.g., Scholtes [20]). Furthermore, it was shown in Scholtes [21] that the MPEC linear independence constraint qualification is a generic condition for the OPCC. Here we prove the importance of the MPEC linearly independence constraint qualification from the aspect of the sensitivity analysis: the value function is Lipschitz continuous, and it is even strictly differentiable in the case where the optimal solution set is unique. Note that the MPEC linear independence constraint qualification is stronger than the partial MPEC linear independence constraint qualification.

COROLLARY 4.14. *In addition to the assumptions of Theorem* 4.11*, assume that the MPEC linear independence constraint qualifications are satisfied at all $(\bar{x},\bar{y}) \in \Sigma(\bar{\alpha})$, i.e.,*

$$\begin{cases} 0 = \nabla\Psi(\bar{x},\bar{y},\bar{\alpha})^{\top}\gamma + \nabla H(\bar{x},\bar{y},\bar{\alpha})^{\top}\beta + \nabla F(\bar{x},\bar{y},\bar{\alpha})^{\top}\eta + (0,0,\xi), \\ \gamma_{J(\Psi)} = 0, \eta_{I_+} = 0, \xi_L = 0, \end{cases}$$

implies that $\gamma = 0, \beta = 0, \eta = 0, \xi = 0$. Then the value function is Lipschitz continuous near $\bar{\alpha}$. Furthermore, if the set of optimal solutions $\Sigma(\bar{\alpha})$ is a singleton, then the value function V is strictly differentiable at $\bar{\alpha}$.

Proof. The MPEC linear independence constraint qualification obviously implies that $M_S^0(\bar{x},\bar{y}) = \{0\}$ and $M_S^1(\bar{x},\bar{y})$ is a singleton. Hence the conclusion follows from Theorem 4.13 and Proposition 2.4. □

4.5. Relationships between the multipliers for the OPCC. Applying the definitions, it is clear that

$$M_S^\lambda(\bar{x},\bar{y}) \subseteq M_{CD}^\lambda(\bar{x},\bar{y}) \subseteq M_C^\lambda(\bar{x},\bar{y}), \qquad M_S^\lambda(\bar{x},\bar{y}) \subseteq M_P^\lambda(\bar{x},\bar{y}). \tag{29}$$

It is not possible to compare the set of NLP multipliers directly with the other multipliers since the spaces they belong to have different dimensions. However, the following interesting relationships can be obtained.

PROPOSITION 4.15 (relationship between an NLP multiplier and an S multiplier).

$$\{(\gamma,\beta,\mu\bar{y}-r^F) : (\gamma,\beta,r^F,r^y,\mu) \in M_{NLP}^\lambda(\bar{x},\bar{y})\} \subseteq M_S^\lambda(\bar{x},\bar{y})$$

for all $\lambda \geq 0$.

Proof. Let $(\gamma,\beta,r^F,r^y,\mu) \in M_{NLP}^\lambda(\bar{x},\bar{y})$. We consider the following cases.

Case $\bar{y}_i > 0, F_i(\bar{x},\bar{y}) = 0$. Then $r_i^y = 0$. So $\xi_i := \mu F_i - r_i^y = 0$.

Case $\bar{y}_i = 0, F_i(\bar{x},\bar{y}) > 0$. Then $r_i^F = 0$. So $\eta_i = \mu\bar{y}_i - r_i^F = 0$.

Case $\bar{y}_i = 0, F_i(\bar{x},\bar{y}) = 0]$. Then $\xi_i = \mu F_i(\bar{x},\bar{y}) - r_i^y = -r_i^y$ and $\eta_i = \mu\bar{y}_i - r_i^F = -r_i^F$. So $\xi_i = -r_i^y \leq 0$ and $\eta_i = -r_i^F \leq 0$.

Hence (γ,β,η), where $\eta := \mu\bar{y} - r^F$, is an S multiplier, and the proof of the proposition is complete. □

The above relationship indicates that one can arrange the upper estimates of the limiting subdifferentials in Theorems 4.2, 4.13, 4.4, and 4.8 from the smallest to the largest in the order of NLP multipliers, S multipliers, CD multipliers, and C multipliers.

One may try to use the smallest multiplier set in sensitivity analysis. However, the smaller multipler sets tend to require stronger constraint qualifications and hence may be empty. In such a case, where the smaller multipler set is empty, one may have to use the larger multiplier set.

We now use the following example to show that in some cases the smaller multiplier sets such as the NLP and the S multiplier sets may be empty while the CD multiplier provides the tightest bound.

Example. Consider the OPCC

$$(P) \qquad \begin{aligned} \text{minimize} \quad & -y \\ \text{subject to} \quad & x - y = 0, \\ & x \geq 0, y \geq 0, xy = 0, \end{aligned}$$

where $x \in R$ and $y \in R$, and its perturbed problem

$$P(q,r) \qquad \begin{aligned} \text{minimize} \quad & -y \\ \text{subject to} \quad & x - y = q, \\ & x - r \geq 0, y \geq 0, (x-r)y = 0, \end{aligned}$$

which is OPCC (α) with $\alpha = (q,r), f = -y, H = x - y - q, F = x - r$. Let $\bar{\alpha} = (0,0)$. It is clear that the only feasible solution for problem $(P) = P(0,0)$ is $(0,0)$. Hence the only optimal solution for *(P)* is $(0,0)$. The set of index λ NLP multipliers (β, r^y, r^F, μ)

at $(0,0)$ satisfy

$$\begin{cases} 0 = \lambda(0,-1) + \beta(1,-1) - (r^F, 0) - (0, r^y) + \mu(0,0), \\ r^F, r^y \geq 0. \end{cases}$$

It is clear that any $(\beta, r^y, r^F, \mu) = (0,0,0,\mu)$ with $\mu \neq 0$ is a nonzero NLP abnormal multiplier and there is no NLP normal multiplier. Hence $M^0_{NLP}(0,0) = \{(0,0,0)\} \times (-\infty, +\infty) \neq \{(0,0,0,0)\}$ and $M^1_{NLP}(0,0) = \emptyset$.

Since $\bar{y} = 0$ and $F(\bar{x}, \bar{y}, 0) = \bar{x} = 0$, the index λ CD multipliers (β, η) at $(0,0)$ satisfy

$$\begin{aligned} &0 = \lambda(0,-1) + \beta(1,-1) + \eta(1,0) + (0,\xi), \\ &\text{either } \xi < 0, \eta < 0, \text{ or } \xi\eta = 0. \end{aligned}$$

When $\lambda = 0$, the above condition implies that $\beta = \eta = \xi = 0$, while when $\lambda = 1$, either $\eta = 1, \beta = -1, \xi = 0$, or $\beta = \eta = 0, \xi = 1$. So $M^1_{CD}(0,0) = \{(0,0)\} \cup \{(-1,1)\}$ and $M^0_{CD}(0,0) = \{(0,0)\}$.

The set of index λ C multipliers (β, η) at $(0,0)$ satisfy

$$\begin{aligned} &0 = \lambda(0,-1) + \beta(1,-1) + \eta(1,0) + (0,\xi), \\ &\xi\eta \geq 0. \end{aligned}$$

When $\lambda = 0$, the above condition implies that $\beta = \eta = \xi = 0$, while for $\lambda = 1$, $-\beta = \eta \in [0,1]$. So $M^1_C(0,0) = \{(\beta, \eta) : \eta = -\beta \in [0,1]\}$ and $M^0_C(0,0) = \{(0,0)\}$.

Since the optimal solution for (P) is $(\bar{x}, \bar{y}) = (0,0)$, $(0,0)$ is also optimal for the subproblem associated with $\nu = \{1\}$,

$$\begin{aligned} (P_1) \qquad & \text{minimize} && -y \\ & \text{subject to} && x - y = 0, \\ & && y \geq 0, x = 0, \end{aligned}$$

and the subproblem associated with $\nu = \emptyset$,

$$\begin{aligned} (P_2) \qquad & \text{minimize} && -y \\ & \text{subject to} && x - y = 0, \\ & && y = 0, x \geq 0. \end{aligned}$$

The index λ multiplier set for (P_1) consists of vectors (β, η) satisfying

$$\begin{cases} 0 = \lambda(0,-1) + \beta(1,-1) + \eta(1,0) + (0,\xi), \\ \xi \leq 0, \end{cases}$$

and the index λ multiplier set for (P_2) consist of vectors (β, η) satisfying

$$\begin{cases} 0 = \lambda(0,-1) + \beta(1,-1) + \eta(1,0) + (0,\xi), \\ \eta \leq 0. \end{cases}$$

Therefore, the abnormal P multiplier set is

$$\begin{aligned} M^0_P(0,0) = M^0_1(0,0) \cup M^0_2(0,0) &= \{(\beta,\eta) : \beta = -\eta \leq 0\} \cup \{(\beta,\eta) : \beta = -\eta \geq 0\} \\ &= \{(\beta,\eta) : \beta = -\eta\}, \end{aligned}$$

and the normal P multiplier set is

$$\begin{aligned} M_P^1(0,0) = M_1^1(0,0) \cup M_2^1(0,0) &= \{(\beta,\eta) : \beta = -\eta \le -1\} \cup \{(\beta,\eta) : \beta = -\eta \ge 0\} \\ &= \{(\beta,\eta) : \beta = -\eta \in (-\infty,-1] \cup [0,\infty)\}. \end{aligned}$$

The index λ S multiplier set consists of vectors (β,η) satisfying

$$\begin{cases} 0 = \lambda(0,-1) + \beta(1,-1) + \eta(1,0) + (0,\xi), \\ \xi \le 0, \eta \le 0, \end{cases}$$

i.e.,

$$\begin{aligned} &\beta = -\eta,\ \beta = -\lambda + \xi, \\ &\xi \le 0,\ \eta \le 0. \end{aligned}$$

That is, $M_S^0(0,0) = \{0\}$, and $M_S^1(0,0) = \emptyset$.

Consider the value function

$$V(q,r) := \inf\{-y : x - r \ge 0, y \ge 0, (x-r)y = 0, x - y = q\}.$$

Then by Theorem 4.4, since the only abnormal CD multiplier is the zero vector, we conclude that the value function is Lipschitz near $(0,0)$, and

$$\begin{aligned} \emptyset \ne \partial V(0,0) &\subseteq \{\beta(-1,0) + \eta(0,-1) : (\beta,\eta) \in M_{CD}^1(0,0)\} \\ &= -M_{CD}^1(0,0) = \{(0,0)\} \cup \{(1,-1)\}. \end{aligned}$$

In fact, we can easily find the expression for the value function for this simple example since the feasible set of the perturbed problem $P(q,r)$ still reduces to one point. Indeed, we have

$$\begin{cases} \Sigma(q,r) = \{(r, r-q)\} \text{ and } V(q,r) = q - r & \text{if } q < r, \\ \Sigma(q,r) = \{(q,0)\} \text{ and } V(q,r) = 0 & \text{if } q \ge r. \end{cases}$$

So $V(q,r) = \min(0, q-r)$, which is Lipschitz continuous everywhere. By definition of the limiting subdifferentials, it is easy to see that

$$\begin{aligned} \partial V(0,0) &= \{(0,0)\} \cup \{(1,-1)\}, \\ \partial^\infty V(0,0) &= \{(0,0)\}. \end{aligned}$$

Therefore, the inclusions in Theorem 4.4 are actually equalities here, i.e.,

$$\begin{aligned} \partial V(0,0) &= \{(0,0)\} \cup \{(1,-1)\} = -M_{CD}^1(0,0), \\ \partial^\infty V(0,0) &= \{(0,0)\} = -M_{CD}^0(0,0). \end{aligned}$$

Using Theorem 4.8, since the only abnormal C multiplier is the zero vector, one also concludes that the value function is Lipschitz. However, the upper estimate for the limiting subdifferentials of the value function in terms of the C multiplier set is a strict inclusion here:

$$\begin{aligned} \partial V(0,0) &= \{(0,0)\} \cup \{(1,-1)\} \subset \{(\beta,\eta) : \beta = -\eta \in [0,1]\} = -M_C^1(0,0), \\ \partial^\infty V(0,0) &= \{(0,0)\} = -M_{CD}^0(0,0). \end{aligned}$$

The upper estimate for both the limiting and the singular limiting subdifferentials of the value function in Theorem 4.11 are both strict:

$$\begin{aligned}\partial V(0,0) &= \{(0,0)\} \cup \{(1,-1)\} \\ &\subset \{(\beta,\eta) : \beta = -\eta \in (-\infty, 0] \cup (1,\infty)\} = -M_P^1(0,0), \\ \partial^\infty V(0,0) &= \{(0,0)\} \\ &\subset \{(\beta,\eta) : \beta = -\eta\} = -M_P^0(0,0).\end{aligned}$$

These inclusions are not very helpful since the Lipschitz continuity of the value function cannot be detected and the upper estimate is unbounded.

Since there is no S multiplier for this problem, the limiting subdifferential of the value function cannot be estimated in terms of the S multiplier. In fact, the assumptions in Theorem 4.13 are not satisfied for this problem. Indeed,

$$(0,0) = \beta(1,-1) + \eta(1,0) + (0,\xi)$$

does not imply that $\eta = 0, \xi = 0$.

Note that by Theorem 4.2, if the growth hypotheses were satisfied, then

$$\begin{aligned}\partial V(0,0) &\subseteq \{\beta(-1,0) - r^F(0,-1) : (\beta, r^F, r^y, \mu) \in M_{NLP}^1(\Sigma)\}, \\ \partial^\infty V(0,0) &\subseteq \{\beta(-1,0) - r^F(0,-1) : (\beta, r^F, r^y, \mu) \in M_{NLP}^0(\Sigma)\}.\end{aligned}$$

But this is not possible since $M_{NLP}^1(\Sigma) = \emptyset$. Indeed, (GH) is not satisfied for this example.

In the above example, $M_{NLP}^0(\Sigma) \neq \{0\}$, while $M_{CD}^0(\Sigma) = \{0\}$. In fact, it is not just a coincidence that $M_{NLP}^0(\Sigma) \neq \{0\}$. In general, the Mangasarian–Fromovitz constraint qualification satisfying at a feasible solution $(\bar{x}, \bar{z}, \bar{u})$ implies that $M_{NLP}^0((\bar{x}, \bar{z}, \bar{u})) = \{0\}$, and in the case of no abstract constraint, the two conditions are equivalent (see, e.g., [6] and [25, Proposition 4.5] for details). It is well known that in the case of no abstract constraint, the Mangasarian–Fromovitz constraint qualification fails to hold at every feasible point of the OPCCs. (The proof for the case where the complementarity constraint comes from the KKT condition of a lower level quadratic programming problem was given in Chen and Florian [1, Lemma 3.1], and the proof for the general case was given in [25, Proposition 1.1].) We now prove that even for the case when the abstract constraint set C is present, there always exist nonzero abnormal NLP multipliers for the OPCC.

PROPOSITION 4.16. *Let $(\bar{x}, \bar{y}) \in R^{n+m}$ be any feasible solution of the* OPCC. *Then $M_{NLP}^0(\bar{x}, \bar{y}) \backslash \{0\} \neq \emptyset$.*

Proof. The point $(\bar{x}, \bar{y})$ is obviously a solution to the following optimization problem:

$$\begin{aligned}\text{minimize} \quad & \langle y, F(x,y)\rangle \\ \text{subject to} \quad & y \geq 0, F(x,y) \geq 0.\end{aligned}$$

By the multiplier rule, there exists $\mu \geq 0, r^y \in R_+^m, r^F \in R_+^m$ not all zero such that

$$\begin{aligned}&0 = \mu \nabla \langle y, F\rangle(\bar{x}, \bar{y}) - (0, r^y) - \nabla F(\bar{x}, \bar{y})^\top r^F, \\ &\langle \bar{y}, r^y\rangle = 0, \langle r^F, F(\bar{x}, \bar{y}) = 0.\end{aligned}$$

Therefore, taking $\gamma = 0, \beta = 0$ $(\gamma = 0, \beta = 0, r^F, r^y, \mu)$ is a nonzero NLP abnormal multiplier of the OPCC. □

4.6. Applications to the bilevel programming problem. One of the motivations to consider OPCCs is to solve the following bilevel programming problem:

$$\text{(BLPP)} \qquad \begin{array}{ll} \text{minimize} & f(x,z) \\ \text{subject to} & z \in S(x), \Psi(x,z) \le 0, (x,z) \in C, \end{array} \tag{30}$$

where $S(x)$ is the solution of the lower-level problem

$$P_x \qquad \begin{array}{ll} \text{minimize} & g(x,z) \\ \text{subject to} & \psi(x,z) \le 0, \end{array} \tag{31}$$

where $f : R^{n+a} \to R, g : R^{n+a} \to R, \psi : R^{n+a} \to R^b, \Psi : R^{n+a} \to R^d$. Under suitable convexity assumptions, we can replace the lower problem by its KKT conditions. As in [24], we find that any (x,z) is solution of (BLPP) if and only if there is u such that (x,z,u) is solution of the problem

$$\begin{array}{ll} \text{minimize} & f(x,z) \\ \text{subject to} & \psi(x,z) \le 0 \text{ and } u \ge 0, \\ & \langle \psi(x,z), u \rangle = 0, \\ & \nabla_z g(x,z) + \nabla_z \psi(x,z)^\top u = 0, \\ & \Psi(x,z) \le 0, (x,z) \in C, \end{array} \tag{32}$$

which is an OPCC.

Consider the perturbed bilevel programming problem

$$\text{BLPP}(\alpha) \quad \begin{array}{ll} \text{minimize} & f(x,z,\alpha) \\ \text{subject to} & z \in S(x,\alpha), \Psi(x,z,\alpha) \le 0, (x,z) \in C, \end{array} \tag{33}$$

where $S(x,\alpha)$ is the solution of the lower-level problem

$$\begin{array}{ll} \text{minimize} & g(x,z,\alpha) \\ \text{subject to} & \psi(x,z,\alpha) \le 0. \end{array} \tag{34}$$

Under suitable assumptions, BLPP(α) is equivalent to

$$\begin{array}{ll} \text{minimize} & f(x,z,\alpha) \\ \text{subject to} & \psi(x,z,\alpha) \le 0 \text{ and } u \ge 0, \\ & \langle \psi(x,z,\alpha), u \rangle = 0, \\ & \nabla_z g(x,z,\alpha) + \nabla_z \psi(x,z,\alpha)^T u = 0, \\ & \Psi(x,z,\alpha) \le 0, (x,z) \in C. \end{array} \tag{35}$$

Hence the results in this section allow us to derive the properties of the value function and compute the upper estimates of the limiting subdifferentials of V by the various kinds of multipliers for the above problem. For example, we can conclude that the value function is Lipschitz continuous when the strong second order sufficient condition and the linear independence of the binding constraints hold for the lower level problem. Indeed, in this case the corresponding generalized equation is strongly regular; hence the set of abnormal CD multipliers contains only the zero vector (see Ye [23, Theorem 5.1]).

Acknowledgments. The authors would like to thank the anonymous referees whose constructive suggestions led to the better presentation of the results and the consideration of the upper estimates in terms of various multipliers other than the CD multipliers in the last section.

REFERENCES

[1] Y. Chen and M. Florian, *The nonlinear bilevel programming problem: Formulations, regularity and optimality conditions*, Optimization, 32 (1995), pp. 193–209.

[2] F. H. Clarke, *Optimization and Nonsmooth Analysis*, 2nd ed., Classics in Appl. Math. 5, SIAM, Philadelphia, 1990.

[3] F. H. Clarke, *Methods of Dynamic and Nonsmooth Optimization*, CBNS-NSF Regional Conf. Ser. in Appl. Math. 57, SIAM, Philadelphia, 1989.

[4] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Nonsmooth Analysis and Control Theory*, Grad. Texts in Math. 178, Springer-Verlag, New York, 1998.

[5] A. Dontchev and R. T. Rockafellar, *Characterizations of strong regularity for variational inequalities over polyhedral convex sets*, SIAM J. Optim., 6 (1996), pp. 1087–1105.

[6] A. Jourani, *Constraint qualifications and Lagrange multipliers in nondifferentiable programming problems*, J. Optim. Theory Appl., 81 (1994), pp. 533–548.

[7] P. D. Loewen, *Optimal Control via Nonsmooth Analysis*, AMS, Providence, RI, 1993.

[8] Z. Q. Luo, J. S. Pang, and D. Ralph, *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press, New York, 1996.

[9] B. S. Mordukhovich, *Metric approximations and necessary optimality conditions for general classes of nonsmooth extremal problems*, Soviet Math. Dokl., 22 (1980), pp. 526–530.

[10] B. S. Mordukhovich, *Approximation Methods in Problems of Optimization and Control*, Nauka, Moscow, 1988.

[11] B. S. Mordukhovich, *Sensitivity analysis in nonsmooth optimization*, in Theoretical Aspects of Industrial Design, D. A. Field and V. Komkov, eds., SIAM, Philadelphia, 1992, pp.32–46.

[12] B. S. Mordukhovich, *Generalized differential calculus for nonsmooth and set-valued mappings*, J. Math. Anal. Appl., 183 (1994), pp. 250–288.

[13] B. S. Mordukhovich, *Nonsmooth sequential analysis in Asplund spaces*, Trans. Amer. Math. Soc., 348 (1996), pp. 1235–1280.

[14] J. V. Outrata, *Optimality conditions for a class of mathematical programs with equilibrium constraints*, Math. Oper. Res., 24 (1999), pp. 627–644.

[15] J. V. Outrata, M. Kočvara, and J. Zowe, *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints: Theory, Applications and Numerical Results*, Kluwer, Dordrecht, The Netherlands, 1998.

[16] R. A. Poliquin and R. T. Rockafellar, *Tilt stability of a local minimum*, SIAM J. Optim., 8 (1998), pp. 287–299.

[17] S. M. Robinson, *Strongly regular generalized equation*, Math. Oper. Res., 5 (1980), pp. 43–62.

[18] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.

[19] H. Scheel and S. Scholtes, *Mathematical programs with complementarity constraints: Stationarity, optimality, and sensitivity*, Math. Oper. Res., 25 (2000), pp. 1–22.

[20] S. Scholtes, *Convergence Properties of a Regularization Scheme for Mathematical Programs with Complementarity Constraints*, Working paper, Judge Institute of Management Studies, University of Cambridge, Cambridge, UK, 1999.

[21] S. Scholtes, *How Stringent is the Linear Independence Assumption for Mathematical Programs with Stationary Constraints?*, Working paper, Judge Institute of Management Studies, University of Cambridge, Cambridge, UK, 1999.

[22] J. J. Ye, *Optimality conditions for optimization problems with complementarity constraints*, SIAM J. Optim., 9 (1999), pp. 374–387.

[23] J. J. Ye, *Constraint qualifications and necessary optimality conditions for optimization problems with variational inequality constraints*, SIAM J. Optim., 10 (2000), pp. 943–962.

[24] J. J. Ye and X. Y. Ye, *Necessary optimality conditions for optimization problems with variational inequality constraints*, Math. Oper. Res., 22 (1997), pp. 977–997.

[25] J. J. Ye, D. L. Zhu, and Q. J. Zhu, *Exact penalization and necessary optimality conditions for generalized bilevel programming problems*, SIAM J. Optim., 7 (1997), pp. 481–507.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 3, pp. 724–740

A STATE-SPACE CALCULUS FOR RATIONAL PROBABILITY DENSITY FUNCTIONS AND APPLICATIONS TO NON-GAUSSIAN FILTERING*

BERNARD HANZON† AND RAIMUND J. OBER‡

Abstract. We propose what we believe to be a novel approach to performing calculations for rational density functions using state-space representations of the densities. By standard results from realization theory, a rational probability density function is considered to be the transfer function of a linear system with generally complex entries. The stable part of this system is positive-real, which we call the density summand. The existence of moments is investigated using the Markov parameters of the density summand. Moreover, explicit formulae are given for the existing moments in terms of these Markov parameters. Some of the main contributions of the paper are explicit state-space descriptions for products and convolutions of rational densities.

As an application which is of interest in its own right, the filtering problem is investigated for a linear time-varying system whose noise inputs have rational probability density functions. In particular, state-space formulations are derived for the calculation of the prediction and update equations. The case of Cauchy noise is treated as an illustrative example.

Key words. probability theory, realization theory for linear systems, non-Gaussian filtering, rational functions, linear algebra

AMS subject classifications. 93, 60, 15, 62, 90

PII. S036301299731610X

1. Introduction.

We are going to consider the filtering problem for the first order system

$$x_{t+1} = f_t x_t + \eta_t,$$
$$y_t = h_t x_t + \epsilon_t,$$

$t = 0, 1, 2, \ldots$, where f_t, h_t are assumed to be known real numbers and, for ease of exposition, are assumed to be such that $f_t \neq 0$ and $h_t > 0$, $t \geq 0$. The noise sequences $\{\eta_t\}_{t\geq 0}$ and $\{\epsilon_t\}_{t\geq 0}$ are assumed to be mutually independent sequences of independent random variables whose probability density functions are rational. The initial state x_0 is also assumed to be a random variable which is independent of the noise sequences and also has a rational density. No assumption is made that any of the random variables are identically distributed.

This filtering problem with non-Gaussian noise has applications in econometrics, for example in the analysis of financial time series. Studies have shown that the quantities that are encountered there often do not admit a Gaussian distribution ([7], [5], and see also [12]), since these distributions have "heavy tails." As one of the consequences, higher order moments may not exist. It has therefore been proposed (see, e.g., [11]) that these distributions be modelled by rational densities, both because

*Received by the editors February 7, 1997; accepted for publication (in revised form) November 27, 2000; published electronically September 7, 2001. This research was supported by a NATO collaborative research grant CRG 940733. The work of the second author was also supported in part by NSF grants DMS 9501223 and DMS 9803186.

http://www.siam.org/journals/sicon/40-3/31610.html

†Department of Econometrics, Vrije Universiteit, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands (bhanzon@econ.vu.nl).

‡Center for Engineering Mathematics EC35, University of Texas at Dallas, Richardson, TX 75083 (ober@utdallas.edu).

they do have "heavy tails" and because of the richness of the class of distributions. Examples of rational probability densities which have been used in the literature are Cauchy densities and Student densities with odd number of degrees of freedom.

The state filtering problem is defined as the problem of finding the best estimate $\hat{x}_t$ of x_t for the quadratic loss function given knowledge of the distribution of x_0 and the values of $y_0, y_1, \ldots, y_t$. Since

$$\hat{x}_t = \int_{-\infty}^{\infty} xp(x)_{x_t|y_t,y_{t-1},\ldots,y_0} dx,$$

this estimate can be found if the conditional density $p_{x_t|y_t,y_{t-1},\ldots,y_0}$ of x_t is known and the first moment exists, given the measured values of $y_t, y_{t-1}, \ldots, y_0$ and knowledge of the distribution of x_0.

In principle, the calculation of the conditional densities is not difficult. The unnormalized conditional densities, denoted by ρ instead of p, are given by the following.

Update step. For $t = 0$,

$$\rho_{x_0|\mathcal{Y}_0}(x) = \rho_{x_0|y_0}(x) = \rho_{y_0|x}(y_0)\rho_{x_0}(x) = \rho_{\epsilon_0}(y_0 - h_0 x)\rho_{x_0}(x);$$

for $t \geq 1$,

$$\rho_{x_t|\mathcal{Y}_t}(x) = \rho_{y_t|x}(y_t)\rho_{x_t|\mathcal{Y}_{t-1}}(x) = \rho_{\epsilon_t}(y_t - h_t x)\rho_{x_t|\mathcal{Y}_{t-1}}(x),$$

$x \in \Re$.

Prediction step. For $t \geq 0$,

$$\rho_{x_{t+1}|\mathcal{Y}_t}(x) = (\rho_{f_t x_t|\mathcal{Y}_t} * \rho_{\eta_t})(x) = \int_{-\infty}^{\infty} \rho_{x_t|\mathcal{Y}_t}\left(\frac{\xi}{f_t}\right) \rho_{\eta_t}(x - \xi)\, d\xi, \quad x \in \Re.$$

Here we have set $\mathcal{Y}_t$ to be the collection of observations $y_t, y_{t-1}, \ldots, y_0$.

In [11] it was noted that the various probability densities occurring in the filtering problem are all rational functions if the noise variables and the initial state have rational probability densities and if explicit formulas are given. The practical problem in doing these calculations for large numbers of observations is that the conditional densities are fairly complicated to calculate. To alleviate this problem we propose the use of state-space techniques for these calculations. Since by assumption the initial state and the noise sequences have rational densities, this is indeed possible. For this purpose we are going to develop a "state-space calculus" for rational probability density functions. We believe that the use of linear system theory to analyze rational probability densities is novel and may be of relevance beyond the application to non-Gaussian filtering as discussed here. Since the approach is valid in general, we develop the state-space approach for general probability density functions as well as for conditional probability density functions.

Let ρ be a not necessarily normalized rational probability density, i.e., $\rho(x)$ is a rational function in the independent variable x, such that $\rho(x) \geq 0$, $x \in \Re$, and $0 < \int_{-\infty}^{\infty} \rho(x)dx < \infty$. This implies that ρ is *strictly proper*, i.e., $\lim_{|x|\to\infty} \rho(x) = 0$. To speak of a not necessarily normalized or unnormalized probability density function is an abuse of the standard notion of a probability density function, since this term implies that its integral is 1. For ease of notation we use the notion of a not necessarily normalized or unnormalized density function to imply that all properties of a density function are given, with the possible exception of the normalization of its integral.

By standard realization theory there exists a minimal state-space realization such that

$$\rho(x) = c(ixI - A)^{-1}b, \qquad x \in \Re.$$

In particular, we will present here state-space formulae for the translation, scaling, product, and convolution of rational probability density functions. Most of our results will be formulated in terms of state-space realizations for the density summand, which is defined to be the "stable" part of the probability density function. One reason for doing this is that in this way the dimensions of the realizations are typically half of what they would be otherwise. For actual implementations of our results, this could lead to significant computational advantages, in particular when repeated applications are necessary such as can be expected for the filtering case. Moreover, we will investigate the existence of moments from the state-space point of view and give state-space formulae for the existing moments in terms of the Markov parameters of the density summand. A major part of the investigation will be built on a careful analysis of the connections between impulse responses, transfer functions, and characteristic functions of the various objects. In a result that may be of independent interest, a state-space formula is given for the system whose impulse response is the product of impulse responses of two systems.

2. Notation and preliminaries. The symbol $\mathcal{C}$ stands for the complex field, and the symbol $\Re$ stands for the real field. If (A, b, c) is a linear state-space system, we also often use the notation $(\frac{A\,|\,b}{c\,|\,0})$. If M is a complex matrix, M^* denotes the adjoint matrix. If G is a rational function, G^* is defined by $G^*(s) = \overline{(G(-\bar{s}))}$, $s \in \mathcal{C}$. If G has the realization (A, b, c) (i.e., $G(s) = c(sI - A)^{-1}b$ for $s \in \mathcal{C}\backslash\sigma(A)$, where $\sigma(A)$ is the spectrum of A), then G^* has the realization $(-A^*, c^*, -b^*)$. We call a system (A, b, c) *stable* if all eigenvalues of A are in the open left half plane. Note that such systems are often also called asymptotically stable. A rational function G is called strictly proper if $\lim_{|s|\to\infty} G(s) = 0$. An unnormalized probability density function ρ is a nonnegative integrable function on $\Re$ such that $\int_{-\infty}^{\infty} \rho(x)dx > 0$, but not necessarily 1. Then $p = \rho/\int_{-\infty}^{\infty} \rho(x)dx$ is a normalized density function. The set of functions $\mathcal{P}$ is defined in section 3.

3. State-space representations of rational densities. If ρ is not a necessarily normalized rational probability density function, then ρ is strictly proper, i.e., $\lim_{|x|\to\infty} \rho(x) = 0$. Therefore, by standard realization theory (see, e.g., [4, Section 2.1], [10, Sections 10–11]), there exists a minimal linear state-space system (A, b, c) such that

$$\rho(x) = c(ixI - A)^{-1}b, \qquad x \in \Re.$$

It should be noted that the system matrices A, b, c will be, in general, complex matrices. A rational probability density function which is symmetric with respect to 0, however, could be realized with real system matrices.

Note also that we have set up the realization in such a way that we consider the rational function to be defined on the imaginary axis. While in principle the choice of axis is arbitrary, it is convenient to choose the imaginary axis since then standard realization theoretic methods can be adopted without having to change the axis. In particular, we will be using the formal analogy of methods developed for spectral densities which are most naturally considered to be defined on the imaginary axis. To

make this convention clear, set

$$\Phi(ix) := \rho(x), \qquad x \in \Re.$$

Since Φ is a rational function defined on the imaginary axis, it can be extended as a rational function to the whole complex plane. This rational function has the following properties.

1. $\Phi(s) = \Phi^*(s), s \in \mathcal{C}$.
2. Φ has no poles on the imaginary axis.
3. $\Phi(ix) \geq 0$, $x \in \Re$.
4. $\lim_{|s|\to\infty} \Phi(x) = 0$.

The set of rational functions that satisfies properties 1, 2, 3, and 4 is denoted by $\mathcal{P}$. Many of our calculations are going to be based on the following well-known additive decomposition (see Lemma 3.1) of Φ:

$$\Phi(s) = Z(s) + Z^*(s), \qquad s \in \mathcal{C},$$

where Z is a stable rational function, i.e., all poles of Z are in the open left half plane. This decomposition is unique if we assume that $Z(\infty) = 0$, which can be done since $\Phi(\infty) = 0$. The function Z is called the *spectral summand* or Φ. We will also call Z the *density summand* of ρ.

In the following lemma some elementary and standard state-space properties are collected concerning this additive decomposition of Φ. For the sake of completeness, a short proof is added for this standard result.

LEMMA 3.1. *Let $\Phi \in \mathcal{P}$. Then there exists a stable rational function Z such that*

$$\Phi = Z + Z^*.$$

Let (A, b, c) be a minimal realization of Φ, i.e., $\Phi(s) = c(sI - A)^{-1}b$, and (A, b, c) is minimal. There exists an equivalent realization

$$\left(\begin{array}{cc|c} A_1 & 0 & b_1 \\ 0 & A_2 & b_2 \\ \hline c_1 & c_2 & 0 \end{array}\right)$$

of (A, b, c) such that all eigenvalues of A_1 are in the open left half plane and all eigenvalues of A_2 are in the open right half plane. The state-space system (A_1, b_1, c_1) is a minimal realization of Z, and (A_2, b_2, c_2) is a minimal realization of Z^*.

Moreover, (A_2, b_2, c_2) is equivalent to $(-A_1^*, c_1^*, -b_1^*)$. In particular, there exists a minimal realization of Φ such that

$$\left(\begin{array}{cc|c} A_1 & 0 & b_1 \\ 0 & -A_1^* & c_1^* \\ \hline c_1 & -b_1^* & 0 \end{array}\right).$$

Proof. Let $\Phi = Z_s + Z_u$ be a stable-unstable partial fraction decomposition of Φ, i.e., the partial fraction decomposition of Φ such that Z_s is stable, meaning that all its poles are in the open left half plane, and Z_u is unstable, meaning that all its poles are in the open right half plane. Note that this decomposition is unique. Let

(A_1, b_1, c_1) be a minimal realization of Z_s, and let (A_2, b_2, c_2) be a minimal realization of Z_u. Then

$$\left(\begin{array}{cc|c} A_1 & 0 & b_1 \\ 0 & A_2 & b_2 \\ \hline c_1 & c_2 & 0 \end{array}\right)$$

is a minimal realization of Φ and hence equivalent to (A, b, c). Set $Z := Z_s$. We need to show that $Z_u = Z^*$. Now consider $Z_s + Z_u = \Phi = \Phi^* = (Z_s + Z_u)^* = Z_s^* + Z_s^*$. Note that Z_s^* has all its roots in the open right half plane, and Z_u^* has all its roots in the open left half plane. By the above-mentioned uniqueness of the stable-unstable partial fraction decomposition, we have that $Z_u = Z_s^* = Z^*$. The remaining parts of the lemma follow immediately. □

Example. As a special case we are going to consider the Cauchy density, which was suggested, for example in [7], as a suitable density to study financial time series. The normalized Cauchy density is defined as

$$p(x) = \frac{1}{\pi} \frac{k}{(x - x_0)^2 + k^2}, \quad x \in \Re,$$

where $x_0 \in \Re$ and $k > 0$. A state-space realization of $\Phi(ix) := p(x)$, $x \in \Re$, is given by

$$\left[\begin{array}{c|c} A_\Phi & b_\Phi \\ \hline c_\Phi & 0 \end{array}\right] := \left[\begin{array}{cc|c} -k + ix_0 & 0 & \frac{1}{2\pi} \\ 0 & k + ix_0 & 1 \\ \hline 1 & -\frac{1}{2\pi} & 0 \end{array}\right].$$

The density summand of p is

$$Z(s) = \frac{1}{2\pi} \frac{1}{s - (-k + ix_0)},$$

which has one pole at $-k + ix_0$. A state-space realization of Z is given by

$$\left[\begin{array}{c|c} A & b \\ \hline c & 0 \end{array}\right] := \left[\begin{array}{c|c} -k + ix_0 & \frac{1}{2\pi} \\ \hline 1 & 0 \end{array}\right].$$

4. Fourier transforms, moments, and Markov parameters. In order to obtain state space formulae for the moments of probability density functions and for the convolution of such densities, we need to employ the Fourier transform. The main tool will be to interpret the density summand as the Fourier transform of the impulse response of a stable linear state-space system. Actually, we introduce the Fourier transform as the Laplace transform evaluated on the imaginary axis. For a general reference on Fourier transforms see, e.g., [9], [6]. This way of proceeding is of course closely related to the use of the characteristic function in statistics, but there are a few more minor technical differences.

For an integrable function f on $\Re$ define the Fourier transform as usual by

$$(\mathcal{F}(f))(iw) = \int_{-\infty}^{\infty} f(t) e^{-iwt} dt, \qquad iw \in i\Re.$$

If (A, b, c) is a stable system, let $m^+(t) := ce^{tA}b$ for $t \geq 0$, and $m^+(t) := 0$ for $t < 0$. Then the Fourier transform of m^+ is given by

$$(\mathcal{F}m^+)(iw) = \int_0^\infty ce^{tA}be^{-itw}dt = c(-iwI + A)^{-1}e^{(-iwI+A)t}|_0^\infty b = c(iwI - A)^{-1}b$$
$$=: G(iw), \qquad iw \in i\Re.$$

If we set $m^-(t) := b^*e^{-tA^*}c^*$ for $t < 0$ and $m^-(t) := 0$ for $t \geq 0$, then the Fourier transform of m^- is given by

$$(\mathcal{F}m^-)(iw) = \int_{-\infty}^0 b^*e^{-tA^*}c^*e^{-itw}dt = b^*(-iwI - A^*)^{-1}e^{(-iwI-A^*)t}|_{-\infty}^0 c^*$$
$$= -b^*(iwI - (-A)^*)^{-1}c^* = G^*(iw), \quad iw \in i\Re.$$

The lth derivative of m^+ at $t > 0$ is given by $(m^+)^{(l)}(t) = cA^le^{tA}b$. Hence the right-hand side limit of the lth derivative at 0 is given by $(m^+)^{(l)}(0+) = cA^lb$. The lth derivative of m^- at $t < 0$ is given by $(m^-)^{(l)}(t) = b^*(-A^*)^le^{-tA^*}c^*$. Hence the left-hand side limit of the lth derivative at 0 is given by $(m^-)^{(l)}(0-) := b^*(-A^*)^lc^* = (-1)^l(cA^lb)^* = (-1)^l((m^+)^{(l)}(0+))^*$, $l \geq 0$.

Assume now that (A, b, c) is a realization of the spectral summand Z of the function $\Phi \in \mathcal{P}$. Then $(\mathcal{F}m^+)(iw) = Z(iw)$, $(\mathcal{F}m^-)(iw) = Z^*(iw)$, and for $m := m^+ + m^-$ we have that $(\mathcal{F}m)(iw) = \Phi(iw)$, $iw \in i\Re$. Hence m is the inverse Fourier transform of Φ. Note that m is l times continuously differentiable at $t = 0$, $l \geq 0$, if and only if $cA^kb = (-1)^k(cA^kb)^*$, $k = 0, 1, \ldots, l$.

If G is a strictly proper rational function on $\mathcal{C}$, then G admits a Laurent expansion around ∞ such that

$$G(s) = \sum_{n=1}^\infty M(n)\frac{1}{s^n}$$

for $s \in \mathcal{C}$ with $|s|$ large enough. The parameters $M(n)$, $n = 1, 2, \ldots$, are the *Markov parameters* of G (see, e.g., [10, p. 194]). If (A, b, c) is a realization of G, then

$$G(s) = c(sI - A)^{-1}b = \frac{1}{s}c\left(I - \frac{A}{s}\right)^{-1}b = \frac{1}{s}c\sum_{k=0}^\infty\left(\frac{1}{s}A\right)^k b = \sum_{n=1}^\infty \frac{1}{s^n}cA^{n-1}b.$$

Hence the Markov parameters of G are given by

$$M(n) = cA^{n-1}b, \quad n = 1, 2, 3, \ldots.$$

The Markov parameters of a rational strictly proper function of $\mathcal{P}$ and its spectral summand are easily determined.

LEMMA 4.1. *Let Φ be a strictly proper rational function in $\mathcal{P}$ with spectral summand Z. If (A, b, c) is a realization of Z, then*

1. *the Markov parameters of Z are given by*

$$cA^{n-1}b, \qquad n = 1, 2, 3, \ldots,$$

2. *the Markov parameters of Z^* are given by*

$$(-1)^nb^*(A^*)^{n-1}c^* = (-1)^n(cA^{n-1}b)^*, \qquad n = 1, 2, 3, \ldots, \text{ and}$$

3. *the Markov parameters of* Φ *are given by*

$$cA^{n-1}b - (-1)^{n-1}(cA^{n-1}b)^*, \qquad n = 1, 2, 3, \ldots.$$

In the following lemma, a basic result on the integrability of rational functions is summarized.

LEMMA 4.2. *Let* $G = \frac{n}{d}$ *with* n *and* d *as a pair of coprime polynomials. Then*

$$\int_{-\infty}^{\infty} |G(x)|dx < \infty$$

if and only if degree$(n) \leq$ *degree*$(d) - 2$ *and* $d(x) \neq 0$ *for all* $x \in \Re$.

If G is as defined in the lemma, then *degree*(d) − *degree*(n) is called the *codegree* of the rational function G. Therefore, G is integrable if and only if the codegree of G is greater than or equal to 2. This lemma also implies that if the random variable X has the rational probability density function $p = \frac{n}{d}$, then the moments EX^k exist for $k = 0, 1, 2, \ldots$, codegree$(p) - 2$.

Let k be such that $M(n) = 0$ for $n = 1, 2, \ldots, k-1$ and $M(k) \neq 0$. Then the codegree of G is k [10, p. 254].

Summarizing the previous remarks, we obtain the following proposition.

PROPOSITION 4.1. *Let* Φ *be a strictly proper rational function in* $\mathcal{P}$ *with spectral summand* Z. *Let* (A, b, c) *be a minimal realization of* Z. *Let* $m(t) := ce^{tA}b$ *for* $t \geq 0$ *and* $m(t) := b^*e^{-tA^*}c^*$ *for* $t < 0$. *Then the following hold.*

1. *The codegree of* Φ *is* k *if and only if* $M(n) = 0$ *for all* $n \in \{1, \ldots, k-1\}$ *and* $M(k) \neq 0$, *where* $M(n)$ *is the* nth *Markov parameter of* Φ.

2. *The codegree of* Φ *is* k *if and only if*

$$cA^{n-1}b = (-1)^{n-1}(cA^{n-1}b)^*$$

for all $n \in \{1, \ldots, k-1\}$ *and*

$$cA^{k-1}b \neq (-1)^{k-1}(cA^{k-1}b)^*.$$

3. m *is* $k-1$ *times continuously differentiable at* 0 *if and only if the first* k *Markov parameters of* Φ *are zero.*

4. Φ *has codegree* k *if and only if* m *is* $k-2$ *times continuously differentiable but not* $k-1$ *times continuously differentiable at* 0.

The following theorem provides important results concerning moments of a random variable with rational probability density.

THEOREM 4.1. *Let* X *be a random variable with unnormalized rational probability density function* ρ. *Let* (A, b, c) *be a realization of the density summand* Z *of* ρ. *Then the following hold.*

1. *The codegree of* ρ *is* k *if and only if*

$$cA^{n-1}b = (-1)^{n-1}(cA^{n-1}b)^*$$

for all $n \in \{1, \ldots, k-1\}$ *and* $cA^{k-1}b \neq (-1)^{k-1}(cA^{k-1}b)^*$.

2. *The* lth *moment* EX^l *of* X *with* l *a nonnegative integer exists if and only if* $l \in \{0, 1, \ldots, k-2\}$.

3. $EX^l = (-i)^l \frac{cA^lb}{cb}$ *for all* $l \in \{0, 1, \ldots, k-2\}$.

Proof. (1) The proof follows immediately from Proposition 4.1.

(2) Recall that the lth moment of X is given by

$$EX^l = \frac{1}{R}\int_{-\infty}^{\infty} x^l \rho(x)\,dx,$$

where $R := \int_{-\infty}^{\infty} \rho(x)\,dx$. The codegree of the integrand is $k-l$. By Lemma 4.2 the integrand is integrable if and only if its codegree is greater than or equal to 2. Hence the claim.

(3) Let $0 \le l \le k-2$. Set $\Phi(ix) := \rho(x)$, $x \in \Re$, and use the notation of Proposition 4.1. Then m is $k-2$ times continuously differentiable at 0 and therefore on $\Re$. Since the codegree of ρ is greater than or equal to 2, m is continuous on $\Re$. Since ρ and m are continuous and integrable, we have by the inversion theorem for Fourier transforms (see, e.g., [6, Theorem 60.1, p. 296]) that

$$m(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \Phi(iw)e^{iwt}dw, \qquad t \in \Re.$$

Note that differentiation up to order $k-2$ under this integral is justified by the usual argument (see, e.g. [6, Theorem 53.5, p. 268]) as $|\omega^l \Phi(i\omega)e^{i\omega t}| = |\omega^l \Phi(i\omega)|$ is integrable for each $t \in \Re$ and $0 \le l \le k-2$. Hence for $t \in \Re$,

$$\frac{d^l}{dt^l}m(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \Phi(iw)\frac{d^l}{dt^l}e^{iwt}dw = (i)^l \frac{1}{2\pi}\int_{-\infty}^{\infty} w^l \Phi(iw)e^{iwt}\,dw.$$

Evaluating at $t=0$, we have

$$\frac{d^l}{dt^l}m(t)|_{t=0} = \frac{1}{2\pi}(i)^l \int_{-\infty}^{\infty} w^l \Phi(iw)e^{iwt}dw|_{t=0} = R(i)^l \frac{1}{2\pi}EX^l.$$

Since $\frac{d^l}{dt^l}m(t)|_{t=0} = cA^l b$, $l = 0, \ldots, k-2$, we have that

$$EX^l = \frac{2\pi}{R}(-i)^l cA^l b.$$

The constant R is determined by considering this equation for $l=0$. Since $EX^0 = 1$, we have that $R = 2\pi cb$. Hence $EX^l = (-i)^l \frac{cA^l b}{cb}$. □

In most of this paper we will be dealing with unnormalized rational probability densities ρ. If (A,b,c) is a state-space realization of the density summand of ρ, the normalized probability density function is given by $p := \rho / \int_{-\infty}^{\infty} \rho(x)dx$. By the above proposition $\int_{-\infty}^{\infty} \rho(x)dx = 2\pi cb$, which provides a state-space formula for the normalization constant.

If X is a random variable with rational probability density function ρ whose density summand has the state-space realization (A,b,c), then the first moment exists if the codegree of ρ is at least 3. This is the case if and only if

$$cb = (cb)^*$$

and

$$cAb = -(cAb)^*.$$

If the first moment, i.e., the mean, exists, then by the theorem it is given by

$$EX = -i\frac{cAb}{cb}.$$

In the above discussion we gave a state-space construction for the inverse Fourier transform m of a not necessarily normalized rational probability density function ρ, i.e.,

$$m(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \rho(\omega)e^{i\omega t}d\omega, \qquad \omega \in \Re.$$

In the statistical literature an important object is the characteristic function of a random variable X which is defined by $E(e^{itX})$, $t \in \Re$. If X has the unnormalized probability density function ρ, then

$$E(e^{itX}) = \frac{1}{\int_{-\infty}^{\infty}\rho(x)dx}\int_{-\infty}^{\infty} e^{itx}\rho(x)dx = \frac{2\pi}{\int_{-\infty}^{\infty}\rho(x)dx}m(t), \qquad t \in \Re.$$

Hence up to a (known) scaling factor the function m is identical to the characteristic function.

Example continued. We continue the discussion of the Cauchy density from section 3. Note that for all $x_0 \in \Re$ and $k > 0$

$$cAb = \frac{1}{2\pi}(-k + ix_0) \neq -\frac{1}{2}(-k - ix_0) = -(cAb)^*.$$

Hence by the theorem the mean EX does not exist. This is of course also directly evident by consideration of the integral $\int_{-\infty}^{\infty} xp(x)dx$.

If $m^+(\tau) := \frac{1}{2\pi}e^{\tau(-k+ix_0)}$ for $t \geq 0$ and $m^+(\tau) := 0$ for $t < 0$, then $\mathcal{F}(m^+)(iw) = \frac{1}{2\pi}\frac{1}{iw-(-k+ix_0)}$, $iw \in i\Re$. If $m^-(\tau) := \frac{1}{2\pi}e^{-\tau(-k-ix_0)}$ for $t < 0$ and $m^-(\tau) := 0$ for $t \geq 0$, then $\mathcal{F}(m^-)(iw) = \frac{1}{2\pi}\frac{1}{iw+(k-ix_0)}$, $iw \in i\Re$. With $m := m^+ + m^-$, we have that m is continuous at 0. The derivative is given by

$$\frac{d}{dt}m(t) = \frac{1}{2\pi}(-k + ix_0)e^{\tau(-k+ix_0)}, \qquad \tau > 0,$$
$$\frac{d}{dt}m(t) = \frac{1}{2\pi}(k + ix_0)e^{-\tau(-k-ix_0)}, \qquad \tau < 0.$$

Note that the left-hand side limit and the right-hand side limit do not agree at 0. Hence m is not differentiable at 0. As the codegree of p is 2, this is in agreement with Proposition 4.1. The first two Markov parameters of Φ are

$$c_\Phi b_\Phi = 0, \qquad c_\Phi A_\Phi b_\Phi = \frac{-k}{\pi}.$$

Hence the second Markov parameter is nonzero, which is also in agreement with Proposition 4.1. □

5. Operations on probability densities. In this section we are going to discuss state-space formulations of operations on rational probability densities. Given state-space realizations for the density summands of two probability densities, we will give state-space realizations for the density summand of the translation, scaling, product, and convolution of the densities.

5.1. Translation and scaling of a probability density. In the next straightforward lemma the effect of translation and scaling of a random variable on the state-space realization of the density is considered.

LEMMA 5.1. *Let X be a random variable with unnormalized rational density ρ. Let (A, b, c) be a minimal realization such that $\rho(x) = c(ixI - A)^{-1}b$, $x \in \Re$.*

Let $x_0 \in \Re$. Then the random variable $X + x_0$ has an unnormalized probability density function $q(x) = \rho(x - x_0)$, which has a realization $(A + ix_0 I, b, c)$, so

$$q(x) = c(ixI - (A + ix_0 I))^{-1}b, \qquad x \in \Re.$$

Let $a \in \Re$, $a \neq 0$; then the random variable aX has the unnormalized probability density function $q(x) = \frac{1}{a}\rho(\frac{x}{|a|})$ which has a realization $(aA, b, \frac{a}{|a|}c)$, so

$$q(x) = \frac{a}{|a|}c(ixI - aA)^{-1}b, \qquad x \in \Re.$$

In the following lemma, we are going to write down the analogous results for the case when a state-space realization is given for the density summand of the probability density. The proof is elementary.

LEMMA 5.2. *Let X be a random variable with unnormalized rational density ρ. Let (A, b, c) be a realization of the density summand Z of ρ.*

Let $x_0 \in \Re$; then the random variable $X + x_0$ has the unnormalized probability density function $q(x) = \rho(x - x_0)$, $x \in \Re$, whose density summand has a realization

$$\left(\begin{array}{c|c} A + ix_0 I & b \\ \hline c & 0 \end{array}\right).$$

Let $a \in \Re$, $a \neq 0$; then the random variable aX has the unnormalized probability density function $q(x) = \frac{1}{|a|}\rho(\frac{x}{a})$, whose density summand has a realization

$$\left(\begin{array}{c|c} aA & b \\ \hline c & 0 \end{array}\right)$$

if $a > 0$ and

$$\left(\begin{array}{c|c} -aA^* & c^* \\ \hline b^* & 0 \end{array}\right)$$

if $a < 0$.

5.2. Product of two rational probability densities. In the update step of the filtering problem, it is necessary to calculate the product of two density functions. We are going to do this also by state-space techniques using the decomposition into density summands. The following lemmas will be useful.

LEMMA 5.3. *Let G_1 and G_2 be two stable strictly proper rational functions with minimal state-space realizations (A_1, b_1, c_1) and (A_2, b_2, c_2). Then the product $G_1^* G_2$ can be decomposed as*

$$G_1^* G_2 = F + H^*,$$

where F, H are stable strictly proper rational functions such that F has the realizations given by

$$\left(\begin{array}{c|c} A_2 & b_2 \\ \hline b_1^* T_1 & 0 \end{array}\right), \qquad \left(\begin{array}{c|c} A_2 & T_2 c_1^* \\ \hline c_2 & 0 \end{array}\right),$$

and H^* has the realizations given by

$$\left(\begin{array}{c|c} -A_1^* & T_1 b_2 \\ \hline -b_1^* & 0 \end{array}\right), \qquad \left(\begin{array}{c|c} -A_1^* & -c_1^* \\ \hline c_2 T_2 & 0 \end{array}\right),$$

where T_1 is the unique solution to the Sylvester equation

$$A_1^* T_1 + T_1 A_2 + c_1^* c_2 = 0$$

and T_2 is the unique solution to the Sylvester equation

$$A_2 T_2 + T_2 A_1^* + b_2 b_1^* = 0.$$

Proof. Note that a realization of G_1^* is given by

$$(-A_1^*, c_1^*, -b_1^*),$$

and a realization of $G_1^* G_2$ is given by

$$\left(\begin{array}{cc|c} -A_1^* & c_1^* c_2 & 0 \\ 0 & A_2 & b_2 \\ \hline -b_1^* & 0 & 0 \end{array}\right).$$

Performing a state-space basis transformation with transformation matrix $\left(\begin{smallmatrix} I & T_1 \\ 0 & I \end{smallmatrix}\right)$, we obtain the equivalent realization

$$\left(\begin{array}{cc|c} -A_1^* & A_1^* T_1 + T_1 A_2 + c_1^* c_2 & T_1 b_2 \\ 0 & A_2 & b_2 \\ \hline -b_1^* & b_1^* T_1 & 0 \end{array}\right) = \left(\begin{array}{cc|c} -A_1^* & 0 & T_1 b_2 \\ 0 & A_2 & b_2 \\ \hline -b_1^* & b_1^* T_1 & 0 \end{array}\right)$$

since T_1 is such that $A_1^* T_1 + T_1 A_2 + c_1^* c_2 = 0$. Note that such a T_1 exists and is unique since both A_1^* and A_2 have all their eigenvalues in the open left half plane (see, e.g., [1, Vol. I, p. 225]). This representation implies the first set of realizations. The other set of realizations follows analogously by considering the state-space formula which corresponds to $G_2 G_1^*$. ∎

Remark. A method to generate explicit formulas for the solutions of Sylvester equations is presented in [3].

We can now derive the desired representation for the density summand of the product of two rational probability density functions.

PROPOSITION 5.1. *Let ρ_1 and ρ_2 be two unnormalized rational probability density functions with density summands Z_1 and Z_2. Let (A_i, b_i, c_i) be a minimal realization of Z_i, $i = 1, 2$. Then the density summand Z of the unnormalized rational probability density function $\rho = \rho_1 \rho_2$ has a realization given by*

$$\left(\begin{array}{cc|c} A_1 & b_1 c_2 & T_2^* c_2^* \\ 0 & A_2 & b_2 \\ \hline c_1 & b_1^* T_1 & 0 \end{array}\right),$$

where T_1, T_2 are the unique solutions to the Sylvester equations

$$A_1^* T_1 + T_1 A_2 + c_1^* c_2 = 0,$$
$$A_2 T_2 + T_2 A_1^* + b_2 b_1^* = 0.$$

Proof. We have that

$$\rho = \rho_1\rho_2 = (Z_1 + Z_1^*)(Z_2 + Z_2^*) = Z_1Z_2 + Z_1Z_2^* + (Z_1Z_2^*)^* + (Z_1Z_2)^*.$$

By Lemma 5.3 a state-space realization for the stable part of this expression is given by

$$\left(\begin{array}{cccc|c} A_1 & b_1c_2 & 0 & 0 & 0 \\ 0 & A_2 & 0 & 0 & b_2 \\ 0 & 0 & A_1 & 0 & T_2^*c_2^* \\ 0 & 0 & 0 & A_2 & b_2 \\ \hline c_1 & 0 & c_1 & b_1^*T_1 & 0 \end{array}\right),$$

where T_1 is the unique solution of the equation

$$A_1^*T_1 + T_1A_2 + c_1^*c_2 = 0$$

and T_2 is the unique solution of the equation

$$A_2T_2 + T_2A_1^* + b_2b_1^* = 0.$$

Performing a state-space basis transformation with transformation matrix

$$T = \left(\begin{array}{cccc} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & -I & 0 & I \end{array}\right),$$

we obtain the equivalent realization

$$\left(\begin{array}{cccc|c} A_1 & b_1c_2 & 0 & 0 & 0 \\ 0 & A_2 & 0 & 0 & b_2 \\ 0 & 0 & A_1 & 0 & T_2^*c_2^* \\ 0 & 0 & 0 & A_2 & 0 \\ \hline c_1 & b_1^*T_1 & c_1 & b_1^*T_1 & 0 \end{array}\right),$$

which is equivalent to

$$\left(\begin{array}{ccc|c} A_1 & b_1c_2 & 0 & 0 \\ 0 & A_2 & 0 & b_2 \\ 0 & 0 & A_1 & T_2^*c_2^* \\ \hline c_1 & b_1^*T_1 & c_1 & 0 \end{array}\right).$$

On this realization perform another state-space basis transformation with transformation matrix

$$T = \left(\begin{array}{ccc} I & 0 & I \\ 0 & I & 0 \\ 0 & 0 & I \end{array}\right)$$

to obtain

$$\left(\begin{array}{ccc|c} A_1 & b_1c_2 & 0 & T_2^*c_2^* \\ 0 & A_2 & 0 & b_2 \\ 0 & 0 & A_1 & T_2^*c_2^* \\ \hline c_1 & b_1^*T_1 & 0 & 0 \end{array}\right),$$

which is equivalent to

$$\left(\begin{array}{cc|c} A_1 & b_1c_2 & T_2^*c_2^* \\ 0 & A_2 & b_2 \\ \hline c_1 & b_1^*T_1 & 0 \end{array}\right). \qquad \square$$

It was noted before that the codegree of a rational probability density function is at least 2. Therefore, the product of two such probability density functions has codegree at least 4. Hence for a random variable whose density is given by such a product, at least the first and second moments exist. This will be used in the next section to show the existence of a conditional mean and variance.

5.3. Convolution of probability densities. We now come to determine a state-space formulation for the convolution of two probability densities. Recall that if X and Y are two random variables with rational probability densities ρ_X and ρ_Y, then the probability density of $X+Y$ is given by the convolution $\rho_X * \rho_Y$.

Let ρ_1 and ρ_2 be two unnormalized rational probability functions with corresponding spectral summands Z_1 and Z_2. Let (A_i, b_i, c_i) be a realization of Z_i, $i = 1, 2$. Let, for $i = 1, 2$,

$$m_i^+(\tau) := \begin{cases} c_i e^{\tau A_i} b_i, & \tau \geq 0, \\ 0, & \tau < 0, \end{cases}$$

$$m_i^-(\tau) := \begin{cases} b_i^* e^{-\tau A_i^*} c_i^*, & \tau < 0, \\ 0, & \tau \geq 0. \end{cases}$$

Then $(\mathcal{F}m_i^+)(iw) = Z_i(iw)$, $(\mathcal{F}m_i^-)(iw) = Z_i^*(iw)$, $iw \in i\Re$, and

$$\begin{aligned}(\rho_1 * \rho_2)(w) &= \int_{-\infty}^{\infty} \Phi_1(iw - i\nu)\Phi_2(i\nu)d\nu = \mathcal{F}\left(\mathcal{F}^{-1}\int_{-\infty}^{\infty} \Phi_1(iw - i\nu)\Phi_2(i\nu)d\nu\right) \\ &= \mathcal{F}((\mathcal{F}^{-1}\Phi_1)(\mathcal{F}^{-1}\Phi_2))(iw) = \mathcal{F}((\mathcal{F}^{-1}(Z_1 + Z_1^*))(\mathcal{F}^{-1}(Z_2 + Z_2^*)))(iw) \\ &= \mathcal{F}((m_1^+ + m_1^-)(m_2^+ + m_2^-))(iw) = \mathcal{F}(m_1^+ m_2^+ + m_1^- m_2^-)(iw) \\ &= \mathcal{F}(m_1^+ m_2^+)(iw) + \mathcal{F}(m_1^- m_2^-)(iw).\end{aligned}$$

It follows that the spectral summand Z of $\rho_1 * \rho_2$ is given by $Z(iw) = \mathcal{F}(m_1^+ m_2^+)(iw)$.

In the following proposition we are going to give the state-space formulae for the product of the impulse responses of two single-input single-output state-space systems. This will be the key step to determine a state-space realization for the convolution of two rational probability density functions.

PROPOSITION 5.2. *Let $m_i^+(\tau) := c_i e^{\tau A_i} b_i$ for $\tau \geq 0$, and $m_i^+(\tau) := 0$ for $\tau < 0$, where (A_i, b_i, c_i) is an n_i-dimensional single-input single-output system, $i = 1, 2$. Then*

$$m^+(\tau) := m_1^+(\tau) m_2^+(\tau), \quad \tau \geq 0,$$

has a realization $m^+(\tau) = ce^{\tau A}b$ *for* $\tau \geq 0$ *and* $m^+(\tau) = 0$ *for* $\tau < 0$, *where*

$$A = A_1 \otimes I_{n_2} + I_{n_1} \otimes A_2,$$
$$b = b_1 \otimes b_2,$$
$$c = c_1 \otimes c_2.$$

(Here $\otimes$ *denotes the Kronecker product.)*

Proof. This follows immediately from basic rules on the Kronecker product (see, e.g., [8]), since for $\tau \geq 0$

$$\begin{aligned} m^+(\tau) &= ce^{\tau A}b = (c_1 \otimes c_2)e^{\tau(A_1 \otimes I_{n_2} + I_{n_1} \otimes A_2)}(b_1 \otimes b_2) \\ &= (c_1 \otimes c_2)(e^{\tau A_1} \otimes e^{\tau A_2})(b_1 \otimes b_2) = c_1 e^{\tau A_1} b_1 \otimes c_2 e^{\tau A_2} b_2 = c_1 e^{\tau A_1} b_1 c_2 e^{\tau A_2} b_2 \\ &= m_1^+(\tau) m_2^+(\tau). \qquad \square \end{aligned}$$

The proposition is of interest in its own right, as it allows one to find state-space formulas for products of impulse response functions.

Summarizing, we have the following result.

PROPOSITION 5.3. *Let* ρ_1 *and* ρ_2 *be unnormalized rational probability densities whose spectral summands* Z_1 *and* Z_2 *have the* n_1*-dimensional and* n_2*-dimensional state-space realizations* (A_1, b_1, c_1) *and* (A_2, b_2, c_2). *Then the density summand* Z *of the convolution* $\rho = \rho_1 * \rho_2$ *has the state-space realization*

$$\left(\begin{array}{c|c} A_1 \otimes I_{n_2} + I_{n_1} \otimes A_2 & b_1 \otimes b_2 \\ \hline c_1 \otimes c_2 & 0 \end{array} \right). \tag{1}$$

Proof. Suppose Z has the realization (1). Then the inverse Fourier transform of Z is $m_1^+ m_2^+$, showing that Z is the spectral summand of ρ. $\square$

Note that the state-space dimension of this realization is $n_1 n_2$, which implies that the McMillan degree of Z is at most $n_1 n_2$.

6. State-space expressions for the filtering equations. We are now in a position to derive state-space expressions for the unnormalized conditional densities in the filter equations which were discussed in the introduction.

THEOREM 6.1. *Assume the notation and assumptions for the filtering problem as presented in the introduction.*

Let $t \geq 0$, *and let* $(A_{x_{t|t-1}}, b_{x_{t|t-1}}, c_{x_{t|t-1}})$ *be a minimal* n_{x_t}*-dimensional state-space realization of the density summand of the unnormalized conditional density* $\rho_{x_t|\mathcal{Y}_{t-1}}$. *For* $t = 0$, *set* $\rho_{x_t|\mathcal{Y}_{t-1}} := \rho_{x_0}$, *the density of the initial state* x_0. *Let* $(A_{\eta_t}, b_{\eta_t}, c_{\eta_t})$ *be a minimal* n_{η_t}*-dimensional state-space realization of the density summand of the unnormalized rational density* ρ_{η_t} *of the noise random variable* η_t, *and let* $(A_{\epsilon_t}, b_{\epsilon_t}, c_{\epsilon_t})$ *be a minimal* η_{ϵ_t}*-dimensional state-space realization of the density summand of the unnormalized rational density* ρ_{ϵ_t} *of the noise random variable* ϵ_t, $t \geq 0$.

Let T_1 *be the unique solution to the equation*

$$\left(\frac{1}{h_t} A_{\epsilon_t} + iy_t I \right) T_1 + T_1 A_{x_{t|t-1}} + b_{\epsilon_t} c_{x_{t|t-1}} = 0,$$

and let T_2 *be the unique solution to the equation*

$$A_{x_{t|t-1}} T_2 + T_2 \left(\frac{1}{h_t} A_{\epsilon_t} + iy_t \right) + b_{x_{t|t-1}} c_{\epsilon_t} = 0.$$

Then the density summand of the unnormalized density $\rho_{x_t|\mathcal{Y}_t}$ has state-space realization

$$\left(\begin{array}{c|c} A_{x_{t|t}} & b_{x_{t|t}} \\ \hline c_{x_{t|t}} & 0 \end{array}\right) = \left(\begin{array}{cc|c} \frac{1}{h_t}A^*_{\epsilon_t} - iy_tI & c^*_{\epsilon_t}c_{x_{t|t}} & T_2^*c^*_{x_{t|t}} \\ 0 & A_{x_{t|t}} & t_{x_{t|t}} \\ \hline b^*_{\epsilon_t} & c_{\epsilon_t}T_1 & 0 \end{array}\right).$$

The density summand of $\rho_{x_{t+1}|\mathcal{Y}_t}$ has state-space realization

$$\begin{aligned} &\left(\begin{array}{c|c} A_{x_{t+1|t}} & b_{x_{t+1|t}} \\ \hline c_{x_{t+1|t}} & 0 \end{array}\right) \\ &= \left(\begin{array}{c|c} f_tA_{x_{t|t}} \otimes I_{n_{\eta_t}} + I_{n_{\epsilon_t}+n_{x_t}} \otimes A_{\eta_t} & b_{x_{t|t}} \otimes b_{n_{\eta_t}} \\ \hline c_{x_{t|t}} \otimes c_{\eta_t} & 0 \end{array}\right) \quad \textit{if } f_t > 0, \\ &= \left(\begin{array}{c|c} -f_tA^*_{x_{t|t}} \otimes I_{n_{\eta_t}} + I_{n_{\epsilon_t}+n_{x_t}} \otimes A_{\eta_t} & c^*_{x_{t|t}} \otimes b_{n_{\eta_t}} \\ \hline b^*_{x_{t|t}} \otimes c_{\eta_t} & 0 \end{array}\right) \quad \textit{if } f_t < 0. \end{aligned}$$

Proof. Since by assumption $h_t > 0$, the density summand of the density $q(x) = \rho_{\epsilon_t}(y_t - h_tx)$, $x \in \Re$, has the realization

$$\left(\begin{array}{c|c} \frac{1}{h_t}A^*_{\epsilon_t} - iy_tI & c^*_{\epsilon_t} \\ \hline b^*_{\epsilon_t} & 0 \end{array}\right).$$

As

$$\rho_{x_t|\mathcal{Y}_t}(x) = \rho_{\epsilon_t}(y_t - h_tx)\rho_{x_t|\mathcal{Y}_{t-1}}(x), \quad x \in \Re,$$

by Proposition 5.1 the density summand of ρ has the realization

$$\left(\begin{array}{cc|c} \frac{1}{h_t}A^*_{\epsilon_t} - iy_tI & c^*_{\epsilon_t}c_{x_{t|t-1}} & T_2^*c^*_{x_{t|t-1}} \\ 0 & A_{x_{t|t-1}} & b_{x_{t|t-1}} \\ \hline b^*_{\epsilon_t} & c_{\epsilon_t}T_1 & 0 \end{array}\right),$$

where T_1 is the unique solution to the equation

$$\begin{aligned} &\left(\frac{1}{h_t}A^*_{\epsilon_t} - iy_tI\right)^* T_1 + T_1A_{x_{t|t-1}} + b_{\epsilon_t}c_{x_{t|t-1}} \\ &\quad = \left(\frac{1}{h_t}A_{\epsilon_t} + iy_tI\right)T_1 + T_1A_{x_{t|t-1}} + b_{\epsilon_t}c_{x_{t|t-1}} = 0, \end{aligned}$$

and T_2 is the unique solution to the equation

$$\begin{aligned} &A_{x_{t|t-1}}T_2 + T_2\left(\frac{1}{h_t}A^*_{\epsilon_t} - iy_tI\right)^* + b_{x_{t|t-1}}c_{\epsilon_t} \\ &\quad = A_{x_{t|t-1}}T_2 + T_2\left(\frac{1}{h_t}A_{\epsilon_t} + iy_tI\right) + b_{x_{t|t-1}}c_{\epsilon_t} = 0. \end{aligned}$$

To obtain a state-space formula for the prediction step

$$\rho_{x_{t+1}|\mathcal{Y}_t} = \rho_{f_tx_t|\mathcal{Y}_t} * \rho_{\eta_t},$$

we use Proposition 5.3. We need to introduce two cases depending on the sign of f_t. If $f_t > 0$, the density summand of $\rho_{f_t x_t|\mathcal{Y}_t}$ has the realization

$$\left(\begin{array}{c|c} f_t A_{x_{t|t}} & b_{x_{t|t}} \\ \hline c_{x_{t|t}} & 0 \end{array}\right).$$

If $f_t < 0$, the density summand of $\rho_{f_t x_t|\mathcal{Y}_t}$ has the realization

$$\left(\begin{array}{c|c} -f_t A^*_{x_{t|t}} & c^*_{x_{t|t}} \\ \hline b^*_{x_{t|t}} & 0 \end{array}\right).$$

The remaining parts of the result now follow by Proposition 5.3. □

It should be noted that the presented state-space realizations are data dependent and, in particular, dependent on $\mathcal{Y}_t$. As the formulae that use Kronecker products show, the dimensions of the state-space representation can potentially grow very quickly as the number of data points increases. It should be pointed out, however, that the growth in complexity is inherent in the use of random variables with rational densities (see also [11]). If, however, the density summand corresponding to η_t has only McMillan degree 1, i.e., η_t has Cauchy distribution, then the Kronecker products reduce to standard multiplication and the prediction step does not lead to an increase in dimension. Also, if the density summand corresponding to ϵ_t has McMillan degree 1, i.e., ϵ_t has Cauchy distribution, then the matrix equations can be solved explicitly to give

$$T_1 = -b_{\epsilon_t} c_{x_{t|t-1}} \left(\left(\frac{1}{h_t} A_{\epsilon_t} + i y_t \right) I + A_{x_{t|t-1}} \right)^{-1},$$

$$T_2 = -\left(\left(\frac{1}{h_t} A_{\epsilon_t} + i y_t \right) I + A_{x_{t|t-1}} \right)^{-1} b_{x_{t|t-1}} c_{\epsilon_t}.$$

Note that the inverse exists, since $A_{x_{t|t-1}}$ has all eigenvalues in the open left half plane and $\frac{1}{h_t} A_{\epsilon_t} + i y_t$ has negative real part, because of the stability of A_{ϵ_t} and since $h_t > 0$. From the remark after Proposition 5.1, it follows that the conditional mean $E(x_t|\mathcal{Y}_t)$ and the corresponding conditional variance $E(x_t - E(x_t|\mathcal{Y}_t)^2|\mathcal{Y}_t)$ exist and can be calculated from the density summand realization $(A_{x_{t|t}}, b_{x_{t|t}}, c_{x_{t|t}})$ using the formulas given in Theorem 4.1.

Note that prediction is also possible using the formulas presented here. For example, the unnormalized rational conditional probability density of the output variable at time $t+1$, given the observations of the output until time t, is equal to $\rho_{y_{t+1|t}}(y) = \rho_{h_{t+1} x_{t+1|t}} * \rho_{\epsilon_{t+1}}$, and the spectral summand of this density can be calculated using the formulas of section 5.

7. Conclusions. State-space formulae have been developed for various operations on rational density functions, and it is shown how this can be used to treat the filtering problem in the case of a first order linear stochastic model with stochastically independent noise variables with rational probability densities and stochastically independent initial state with rational probability density. This makes such filters easy to program on present-day computers, using, e.g., a linear algebra package. If the number of observations is not very small, however, the order of the conditional rational densities will tend to grow quickly. Therefore, various schemes of order reduction for positive real functions may be of relevance in practical applications (see, e.g.,

[2]). The formulae presented can also be used for further theoretical research in the behavior of the optimal filter. It follows, for example, that the conditional mean of the present state, given present and past observations, is a rational function of the present and past observations, which could be further investigated. The formula that is presented for the realization of the product of impulse response functions appears to be important in its own right.

REFERENCES

[1] F. R. Gantmacher, *Matrix Theory*, Chelsea, New York, 1959.
[2] B. Hanzon and R. J. Ober, *Overlapping balanced canonical forms for various classes of linear systems*, Linear Algebra Appl., 281 (1998), pp. 171–225.
[3] B. Hanzon and R. L. M. Peeters, *A Faddeev sequence method for solving Lyapunov and Sylvester equations*, Linear Algebra Appl., 241–243 (1996), pp. 401–430.
[4] T. Kailath, *Linear Systems*, Prentice Hall, Englewood Cliffs, NJ, 1980.
[5] K. Koedijk, M. Schafgans, and C. de Vries, *The tail index of exchange rate returns*, J. International Economics, 29 (1990), pp. 93–108.
[6] T. Körner, *Fourier Analysis*, Cambridge University Press, Cambridge, UK, 1989.
[7] B. Mandelbrot, *The variation of certain speculative prices*, J. Business, 36 (1963), pp. 394–419.
[8] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, 2nd ed., Academic Press, New York, 1985.
[9] W. Rudin, *Real and Complex Analysis*, McGraw Hill, New York, 1987.
[10] W. J. Rugh, *Linear System Theory*, 2nd ed., Prentice Hall, Upper Saddle River, NJ, 1996.
[11] I. J. Steyn, *State Space Models in Econometrics*, Ph.D. thesis, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands, 1996.
[12] S. J. Taylor, *Modelling Financial Time Series*, Wiley, New York, 1986.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 3, pp. 741–776

SEQUENCING AND ROUTING IN MULTICLASS QUEUEING NETWORKS PART I: FEEDBACK REGULATION*

SEAN P. MEYN†

Abstract. This paper establishes new criteria for stability and for instability of multiclass network models under a given stationary policy. It also extends previous results on the approximation of the solution to the average cost optimality equations through an associated fluid model: It is shown that an optimized network possesses a fluid limit model which is itself optimal with respect to a total cost criterion.

A general framework for constructing control algorithms for multiclass queueing networks is proposed based on these general results. Network sequencing and routing problems are considered as special cases. The following aspects of the resulting *feedback regulation policies* are developed in the paper:

(i) The policies are stabilizing and are, in fact, geometrically ergodic for a Markovian model.

(ii) Numerical examples are given. In each case it is shown that the feedback regulation policy closely resembles the average-cost optimal policy.

(iii) A method is proposed for reducing variance in simulation for a network controlled using a feedback regulation policy.

Key words. queueing networks, routing, scheduling, optimal control

AMS subject classifications. Primary, 90B35, 68M20, 90B15; Secondary, 93E20, 60J20

PII. S0363012999362724

1. Introduction.

This paper concerns the effective management of large networks through scheduling and routing.

Specific applications of interest include cellular and internet communication systems, large scale manufacturing processes, and computer systems (see, e.g., [5, 55, 22]). In spite of the diversity of these applications, one can find many common goals:

(i) Controlling delay, throughput, inventory, and loss. The crudest issue is *stability*: do queue lengths remain bounded for all time?

(ii) Estimating performance, or comparing the performance of one policy over another one. *Performance* is context-dependent, but common metrics are average delay and loss probabilities.

(iii) Prescriptive approaches to policy synthesis which are intuitive, flexible, robust, and reasonable in complexity. *Robustness* means that the policy will be effective even under significant modeling error. By *flexibility* we mean that the policies will react appropriately to changes in network topology or other gross structural changes.

(iv) In applications to telecommunications or power systems one may have limited information. Routing or sequencing decisions must then be determined using only that information which can be made available. This issue is becoming less critical with ever-increasing information processing power. In the future internet it may be possible to assume essentially complete information at every node in the network through current flooding algorithms and proposed *explicit congestion notification algorithms* [21].

There are currently several popular classes of models which describe in various

*Received by the editors October 11, 1999; accepted for publication (in revised form) March 26, 2001; published electronically September 28, 2001. This research was supported in part by NSF grants ECS 9403742 and ECS 9972957.

http://www.siam.org/journals/sicon/40-3/36272.html

†Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, 1308 W. Main Street, Urbana, IL 61801 (s-meyn@uiuc.edu, http://black.csl.uiuc.edu:80/~meyn).

levels of detail the dynamics of a queueing network to address the range of issues in (i)–(iv). The utility of a particular model depends upon one's particular goals.

A traditional academic approach to policy synthesis is to construct a Markov decision process (MDP) model for the network. This involves constructing a controlled transition operator $P_a(x, y)$, which gives the probability of moving from state x to state y when the control decision $a \in \mathcal{A}(x)$ is applied. The state space X (where x and y live) is typically taken as the set of all possible buffer levels at the various stations in the network; $\mathcal{A}(x)$ is then the set of feasible control actions when the state takes the value $x \in \mathsf{X}$. Given an MDP model and a one step cost function $c\colon \mathsf{X} \to \mathbb{R}_+$, a solution to the average cost optimal control problem is found by solving the resulting dynamic programming equations,

$$\eta^* + h^*(x) = \min_{a \in \mathcal{A}(x)} [c(x) + P_a h^*(x)], \tag{1.1}$$

$$F^*(x) = \arg\min_{a \in \mathcal{A}(x)} P_a h^*(x), \qquad x \in \mathsf{X}. \tag{1.2}$$

The function F^* on X then defines an optimal stationary policy.

The difficulty with this approach is very well known: When buffers are infinite, this becomes an infinite dimensional optimization problem. Even when considering finite buffers, the complexity grows exponentially with the dimension of the state space. Some form of aggregation is necessary—the Markovian model is simply too detailed to be useful in optimization.

An elegant approach is to consider the model in heavy traffic where a reflected Brownian motion model is appropriate. The papers [24, 28] develop these ideas for the network scheduling or sequencing problems, and [32] considers routing and other control problems. One is then faced with optimizing a controlled stochastic differential equation (SDE) model. In many examples considered in the literature this control problem has a simple intuitive solution. This is just one example of a fluid model for the physical network. Another popular model is the "σ-ρ" constrained fluid model [11, 51] and the linear fluid model considered here (see, e.g., [8, 7, 56, 41, 42, 2]). Any one of these models is valuable in network design because unimportant details are stripped away.

Justification for considering these various idealizations comes from theory that establishes solidarity between idealized fluid models and more accurate discrete models, when the load is close to capacity [32, 4], or the state of the system is large (e.g., the network is congested [45], or a "large deviation" occurs [54]). Stability theory for networks, as in (i), has essentially reached maturation over the past decade, following counterexamples introduced in [35, 52]. This theory is based on the close ties between a stochastic network model and its linear fluid counterpart [14, 15, 16].

There are, however, several difficulties with these approaches:

- The Brownian motion approximation is based on a model in heavy traffic. If the stations are not balanced then one loses some information at the stations which are not heavily loaded.
- Although the optimal control problem for a Brownian motion or σ-ρ constrained fluid model often has a simple intuitive solution, frequently this is not the case. There is currently no general practical method for generating policies.
- It is not always obvious how to translate an optimal policy for an abstract model to a feasible and efficient policy for the original discrete model.

In this paper, we consider exclusively the linear fluid model (2.2) in design. Theorem 3, an extension of the main result of [45], establishes a strong form of solidarity between the discrete optimization problem (1.1) (1.2) and a related total-cost optimal control problem for the linear fluid model. These results can be generalized to show that an optimized SDE model possesses a fluid limit model which is itself optimal with respect to the total-cost criterion. Hence to optimize the Brownian motion model one must also solve the linear fluid model optimization problem.

A translation of a policy from the fluid model to the original discrete model of interest is again not obvious. This issue is addressed in [42, 2], where it is shown that virtually any "fluid trajectory" can be approximately tracked using a discrete policy for the discrete-stochastic network. In [13, 45] the results from several numerical studies are described. It is found that the optimal discrete policy resembles the optimal fluid policy, but with the origin θ for the model (2.2) shifted to some value $\bar{x} \in \mathbb{R}_+^\ell$. From these results it is argued that one should use the fluid model to attempt to regulate the state to some deterministic value $\bar{x}$. In the numerical studies considered, it was found that the average cost was nearly optimal and that the variance at each buffer was *lower* than that obtained using an optimal policy. The computation of optimal policies for the linear fluid model appears to be feasible for network models of moderate complexity [56, 41, 49].

A final motivation for considering the simplest network model follows on considering our main goal: robust policy synthesis. As described in (iii) above, any policy that we construct should be sufficiently robust so that it will tolerate modeling errors resulting from uncertain variability in service or arrival rates.

The main results of the present paper builds upon those of [45, 44]:

(i) The class of models is extended to include routing and processor sharing models as well as the scheduling models considered earlier. This requires that we introduce a notion of stabilizability for networks, which is a generalization of the usual capacity conditions.

(ii) The underlying theory is improved. It is shown that optimal policies have optimal fluid limits with respect to the total cost criterion. This improves upon the main result of [45, 44], which required specialization to a class of "fluid limit models" obtained via weak convergence. Moreover, the stability theory relating networks and their fluid models is improved to give criteria for geometric ergodicity and simpler conditions implying transience of the controlled network.

(iii) The practical usefulness of the approach is improved by borrowing from the BIGSTEP approach of [28] and by exploring reduced complexity control approaches for the fluid model.

(iv) Numerical examples are given. In each case it is shown that the feedback regulation policy closely resembles the average-cost optimal policy.

(v) Consideration of the fluid model leads to an approach to estimating steady state performance indicators, such as mean delay, through simulation. This requires care since standard Monte Carlo simulation is known to have high variance for highly loaded networks.

The viewpoint arrived at in this paper leads to policies which are similar to those found through a heavy traffic analysis using a Brownian motion approximation. Consider, for example, the models treated in [32]. In each case one could perform designs on the fluid model, translate these policies as in (4.1), and arrive at the same policy that was obtained using a Brownian motion approximation. Given the greater complexity of the Brownian motion model, we conclude that while diffusion

approximations are tremendously useful for analysis and performance approximation, presently they appear to be less useful for the purposes of control design. This will change if more efficient numerical methods can be devised for control synthesis in SDE models [38].

One of the most important benefits of a heavy traffic assumption is that the resulting "state space collapse" can result in a model of reduced complexity. In the models considered in the aforementioned references, in each case one is left with a one dimensional state process which captures all relevant information. This model reduction is obtained for *either* model, fluid or SDE, when the system load approaches a critical value. The aim of Part II, the sequel to the present paper, is to exploit this observation to prove that, under certain geometric conditions, an optimal policy for the fluid model may be translated to form a policy which is approximately optimal for the stochastic model [43].

The remainder of the paper is organized as follows. Section 2 describes the class of models considered and their associated fluid limit model. Here some general stability theory for networks and their fluid models is presented, including criteria for geometric ergodicity. In section 3 this solidarity between the fluid model and the discrete network is extended. It is shown that, provided the fluid model is stabilizable, there exists an average cost optimal policy whose fluid model is optimal with respect to the total-cost criterion. Several examples are given to illustrate the relationship between the two optimization problems for specific models. The *feedback regulation* policies are introduced in section 4. Several classes of stabilizing fluid policies are described, and a stability proof is provided in this section. Conclusions are postponed to Part II.

2. Networks and their fluid limit models. The networks envisioned here consist of a finite set of resources, a finite set of buffers, and various customers classes which arrive to the network for processing. Resources perform various activities, which transform parts or customers at the various buffers. On completion of service, a customer either leaves the network or visits another resource for further processing. This is the intuitive definition of a multiclass queueing network. A popular continuous-time model is given by

$$Q(t;x) = x - S(Z(t;x)) + R(Z(t;x)) + A(t), \qquad t \geq 0. \tag{2.1}$$

The vector-valued stochastic process $Q(t;x)$ denotes the buffer levels at time t with initial condition $Q(0;x) = x \in \mathbb{R}^{\ell}$. Some of these buffers may be *virtual*. In a manufacturing model, such as that shown in Figure 1, virtual buffers may correspond to backlog or excess inventory.

The vector-valued stochastic process $Z(t;x)$ is the *allocation* (or *control*). The ith component $Z_i(t;x)$ gives the cumulative time that the activity i has run up to time t.

The vector-valued process $\boldsymbol{A}$ may denote a combination of exogenous arrivals to the network and exogenous demands for materials *from* the network. The vector-valued functions $S(\,\cdot\,), R(\,\cdot\,)$ represent, respectively, the effects of random service rates and the effects of a combination of possibly uncontrolled, possibly random routing, and random service rates.

The *fluid limit model* is obtained on considering a congested network. When $Q(t;x)$ is large in magnitude, variations in the arrival and service processes appear small when compared with the state. The behavior of $\boldsymbol{Q}$ when viewed on this large spatial scale will appear deterministic and can be approximated by the *mean field*

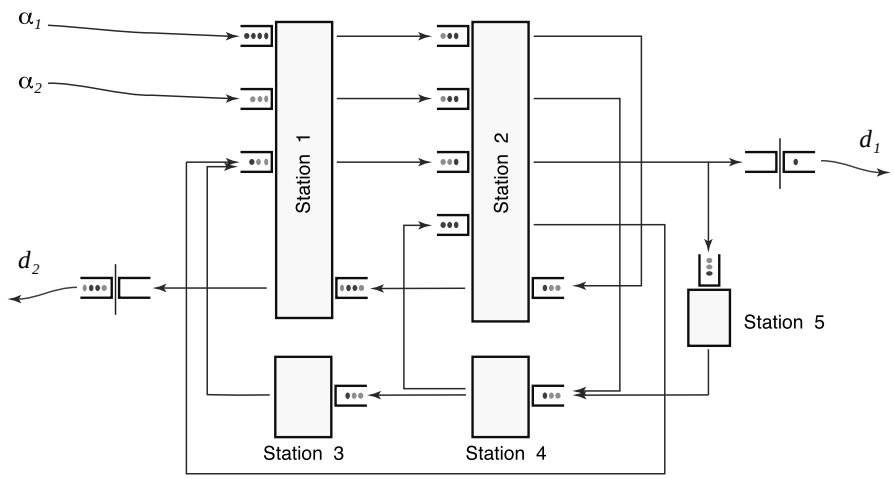


FIG. 1. *A network with many buffers, controlled routing, uncontrolled routing, multiple demands, and virtual buffers.*

equations, or *(linear) fluid model*,

$$q(t;x) = x + Bz(t;x) + \alpha t, \qquad t \geq 0. \tag{2.2}$$

Here B is a matrix of appropriate dimension, interpreted as a long-run average of $\boldsymbol{R} - \boldsymbol{S}$, and α is a long-run average of $\boldsymbol{A}$.

There are several ways of making this precise, and very few assumptions are required to ensure the existence of a well-defined fluid limit model. A construction is provided in section 2.1, and section 2.2 describes a stability theory for (2.1) based on the simpler model (2.2).

Section 2.1 introduces a discrete-time countable state space MDP model. In the special case where all of the driving processes $(\boldsymbol{A}, \boldsymbol{R}, \boldsymbol{S})$ are multidimensional Poisson processes, the discrete-time model is obtained from (2.1) via *uniformization* [40]. The MDP model is convenient for the purposes of optimization and also provides the simplest setting for constructing the fluid limit model through scaling $\boldsymbol{Q}$ and $\boldsymbol{Z}$.

2.1. A Markovian network model and its fluid limit. Consider the M/M/1 queue, described in continuous time by

$$Q(t;x) = x - S(Z(t;x)) + A(t), \qquad t \geq 0,$$

where the *cumulative busy time* $\boldsymbol{Z}$ satisfies $\frac{d}{dt} Z(t;x) = 1$ whenever $Q(t;x) \neq 0$. The stochastic processes $(\boldsymbol{A}, \boldsymbol{S})$ are one dimensional Poisson processes with rates α, μ, respectively. The fluid model is given by the analogous equation,

$$q(t;x) = -\mu z(t) + \alpha t, \qquad t \geq 0. \tag{2.3}$$

To obtain a discrete-time process, one might sample at successive jump times of $\boldsymbol{Q}$, but this would introduce bias. When $Q(t;x) = 0$, only upward jumps are possible; hence sampling is less frequent in this situation, and the overall sampling rate is *nonuniform*. Consequently, the steady-state queue length for the sampled process is strictly larger than that of the unsampled queue.

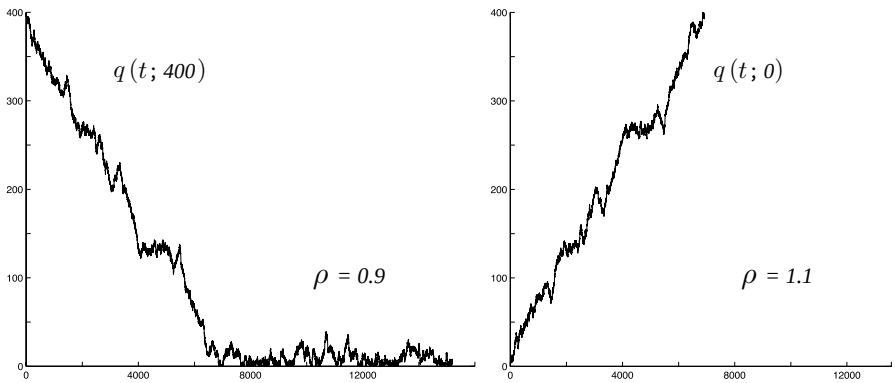


FIG. 2. *The M/M/1 queue: In the stable case on the left we see that the process $Q(t;x)$ appears piecewise linear with a relatively small high frequency "disturbance." The process explodes linearly in the unstable case shown at right.*

Uniformization corrects this by introducing *virtual service times* and by sampling $\boldsymbol{Q}$ at times of arrivals, service completions, or virtual service completions. For example, the first sampling time is of the form $\tau_1 = \min(S_1, T_1)$, where S_1 and T_1 are exponentially distributed random variables with mean $1/\mu$ and $1/\alpha$, respectively. If $Q(0;x) = x > 0$, then S_1 is the remaining service time for the customer currently in service. If $x = 0$, then S_1 is again exponential with mean $1/\mu$, but it is now a random variable which is independent of the original queue length process. Sampling results in a discrete-time Markov chain with transition probabilities:

$$P(x, x+1) = \alpha, \qquad P(x, (x-1)_+) = \mu, \qquad x = 0, 1, 2, 3, \ldots .$$

When $\rho = \alpha/\mu$ is less than one, then the queue is *positive recurrent*, as shown in the left-hand side of Figure 2.

The discrete-time M/M/1 queue model may be viewed as a random linear system,

$$Q(k+1) = Q(k) + \widetilde{B}(k+1)U(k) + \widetilde{\alpha}(k+1), \tag{2.4}$$

where the sequence $\boldsymbol{U}$ is defined again by the nonidling policy $U(k) = \mathbb{I}(Q(k) > 0)$, $k \geq 0$, and $\{\widetilde{B}(k), \widetilde{\alpha}(k) : k \geq 1\}$ is an indendent and identically distributed (i.i.d.) sequence satisfying,

$$\begin{pmatrix} \widetilde{B}(k) \\ \widetilde{\alpha}(k) \end{pmatrix} = \begin{cases} -e^1 & \text{with prob } \mu, \\ e^2 & \text{with prob } \alpha, \end{cases}$$

with e^i equal to the standard basis element in $\mathbb{R}^2$, $i = 1, 2$. The general network model may be sampled in the same way to obtain a similar, multidimensional model whenever $(\boldsymbol{A}, \boldsymbol{R}, \boldsymbol{S})$ are Poisson.

Since we will not consider again the continuous-time model, we will denote by $(\boldsymbol{Q}, \boldsymbol{U}) = \{(Q(k;x), U(k;x)) : k \geq 0\}$ the state-allocation process for a discrete-time model with initial condition x. (This dependency will be suppressed when the particular initial condition is not relevant.) We assume that there are ℓ buffers and ℓ_u activities so that $(\boldsymbol{Q}, \boldsymbol{U})$ evolves on $\mathbb{Z}_+^{\ell} \times \mathbb{Z}_+^{\ell_u}$. As in the continuous-time case, each $U_i(k)$, $1 \leq i \leq \ell_u$, takes binary values. The discrete-time Markovian model is then *defined* as the random linear system (2.4).

We assume that the sequence $\{\widetilde{B}(k), \widetilde{\alpha}(k) : k \geq 1\}$ is i.i.d.. For each k, the matrix $\widetilde{B}(k)$ has dimension $\ell \times \ell_u$, the vector $\widetilde{\alpha}(k)$ has dimension ℓ, and the components of both take on integer values, again strictly bounded. We assume, moreover, that the components of $\widetilde{\alpha}(k)$ are nonnegative.

This is a generalization of the sampled model, in which the entries of $(\widetilde{B}(k), \widetilde{\alpha}(k))$ take on the values $(-1, 0, 1)$ only. However, this discrete-time model covers only a very narrow set of stochastic network models. For example, it is not possible to convert the natural continuous time model into a countable-state, discrete-time MDP if service times are uniformly distributed. We restrict ourselves to the simple discrete-time model for the sake of exposition only. General distributions are considered in [14, 15], where it is shown that stability theory goes through without change. To generalize the results of the present paper, e.g., Theorem 4, one must assume a bounded hazard rate as in [46] to ensure that the mean forward recurrence time is bounded. Part II, which does not require a Markovian description, develops the general model (2.1) [43].

We assume that there is an integer $\ell_m \geq 1$ and an $\ell_m \times \ell_u$ *constituency matrix* C, such that

$$CU(k) \leq \mathbf{1}, \qquad k \geq 0,$$

where $\mathbf{1}$ denotes a vector of ones. The entries of C take on binary values, and each row defines a *resource*: The ith resource $\mathcal{R}_i$ is defined to be the set of activities j such that $C_{ij} = 1$.

There may also be auxiliary constraints on the control sequence $\boldsymbol{U}$ and further constraints on $\boldsymbol{Q}$. For example, buffers may require synchronous processing, or strict limits on buffer levels may be imposed. We assume that these may be expressed through linear constraints,

$$C_a U(k) \leq b_a, \qquad C_s Q(k) \leq b_s, \qquad k \geq 0,$$

for matrices C_a, C_s and vectors b_a, b_s of appropriate dimension.

We have thus restricted $(\boldsymbol{Q}, \boldsymbol{U})$ to lie in the countable sets

$$Q(k) \in \mathsf{X} \cap \mathbb{Z}^{\ell} \quad U(k) \in \mathsf{U} \cap \mathbb{Z}^{\ell_u}, \qquad k \geq 0,$$

where

$$\mathsf{U} := \{\zeta \in \mathbb{R}_+^{\ell_u} : \zeta \geq \theta,\ C\zeta \leq \mathbf{1},\ C_a\zeta \leq b_a\}, \tag{2.5}$$

$$\mathsf{X} := \{x \in \mathbb{R}_+^{\ell} : x \geq \theta,\ C_s x \leq b_s\}. \tag{2.6}$$

We assume throughout the paper that U is bounded. Unless noted otherwise, we assume that C_a and C_s are null.

The sequence $\boldsymbol{U}$ is an adapted (history-dependent) stochastic process. We say that $\boldsymbol{U}$ is defined by a *stationary policy* if there is a feedback function $F\colon \mathsf{X} \to \mathsf{U}$ satisfying

$$\mathsf{P}(U_i(k) = 1 \mid Q(0), \ldots, Q(k)) = F_i(Q(k)), \qquad k \geq 0.$$

The policies we consider are primarily stationary or are based on such policies.

In the classical *scheduling model*, the ℓ_m resources $\{\mathcal{R}_i : 1 \leq i \leq \ell_m\}$ are called *stations*. There is one activity for each customer class, giving $\ell_u = \ell$. Class i customers wait in the ith queue, if necessary, and then receive service via the ith activity. Upon

completion of service, a class i customer becomes a class j customer with probability R_{ij} and exits the system with probability $R_{i0}:=1-\sum_j R_{ij}$. The constraint $CU(k) \le \mathbf{1}$ is the usual condition that no two customers receive service simultaneously at a single station.

To construct the fluid limit model, first consider the time-invariant means, given by

$$B = \mathsf{E}[\widetilde{B}(k)], \quad \alpha = \mathsf{E}[\widetilde{\alpha}(k)], \qquad k \ge 1.$$

We may then write

$$Q(k+1) = Q(k) + BU(k) + \alpha + D(k+1), \tag{2.7}$$

where the process $\boldsymbol{D}$ is bounded, and it is a martingale difference sequence with respect to the natural filtration. In this way, the model (2.4) may be viewed as a deterministic "fluid model" with a bounded "disturbance" $\boldsymbol{D}$. When the initial condition $Q(0)$ is large, then the state dominates this disturbance, and the network behavior appears deterministic.

The fluid limit model considered in this paper is obtained by scaling the process, both temporally and spatially, through a scaling parameter $n \in \mathbb{Z}_+$. For any $x \in \mathbb{R}_+^\ell$, $n \ge 1$, we define

$$q^n(t;x) = \frac{1}{n} Q(nt; nx), \tag{2.8}$$

$$z^n(t;x) = \frac{1}{n} \sum_{i \le nt} U(i;x), \qquad t = \frac{0}{n}, \frac{1}{n}, \frac{2}{n}, \dots, \tag{2.9}$$

where we are taking the integer part of nx whenever necessary so that the initial condition nx lies in $\mathbb{Z}_+^\ell$. We then extend the definition of $\{q^n(t;x), z^n(t;x)\}$ to arbitrary $t \in \mathbb{R}_+$ so that these processes are linear on each time segment $[i/n, (i+1)/n]$ and continuous on $\mathbb{R}_+$.

We have $q^n(0;x) = x$ and $z^n(0;x) = \theta$, and $\{q^n(\,\cdot\,;x), z^n(\,\cdot\,;x)\}$ are Lipschitz continuous for any n and x. Typically, we find that $\{\boldsymbol{q}^n, \boldsymbol{z}^n\}$ converges to a limiting, deterministic function of time $\{\boldsymbol{q}, \boldsymbol{z}\}$ as $n \to \infty$. The limits $\boldsymbol{q}$ and $\boldsymbol{z}$ are piecewise linear functions of t in all of the examples considered below. Figure 2 illustrates the nature of this convergence for the M/M/1 queue, where the limiting process satisfies (2.3).

For each x, the set $\mathcal{L}_x$ denotes all weak limits of $\{(q^n(\,\cdot\,;x), z^n(\,\cdot\,;x)) : n \ge 1\}$ as $n \to \infty$. The *fluid limit model*, denoted $\mathcal{L}$, is the union of $\mathcal{L}_x$ over all initial conditions. Any $(q,z) \in \mathcal{L}$ satisfies (2.2), together with the rate constraints

$$C[z(t) - z(s)] \le (t-s)\mathbf{1}, \quad z(t) - z(s) \ge \theta, \qquad t \ge s \ge 0. \tag{2.10}$$

That is, $\frac{z(t)-z(s)}{t-s} \in \mathsf{U}$ for any $t \ne s$.

2.2. Stability of the models. Stability of the network under some policy requires some assumptions on the model. We say that $\boldsymbol{z}$ is a *feasible allocation* for the fluid model if the resulting state trajectory $\boldsymbol{q}$ satisfying (2.2), (2.10) remains in X for all $t \ge 0$. The fluid model is said to be

- *stabilizable* if, from any initial condition $x \in \mathsf{X}$, there exist $T_\theta < \infty$ and a feasible allocation $\boldsymbol{z}$ such that

$$q(t;x) = x + \alpha t + Bz(t) = \theta, \qquad t \ge T_\theta;$$

- *controllable* if for any pair $x, y \in \mathsf{X}$, there is a feasible allocation $\boldsymbol{z}$ and a time T such that $q(T; x) = y$.

Note that one can assume without loss of generality that an allocation $\boldsymbol{z}$ driving x to y is *linear*.

PROPOSITION 1. *Suppose that $x, y \in \mathsf{X}$, $T > 0$, and $\boldsymbol{z}$ is an allocation satisfying*

$$q(T; x) = x + \alpha T + Bz(T) = y.$$

Then the linear allocation $z^1(t) = \bar{z}t$, $0 \le t \le T$, also brings $\boldsymbol{q}$ to y from x at time T and satisfies (2.10) *on* $[0, T]$.

A necessary condition for stabilizability is that there must exist some solution to the equilibrium equation

$$B\zeta^{\text{ss}} + \alpha = \theta, \qquad \zeta^{\text{ss}} \in \mathsf{U}. \tag{2.11}$$

In the special case of network scheduling, the $\ell \times \ell$ matrix B has the form

$$B = -(I - R^{\text{T}})M, \tag{2.12}$$

where M is the diagonal matrix with diagonal entries $\mu^{\text{T}} = (\mu_1, \dots, \mu_\ell)$. There is a unique solution to the equilibrium equation (2.11), given by $\zeta^{\text{ss}} = -B^{-1}\alpha$, and the standard load condition may be written as

$$\vec{\rho} = -C\zeta^{\text{ss}} = M^{-1}(I - R^{\text{T}})^{-1}\alpha < \mathbf{1}. \tag{2.13}$$

It is readily seen that the load condition implies stabilizability. In routing models and many other examples, the "load" at a station is policy-dependent [32].

To obtain sufficient conditions for stabilizability, it is convenient to envision (2.2) as a differential inclusion,

$$\dot{q} \in \mathsf{V} := \{B\zeta + \alpha : \zeta \in \mathsf{U}\} \subset \mathbb{R}^\ell.$$

The set V is equal is the set of possible velocity vectors for the fluid model. We let $-\mathsf{V}$ denote its reflection. The proof of the following result is obvious and will be omitted. In Proposition 2 the set $B(\theta, \varepsilon)$ denotes the open ball of radius ε, centered at the origin.

PROPOSITION 2. *The fluid model* (2.2) *is*

(i) *stabilizable if and only if there exists $\varepsilon > 0$ such that*

$$B(\theta, \varepsilon) \cap \mathbb{R}^\ell_+ \subset \{-\mathsf{V}\} \cap \mathbb{R}^\ell_+,$$

(ii) *controllable if and only if there exists $\varepsilon > 0$ such that $B(\theta, \varepsilon) \subset \mathsf{V}$.*

Either of the conditions (i) or (ii) can be formulated as a finite linear program. For example, the following set of constraints summarizes the condition that V contains each of the vectors $\{-\varepsilon e^i : 1 \le i \le \ell\}$.

$$\begin{aligned} B\zeta^i + \alpha &= -\varepsilon_i e^i, \\ C\zeta^i &\le \mathbf{1}, \\ \zeta^i &\ge \theta, \qquad 1 \le i \le \ell. \end{aligned}$$

A related linear program is devised to define the system load in [26, 25].

We now turn to the discrete-stochastic network.

When controlled by a randomized stationary policy with feedback law F, the state process becomes a time-homogeneous Markov chain. The state transition matrix is denoted $P_{\scriptscriptstyle F}$—the subscript is supressed when there is no risk of ambiguity. *Stability* of the controlled process is defined as positive recurrence of the resulting Markov chain. Under stability, there exists a unique invariant probability $\pi = \pi_{\scriptscriptstyle F}$, and steady-state performance measures such as mean delay or average total congestion can be described in terms of the invariant probability.

We assume that, under the transition law P, the state process possesses a single communicating class $\mathcal{C}$ which contains the origin θ. We assume, moreover, that the controlled system is ψ-irreducible and aperiodic, as defined in [47]. Defining the first return time to a set $A \subseteq \mathsf{X}$ by

$$\tau_A = \min(k \geq 1 : Q(k) \in A),$$

the ψ-irreducibility condition can be expressed as $\mathsf{P}_x(\tau_\theta < \infty) > 0$ for any $x \in \mathsf{X}$. This is typically a minor constraint on the policy. For the network scheduling problem these conditions hold when the policy is nonidling.

Throughout the paper we use $c\colon \mathbb{R}_+^\ell \to \mathbb{R}_+$ to denote a norm, i.e., it is continuous, convex, vanishes only at θ, and it is radially homogeneous. The function c will be interpreted as a one step cost function for the model. For a particular stationary policy, the controlled chain is called *c-regular* if, for any initial condition x,

$$\mathsf{E}_x \left[\sum_{i=0}^{\tau_\theta - 1} c(Q(i)) \right] < \infty.$$

A c-regular chain always possesses a unique invariant probability π such that

$$\pi(c) := \sum_{x \in \mathsf{X}} c(x)\, \pi(x) < \infty.$$

A stationary Markov policy (and its associated feedback function F) is called *regular* if the controlled chain is c-regular. In this case it follows from the f-norm ergodic theorem of [47, Chapter 14] that the following average cost exists and is independent of the initial condition x:

$$\begin{aligned} &\text{(i)} \qquad J(x,w) = \lim_{k\to\infty} \mathsf{E}_x[c(Q(k))] \;\; = \;\; \pi(c), \\ &\text{(ii)} \qquad\quad \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} c(Q(k)) \;\; = \;\; \pi(c), \qquad \text{almost surely (a.s.).} \end{aligned}$$

The fluid limit model is said to be *stable* if there exist $\varepsilon > 0$ and $T < \infty$ such that $q(T;x) = \theta$ for any $q \in \mathcal{L}_x$ with $\|x\| \leq \varepsilon$. It will be called *L_p-stable* if, for some $\varepsilon > 0$,

$$\lim_{t\to\infty} \sup_{q \in \mathcal{L}_x : \|x\| \leq \varepsilon} \mathsf{E}[\|q(t)\|^p] = 0.$$

The following result is a minor generalization of [45, Theorem 5.2]. Related results are found in [20, 52, 14, 15].

THEOREM 3. *The following stability criteria are equivalent for the network under any nonidling, stationary Markov policy.*

(i) *There exist $b_0 < \infty$ and a function $V\colon \mathbb{R}^\ell \to \mathbb{R}_+$ such that the following drift condition holds:*

$$PV(x) := \mathsf{E}_x[V(Q(k+1)) \mid Q(k) = x] \le V(x) - c(x) + b_0, \qquad x \in \mathsf{X}. \tag{2.14}$$

The function V is equivalent to a quadratic in the sense that, for some $\varepsilon > 0$,

$$1 + \varepsilon\|x\|^2 \le V(x) \le 1 + \varepsilon^{-1}\|x\|^2, \qquad x \in \mathsf{X}.$$

(ii) *For some quadratic function V and some $b_0 < \infty$,*

$$\mathsf{E}_x\left[\sum_{n=0}^{\tau_\theta} c(Q(n))\right] \le V(x) + b_0, \qquad x \in \mathsf{X}.$$

(iii) *For some quadratic function V and some $b_0 < \infty$,*

$$\sum_{n=1}^{N} \mathsf{E}_x[c(Q(n))] \le V(x) + b_0 N \quad \text{for all } x \text{ and } N \ge 1.$$

(iv) *The fluid limit model is L_2-stable.*

(v) *The total cost for the fluid limit is uniformly bounded in the sense that, for some quadratic function V,*

$$\mathsf{E}\left[\int_0^\infty \|q(\tau; x)\|\, d\tau\right] \le V(x), \qquad x \in \mathbb{R}_+^\ell,\ q \in \mathcal{L}_x.$$

If any of these equivalent conditions, hold then the stationary policy is regular.

For a well-designed policy the controlled chain will be stable in a far stronger sense. A Markov chain $\boldsymbol{Q}$ is called *V-uniform ergodic*, with $V\colon \mathbb{R}^\ell \to [1, \infty)$ a given function, if there exist $\gamma < 1$ and $b < \infty$ such that

$$|\mathsf{E}[g(Q(k)) \mid Q(0) = x] - \pi(g)| \le b\gamma^k V(x), \qquad k \in \mathbb{Z}_+,\ x \in \mathsf{X},$$

where g is any function satisfying $|g(x)| \le V(x)$, $x \in \mathsf{X}$ (see [47, Chapter 17]). A Markov chain satisfying this strong form of ergodicity is similar to an i.i.d. process. In particular, a V-uniform Markov chain satisfies a strong form of the large deviations principle [3, 33].

This stronger form of stability holds under uniform convergence to the fluid limit. The following two forms of uniform convergence will be assumed on a given set $S \subset \mathsf{X}$ of initial conditions. For a set $Y \subseteq \mathbb{R}^\ell$ and a point $x \in \mathbb{R}^\ell$, we define

$$d\{x, Y\} = \inf(\|x - y\| : y \in Y).$$

Similarly, if $\mathcal{F} \subseteq C([0,T], \mathbb{R}^{\ell+\ell_u})$ is a set of functions and $q \in C([0,T], \mathbb{R}^{\ell+\ell_u})$ is another function, then we define

$$d\{q, \mathcal{F}\} = \inf_{\psi \in \mathcal{F}} \sup_{0 \le t \le T} \|q(t) - \psi(t)\|.$$

(U1) For any given T, ε there is a sequence $\{\Theta(\varepsilon, T, n)\}$ such that for any $x \in S$,

$$\mathsf{P}\Big(d\{q^n(T; x), Y_x(T)\} > \varepsilon\Big) \le \Theta(\varepsilon, T, n) \to 0, \qquad n \to \infty,$$

where $Y_x(T) = \{q(T; x) : q \in \mathcal{L}_x\} \subseteq \mathbb{R}_+^\ell$.

(U2) For any given T, ε there is a sequence $\{\Theta(\varepsilon, T, n)\}$ such that for any $x \in S$,

$$\mathsf{P}\Big(d\{q^n(\,\cdot\,;x), \mathcal{L}_x\} > \varepsilon\Big) \le \Theta(\varepsilon, T, n) \to 0, \qquad n \to \infty.$$

These conditions are frequently automatic since the functions $\big\{q^n(\,\cdot\,;x) - q(\,\cdot\,;x), z^n(\,\cdot\,,x) : n \ge 1, x \ne \theta\big\}$ are uniformly bounded and uniformly Lipschitz continuous.

A stable fluid limit model always admits a *Lyapunov function* $V\colon \mathbb{R}_+^\ell \to \mathbb{R}_+$ satisfying

$$V(q(t;x)) \le V(x) - t \qquad \text{for } t < \tau_\theta = \min(t : q(t;x) = \theta). \tag{2.15}$$

One can take the maximal emptying time itself and, moreover, this Lyapunov function is radially homogeneous. Conversely, the existence of a Lyapunov function satisfying (2.15) is known to imply stability.

If there exists a *Lipschitz continuous* Lyapunov function, then one can deduce not just stability but robustness with respect to parametric perturbations. It is not surprising then that the existence of a Lipschitz Lyapunov function implies a form of exponential stability.

THEOREM 4. *Suppose that the network is controlled using a stationary policy, and that there exists $b_0 < \infty$ such that, with $S = \{x : \|x\| \ge b_0\}$,*

(i) *assumption* (U1) *holds on S;*

(ii) *there exists a Lipschitz continuous function $V\colon \mathbb{R}_+^\ell \to \mathbb{R}_+$ satisfying* (2.15) *for $x \in S$.*

Then the network is V_ε-uniformly ergodic, where $V_\varepsilon(x) = \exp(\varepsilon V(x))$ for some $\varepsilon > 0$ sufficiently small.

Proof. Note first that V can be taken as radially homogeneous without loss of generality: $V(bx) = bV(x)$ for $b \ge 0$. If this is not the case, we can replace V by $V^1(x) = \inf_{b>0} \frac{1}{b} V(bx)$.

Given the uniform convergence of $\{q^n\}$ and the inequality (2.15) for the limit, we can find an $n_0 > 0$ such that

$$\mathsf{E}[V(Q(n;nx))] = n\mathsf{E}[V(q^n(1;x))] \le n(x - 1/2), \qquad \|x\| \ge b_0,\ n \ge n_0.$$

Hence we can assume that (V1) of [47] is satisfied for the n-step chain:

$$P^n V\,(x) \le V(x) - 1, \qquad \|x\| \ge nb_0.$$

The result then follows from [47, Theorem 16.3.1]. □

A strengthened form of convergence to the fluid limit model also provides a basis for establishing transience of a network model. Note that a large deviations bound would provide a rate of convergence far stronger than assumed in (i).

THEOREM 5. *Suppose that the network is controlled using a stationary policy and that the following hold for the set $S = \mathcal{O}$, where $\mathcal{O}$ is bounded and open as a subset of $\mathbb{R}_+^\ell$.*

(i) *The uniform limit* (U2) *holds, where, for some finite $b(\,\cdot\,)$,*

$$\Theta(\varepsilon, T, n) \le b(\varepsilon, T)/n, \qquad n \ge 1.$$

(ii) *There is an open set $\mathcal{V}$ with $\bar{\mathcal{V}} \subset \mathcal{O}$, and there are an $r > 1$, $T < \infty$, such that*

$$q(T;x) \in r\mathcal{V}, \qquad x \in \mathcal{O},\ q \in \mathcal{L}_x.$$

(iii) *For some $\varepsilon_1 > 0$ we have the uniform lower bound*

$$\|q(t;x)\| \ge \varepsilon_1, \qquad 0 \le t \le T, x \in \mathcal{O},\ q \in \mathcal{L}_x.$$

Then there is a constant b_2 such that for $x \in \mathcal{O}$

$$\mathsf{P}(\boldsymbol{Q} \to \infty \mid Q(0) = nx) \ge 1 - \frac{b_2}{n}.$$

Hence if $\{n\mathcal{O}\} \cap \mathcal{C} \neq \emptyset$ for some $n > b_2$, then the state process $\boldsymbol{Q}$ is a transient Markov chain.

Proof. By (U2) we have, for some $0 < \varepsilon_2 < \varepsilon_1$, some $b_1 < \infty$, and any $x \in \mathcal{O}$,

$$\begin{aligned}
&\mathsf{P}\Big(\|Q(nt;nx)\| \ge \varepsilon_2 n,\ 0 \le t \le T,\ \text{and } Q(nT;nx) \in nr\mathcal{O}\Big) \\
&= \mathsf{P}\Big(\|q^n(t;x)\| \ge \varepsilon_2,\ 0 \le t \le T,\ \text{and } q^n(T;x) \in r\mathcal{O}\Big) \\
&\ge 1 - \frac{b_1}{n}.
\end{aligned}$$

Here we are also using the assumption that $\bar{\mathcal{V}} \subset \mathcal{O}$.

The above bound can be generalized by replacing the integer n with $r^i n$, where $r > 1$ is given in (ii) (again taking integer parts whenever necessary). For any $x \in \mathbb{R}_+^\ell$ and any $i \ge 1, n \ge 1$, define the event $\mathcal{A}(x,i,n) =$

$$\Big\{\|Q(r^i nt;nx)\| \ge \varepsilon_2 r^i n,\ 0 \le t \le T,\ \text{and } Q(r^i nT;nx) \in r^{i+1} n\mathcal{O}\Big\}$$

so that by the previous bound, whenever $x \in r^i\mathcal{O}$,

$$\mathsf{P}(\mathcal{A}(x,i,n)) \ge 1 - \frac{b_1}{n} r^{-i}.$$

By stopping the process at the successive times, $N_0 = 0$, $N_k = N_{k-1} + nr^{k-1}T$, $k \ge 1$, and, using the Markov property, we find that for $x \in \mathcal{O}$ and with $\beta = \varepsilon_2 T^{-1}(1 - r^{-1})$,

$$\begin{aligned}
\mathsf{P}(\|Q(k;nx)\| \ge \beta t,\ 0 \le t \le nN_k) &\ge \sum_{i=0}^{k-1} \inf_{y \in \{r^i\mathcal{O}\}} \mathsf{P}(\mathcal{A}(y,i,n)) \\
&\ge 1 - \frac{b_1}{n}(1 + r^{-1} + \cdots + r^{-k+1}).
\end{aligned}$$

This proves the theorem with $b_2 = \frac{b_1}{1 - r^{-1}}$ since the integer k is arbitrary. ∎

We consider the example illustrated in Figure 10 to show how Theorems 4 and 5 can be applied. This example was introduced in [52, 35] to show how instability can arise in networks even when the traffic conditions are satisfied.

To give one example of a stabilizing policy, suppose that at each time k we choose $U^\circ(k)$ to minimize the conditional mean,

$$U^\circ(k) = \arg\min_a \mathsf{E}[\|Q(k+1)\|^2 \mid Q(k), U(k) = a],$$

where the minimum is over all $a \in \mathsf{U}$, subject to the constraint that $a_i = 0$ if $Q_i(k) = 0$. The minimization can be selected so that $\mathbf{U}^\circ$ is defined by a nonidling, stationary Markov policy defined by a feedback law F°.

The feedback law can be written as

$$F^\circ(x) = \arg\min P_a V_2\,(x), \qquad x \in \mathsf{X},$$

where $V_2(\,\cdot\,) := \|\,\cdot\,\|^2$. The function $F^\circ\colon \mathsf{X} \to \mathsf{U}$ is radially constant and vanishes on the boundaries $F_i^\circ(x) = 0$ when $x_i = 0$. The uniform condition (U2) is readily verified in this case since, for large x, the controlled chain resembles an unreflected random walk (see [6] and also Proposition 8 below).

To evaluate F° note that for any action a, the drift $P_a V_2 - V_2$ is the sum of a linear term $\langle v_a, x\rangle$ and a bounded term. The vector v_a can be expressed as

$$v_a = 2(Ba + \alpha) = 2(\alpha_1 - \mu_1 a_1, \mu_1 a_1 - \mu_2 a_2, \alpha_3 - \mu_3 a_3, \mu_3 a_3 - \mu_4 a_4)^{\scriptscriptstyle T}.$$

The choice $a^{\scriptscriptstyle \rm ss} = (\alpha_1\mu_1^{-1}, \alpha_1\mu_2^{-1}, \alpha_3\mu_3^{-1}, \alpha_3\mu_4^{-1})^{\scriptscriptstyle T}$ makes $v_{a^{\rm ss}} = 0$ and is in the interior of the control space provided that the capacity conditions hold. This gives $P_{a^{\rm ss}} V_2 - V_2 \le b_0$ for some constant b_0. This is a randomized action, which is feasible provided $x_i \neq 0$, $1 \le i \le 4$. If some $x_i = 0$, then the corresponding value $a_i^{\rm ss}$ must also be set to 0, but we still obtain an upper bound of the form $P_{a^{\rm ss}} V_2 - V_2 \le b_0$.

One can conclude that the feedback law

$$F^\varepsilon(x) = \mathbb{I}_+(x)\left(a^{\rm ss} - \varepsilon B^{-1}\frac{x}{\|x\|}\right) \tag{2.16}$$

with B given in (2.12) and $\mathbb{I}_+(x) = \operatorname{diag}(\mathbb{I}(x_1 > 0), \ldots, \mathbb{I}(x_4 > 0))$ is feasible for $\varepsilon > 0$ sufficiently small. For some possibly larger b_0, it satisfies

$$P_{F^\varepsilon} V_2 \le V_2 - 2\varepsilon\|x\| + b_0.$$

By minimality, the feedback law F° exhibits an even larger negative drift,

$$P_{F^\circ} V_2 \le P_{F^\varepsilon} V_2 \le V_2 - 2\varepsilon\|x\| + b_0.$$

Using Jensen's inequality, we find that the function $V(x) = \sqrt{V_2(x)} = \|x\|$ is a Lyapunov function for the network, and it is also a Lyapunov function for the fluid limit model. Applying Theorem 4, we see that F° is a regular policy, and that the controlled network is V_ε-uniformly ergodic.

Suppose that we replace the ℓ_2-norm by the ℓ_1-norm. Letting $c(x) = \sum x_i$, we minimize over all a the conditional mean $P_a c\,(x)$. The resulting policy is the last-buffer-first-served (LBFS) policy (where buffers 2 and 4 have strict priority). This is also one version of the $c\mu$-rule.

This policy is known to lead to a *transient* model for certain parameters, even when (2.13) holds. Specifically, suppose that

$$\frac{\alpha_1}{\mu_2} + \frac{\alpha_2}{\mu_4} > 1.$$

Under the LBFS policy the resulting fluid limit model satisfies, for some T, r,

$$q(T; x_0) = r x_0, \qquad x_0 = (0, 0, 0, 1)^{\scriptscriptstyle T}.$$

This was first shown in [35]. We also have $\|q(t; x_\varepsilon) - q(t; x_0)\| \le \varepsilon$, $0 \le t \le T$, for all $\varepsilon > 0$ sufficiently small and all $x_\varepsilon \in \mathbb{R}_+^\ell$ satisfying $\|x_\varepsilon - x_0\| \le \varepsilon$. Hence the assumptions of Theorem 5 are satisfied with $\mathcal{O} = \{x \in \mathbb{R}_+ : \|x - x_0\| < \varepsilon\}$ and $\mathcal{V} = \{x \in \mathbb{R}_+ : \|x - x_0\| < r^{-1}\varepsilon\}$. Condition (U2) holds with $S = \mathcal{O}$, again using the fact that between emptying times at buffers 2 and 4 the process $\boldsymbol{Q}$ is a simple random walk. We conclude that the network model is transient under the LBFS policy.

3. Optimization.

3.1. The average-cost optimization problem. In this paper we restrict our attention to the average cost problem. For any allocation sequence $\boldsymbol{U}$ and any initial condition x we set

$$J(x,\boldsymbol{U}) = \limsup_{N\to\infty} \frac{1}{N} \sum_{0}^{N-1} \mathsf{E}_x[c(Q(k))].$$

The most common choice is $c(x) = |x|$, where we let $|\cdot|$ denote the ℓ_1-norm. In this case the optimization of J amounts to delay minimization by Little's theorem. We have already seen that the cost $J(x,\boldsymbol{U})$ is finite and independent of x when $\boldsymbol{U}$ is a regular policy.

An optimal policy, if it exists, can be taken to be stationary, where the associated average cost optimality equations are given in (1.1),(1.2). We show below that a solution does exist when the fluid model is stabilizable, and that h^* can be approximated by the value function for a fluid model optimal control problem. For any T and any $x \in \mathbb{R}^\ell_+$, consider the problem of minimizing

$$\int_0^T c(q(t;x))\,dt$$

subject to the constraint that $q\colon [0,T] \to \mathbb{R}^\ell_+$ must satisfy (2.2) for some feasible allocation $\boldsymbol{z}$. The infimum is denoted $V^*(x,T)$.

The following proposition shows that the fluid optimal policy is in some sense *greedy*. That is, the cost as a function of the state $c(q(t))$ is never increasing, and its rate of decrease is maximal when $t \sim 0$. Such behavior is rarely found in dynamic optimization problems. For example, even a second order linear system controlled using optimal linear quadratic regulator (LQR) linear feedback can be expected to exhibit overshoot. An illustration is shown in Figure 3 for the network shown in Figure 12.

The proof of (ii) follows from the aforementioned fact that a state can be reached by following a straight line, provided it is reachable through some control. The result (i) easily follows, and (iii) is well known (see [56, 50]).

PROPOSITION 6. *For any time horizon T,*

(i) *the value function $V^*(\cdot\,,T)$ is convex;*

(ii) *for any $x \in \mathbb{R}^\ell_+$ and any optimal allocation $z^*(\cdot\,;x)$, the function $c(z^*(t;x))$ is decreasing and convex as a function of t;*

(iii) *if the cost c is linear, then $V^*(\cdot\,,T)$ is piecewise quadratic. Moreover, for any $x \in \mathbb{R}^\ell_+$ there exists an optimal state trajectory $q^*(\cdot\,;x)$ and an optimal allocation $z^*(\cdot\,;x)$ which are piecewise linear.*

For any fixed x we evidently have that $V^*(x,T)$ is increasing with T. Moreover, there exists some T_θ such that the optimal trajectories vanish by time T_θ when the time horizon T is at least T_θ, whenever the initial condition satisfies $\|x\| \le 1$. It follows that $V^*(x,T) = V^*(x,T_\theta)$ for any such T and x (see the proof of Theorem 7 (ii) below). Hence for such x and T we have

$$V^*(x,T) = V^*(x) = \min \int_0^\infty c(q(t))\,dt, \tag{3.1}$$

where the minimum is subject to the same constraints on q over the entire positive time axis.

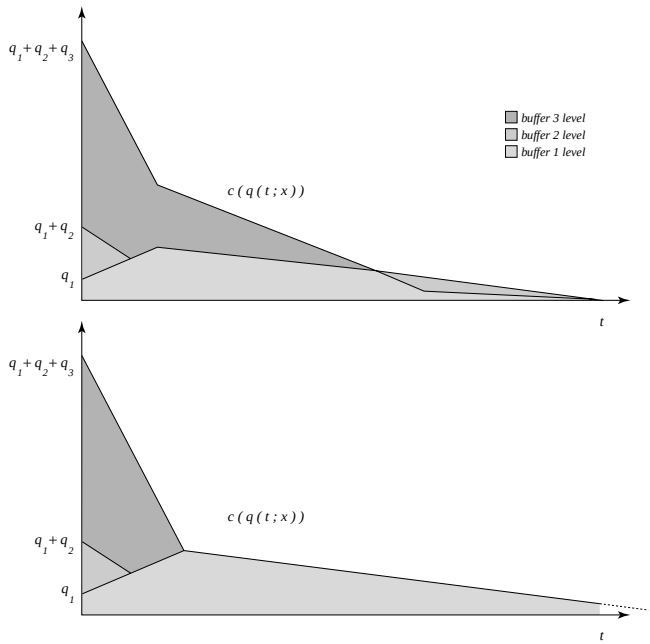


FIG. 3. *The trajectory of buffer levels and evolution of the cost $c(q(t)) = |q(t)|$ for the model of Figure* 12 *for a given set of initial conditions. The first figure illustrates the optimal policy, and the second shows the LBFS priority policy.*

THEOREM 7. *If the fluid model is stabilizable, then for the network* (2.7) *there is a stationary, nonrandomized feedback law F^* with the following properties:*

(i) *It is regular, and hence the average cost $\eta^* = J(x, F^*)$ is finite and independent of x.*

(ii) *The fluid limit model $\mathcal{L}^*$ is stable.*

(iii) *The fluid limit model is optimal with respect to the total cost: With probability one, for any $x \in \mathbb{R}_+^\ell$ and any fluid limit $q^* \in \mathcal{L}_x^*$,*

$$\int_0^\infty c(q^*(t;x))\,dt = V^*(x).$$

(iv) *There exists a solution h^* to the average cost optimality equation* (1.1), (1.2) *which satisfies*

$$\limsup_{\|x\|\to\infty} \left| \frac{h^*(x)}{\|x\|^2} - \frac{V^*(x)}{\|x\|^2} \right| = 0.$$

Proof. Result (i) is a minor generalization of [45, Theorem 5.2]. The existence of a stabilizing policy, as required in this result, is guaranteed by the stabilizability of the fluid model (see Theorem 13 below).

To prove (ii), assume that (iii) holds. We then have for all t, with probability one,

$$V^*(q^*(t;x)) = V^*(x) - \int_0^t c(q^*(s;x))\,ds.$$

From the Lipschitz continuity of the model one can show that V^* is equivalent to a quadratic, in the sense that $V^*(x)/\|x\|^2$ is bounded from above and below for $x \neq \theta$

[45]. It then follows that for some $l_0 < \infty$,

$$\sqrt{V^*(q^*(t;x))} \le \sqrt{V^*(x)} - l_0 t, \qquad t \le \tau_\theta,$$

where τ_θ is the emptying time for $q^*(t;x)$. Thus we have the bound $\tau_\theta \le \sqrt{V^*(x)}/l_0 < \infty$

To prove (iii) we use the following previous results.

(a) It is shown in [45, Theorem 5.2 (i)] that the policy F^* can be chosen so that for any other policy F,

$$\lim_{T\to\infty} \inf \lim_{n\to\infty} \inf \left(\mathsf{E}\left[\int_0^T c(q_F^n(t;x))\,ds\right] - \mathsf{E}\left[\int_0^T c(q_{F^*}^n(s;x))\,ds\right]\right) \ge 0. \tag{3.2}$$

(b) In [42, 2] it is shown that for any T there is a policy F^∞ which attains the optimal cost for the fluid control problem:

$$\lim_{n\to\infty} \mathsf{E}\left[\int_0^T c(q_{F^\infty}^n(s;x))\,ds\right] = V^*(x,T), \qquad \|x\| = 1.$$

Taking $T > T_\theta$, we see that, for any $m \ge 1$,

$$\lim_{n\to\infty} \mathsf{E}\left[\int_0^{mT} c(q_{F^\infty}^n(s;x))\,ds\right] = V^*(x), \qquad \|x\| = m.$$

The point of (b) is that the optimal cost $V^*(x)$ is attainable, and hence the bound given in (a) can be strengthened: For any weak limit $q_{F^*}(\,\cdot\,;x)$ and any T we have by weak convergence

$$\mathsf{E}\left[\int_0^T c(q_{F^*}(s;x))\,ds\right] \le V^*(x), \qquad x \in \mathbb{R}_+^\ell.$$

But for $T \ge T_\theta \|x\|$ we have, with probability one,

$$\int_0^T c(q_{F^*}(s;x))\,ds \ge V^*(x).$$

Combining these two inequalities completes the proof of (iii).

Result (iv) then follows as in the proof of [45, Theorem 5.2 (iii)]. □

From these results we can obtain much insight into the structure of optimal policies for the discrete network when the state is large, i.e., the network is congested. We illustrate this now with several examples.

3.2. Examples. At this stage in the theory there are no general results which are as striking as those that can be found in numerical examples. The general principle appears to be that an optimal policy for the discrete network is equal to the fluid policy, suitably modified along the boundary of the state space. In fact, in several special cases it has been shown that an approximately optimal policy can be constructed in this manner [32, 26, 4], and a general approach is developed in [43].

In computing optimal policies for the examples below, we are forced to truncate the state space to obtain a finite MDP model. Optimization is still difficult due to the large state spaces involved. For example, for a network with four buffers of size

twenty each, the state space contains $m = 160,000$ elements, and computing the optimal policy involves inverting an $m \times m$ matrix.

However, one can use successive approximation to obtain a sequence of approximations $\{h_n : n \geq 0\}$. This is also known as the value iteration algorithm. Theorem 7 (iv) suggests an initialization for the algorithm: $h_0 = V^* \approx h^*$. Numerical results obtained in [9] show that this choice can speed convergence by orders of magnitude. We have used this approach in all of the examples below.

Boundary effects for a truncated model can be severe. For instance, for a loss model, if a buffer is full, then it may be desirable to serve an upstream buffer: the resulting overflow will reduce the steady-state cost. The policies are shown in a truncated region since this behavior has nothing to do with real network dynamics. For example, Figure 5 shows the optimal policy $F^*(x)$ for all x satisfying $\|x\|_\infty < 25$. In this example the value iteration algorithm was used with a network model allowing 39 customers at each buffer. The dimension of the resulting average-cost optimality equations (1.1), (1.2) was 40^3 since there are three buffers in this example.

The $M/M/1$ queue. Recall that the fluid limit model satisfies $q(t;x) = x - \mu z(t) + \alpha t$, $t \geq 0$, where $\alpha + \mu = 1$. Using the notation defined in section 2, we have

$$B = -\mu,\ C = 1, \quad \text{and} \quad \mathbf{1} = 1.$$

The nonidling policy is given by $\zeta(t) = \frac{d}{dt} z(t) = 1$ when $q(t;x) > 0$. It is optimal for any monotone cost function.

For the discrete-stochastic model with cost $c(x) = x$, the relative value function h^* is given by

$$h^*(x) = \frac{1}{2} \frac{x^2 + x}{\mu - \alpha}.$$

The fluid value function is given by

$$\begin{aligned} V^*(x) &= \int_0^\infty q(t;x)\, dt \\ &= \frac{1}{2} \frac{x^2}{\mu - \alpha}. \end{aligned} \tag{3.3}$$

We see that the error in the approximation $h^*(x) \approx V^*(x)$ is linear in this special case.

The "criss-cross" network. Figure 4 shows a model introduced in [27] to illustrate the use of Brownian motion approximation for networks. The system parameters are

$$B = \begin{bmatrix} -\mu_1 & 0 & 0 \\ \mu_1 & -\mu_2 & 0 \\ 0 & 0 & -\mu_3 \end{bmatrix}, \qquad C = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix},$$

and in this model we take $\alpha_2 = 0$. This model has become a standard example.

Optimal policies for the fluid model are easily computed. Take c equal to the ℓ_1-norm, and suppose that $\mu_2, < \mu_3 < \mu_1$. In this case strict priority is given to buffer 3 whenever buffer 2 is nonempty. When this buffer does empty, then the optimal policy sets

$$z_1(t) = \mu_2/\mu_1, \qquad z_3(t) = 1 - z_1(t),$$

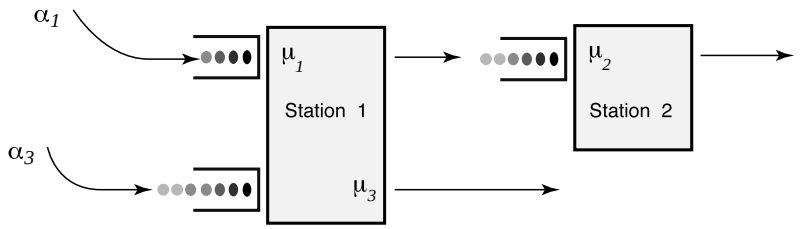


FIG. 4. *A simple two station network with* $\ell_m = 2$ *and* $\ell = \ell_u = 3$.

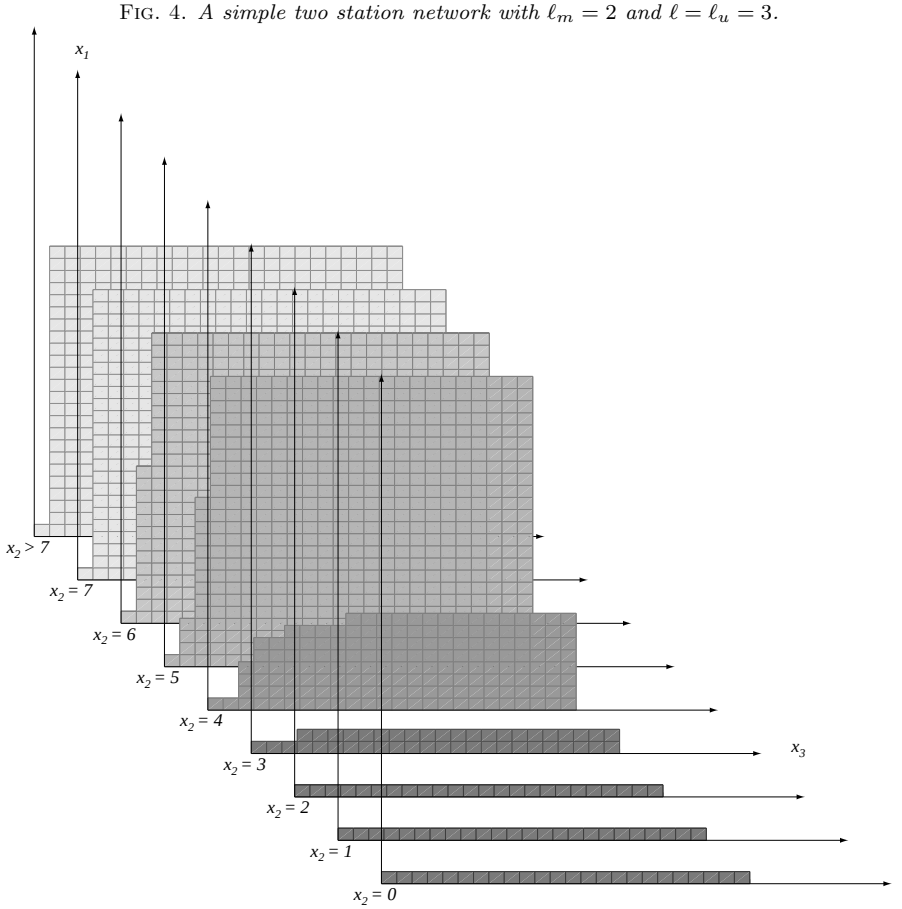


FIG. 5. *The optimal policy for the network above with* $\alpha^T = (9, 0, 9)$ *and* $\mu^T = (25, 10, 20)$. *The grey areas indicate states at which buffer* 3 *is given strict priority.*

provided both buffers 1 and 3 are nonempty. This is a pathwise optimal policy in the sense that it minimizes $c(q(t; x))$, for each $t \geq 0$, over all policies.

The optimal policy for the discrete model with particular parameter values satisfying these constraints is given in Figure 5. As always, the fluid limit of this optimal policy is the optimal policy for the fluid model. The discrete optimal policy is similar to the optimal fluid policy: The critical value $q_2(t) = 0$ has been shifted upward to $Q_2(t) \approx 4$.

A routing model. We now show how the theory applies to a routing problem. The model illustrated in Figure 6 has been considered in several papers; see, in particular,

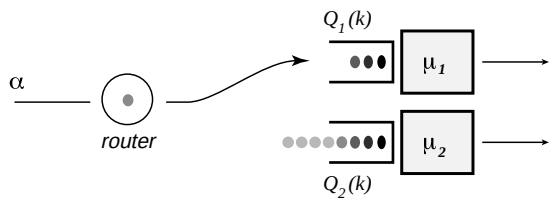


FIG. 6. *A network with controlled routing:* $\ell_m = 3$, $\ell = 2$, *and* $\ell_u = 4$.

[23, 32]. Customers that arrive to the system are routed to one of the two servers. In this example, $\ell = 2$, $\ell_u = 4$, and

$$B = \begin{bmatrix} \alpha & 0 & -\mu_1 & 0 \\ 0 & \alpha & 0 & -\mu_2 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

The router is nonidling in this model; $\zeta_1 + \zeta_2 = 1$. This requirement can be expressed as the additional linear constraint, $C_a\zeta \le b_a$, where $C_a = [-1, -1, 0, 0]$ and $b_a = -1$. Alternatively, one can enlarge the state space to include a buffer at the router but impose the linear constraint that $Q_3(t) \equiv 0$.

Note that in the routing model the arrival stream is absorbed into the random matrix $\widetilde{B}$. Hence in this model we take the two dimensional vector α to be zero. Assume that $\mu_1 = \mu_2 = \mu$ and that $\mu < \alpha < 2\mu$. The fluid model is then stabilizable.

To minimize the total cost for the fluid model

$$\int_0^\infty c(q(t;x))\,dt,$$

one obviously takes z_3 and z_4 to be nonidling. Consider the case where c is linear with $c(x) = (c_1, c_2) \cdot x$ and $c_1 > c_2$. Then the priority policy is optimal, where fluid is routed to buffer 2 as long as buffer 1 is nonempty. As soon as it does empty, then fluid is routed to buffer 1 at rate $\zeta_1(t) = \frac{d}{dt} z_1(t) = \mu_1/\alpha$ so that buffer 1 is nonidling, but empty. The remaining fluid is sent to buffer 2 so that $\zeta_2(t) = 1 - \mu_1/\alpha$. This policy is again pathwise optimal, and it enforces nonidleness so that $\zeta_3(t) = \zeta_4(t) = 1$ for $t < \tau_\theta$.

The discrete-stochastic model is considered in [23] for a general linear cost function. It is shown that an optimal policy exists, and that it is of a nonlinear threshold form: There is a nondecreasing function $\gamma\colon \mathbb{Z}_+ \to \mathbb{Z}_+$ such that when a job arrives, when the queue lengths are x_1 and x_2, then buffer 1 receives the job if and only if $\gamma(x_1) \ge x_2$. The analysis of [58] implies that the function γ is unbounded, but in general no analytic formula is available for the computation of γ.

We see in Figure 7 that the optimal policy for the discrete network is closely approximated by the optimal fluid policy, modified along the boundary. The "thickened boundary" ensures that, with high probability, neither buffer will idle when the network is congested.

A processor-sharing model. Another simple example where the optimal allocation for the fluid model is explicitly computable is the processor-sharing network considered in [4, 26] and illustrated in Figure 8. In this example, $\ell = 2$, $\ell_u = 3$, and the system parameters are

$$B = \begin{bmatrix} -\mu_A & -\mu_B & 0 \\ 0 & 0 & -\mu_C \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}, \quad \alpha = \begin{bmatrix} \alpha_A \\ \alpha_B \end{bmatrix}.$$

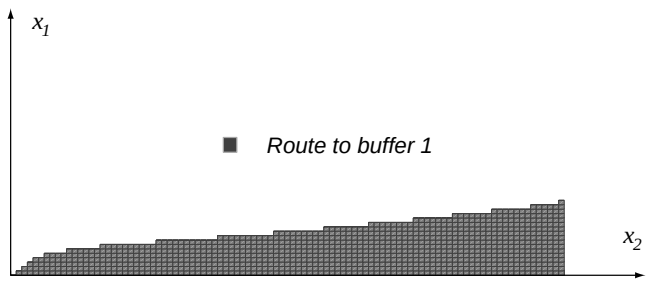


FIG. 7. *Optimal discrete policy for the simple routing model with* $\alpha = 9$ *and* $\mu^T = (5, 5)$. *The one step cost is* $c(x) = 3x_1 + 2x_2$.

As in the previous example, for any cost c we can assume that the optimal policy is nonidling at station 1. The fluid limit model illustrated on the right in Figure 8 is based on this assumption.

We assume that $\alpha_A > \mu_A$. In this case, it is critical that station 1 receive outside assistance. Under this condition and the nonidling assumption at station 1, we arrive at a reduced order model with $\ell = \ell_u = 2$ and

$$B = \begin{bmatrix} -\mu_B & 0 \\ 0 & -\mu_C \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 1 \end{bmatrix}, \quad \alpha = \begin{bmatrix} \alpha_A - \mu_A \\ \alpha_B \end{bmatrix}.$$

For any linear cost the optimal allocation is the $c\mu$-rule priority policy, which is again pathwise optimal in this example. It is shown in [4] that a modification of this policy is nearly optimal in heavy traffic. Figure 9 shows the optimal policy for the discrete model. It is similar to the $c\mu$ priority policy, with priority given to processor B at station 2. However, the boundary $\{x_2 = 0\}$ has been shifted to form the concave region shown in the figure.

A generalized $c\mu$*-rule in scheduling.* We return now to the example illustrated in Figure 10, whose fluid model is defined by the parameters

$$B = \begin{bmatrix} -\mu_1 & 0 & 0 & 0 \\ \mu_1 & -\mu_2 & 0 & 0 \\ 0 & 0 & -\mu_3 & 0 \\ 0 & 0 & \mu_3 & -\mu_4 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix},$$

with $\alpha_2 = \alpha_4 = 0$. Consider for simplicity the symmetric case where $\mu_1 = \mu_3$ and $\mu_2 = \mu_4$. We also assume that $\mu_1 = 2\mu_2$ so that the exit buffers are slow.

It is pointed out in [42] that the optimal policy for the fluid model when c is the ℓ_1-norm is given as follows: The exit buffers have strict priority when $q_2(t) > 0$ and $q_4(t) > 0$. As soon as one of these buffers empties, say, buffer 2, then one sets $\zeta_1(t) = \mu_2/\mu_1$ and $\zeta_4(t) = 1 - \mu_1$ and continues to set $\zeta_2(t) = 1$. This policy maximizes the overall draining rate at each time t, it is pathwise optimal, and it achieves the total cost V^* for the fluid model. Recall that the analogous greedy policy, defined through the discrete model, is destabilizing!

We have computed an optimal policy for the discrete network numerically for this special case. However, noting that the optimal fluid policy is independent of the arrival rates $\alpha^T = (\alpha_1, 0, \alpha_3, 0)$, we have taken an extreme case with $\alpha = 0$ and have considered the total cost problem,

$$V(x) = \min \sum_0^\infty \mathsf{E}_x[c(Q(k))],$$

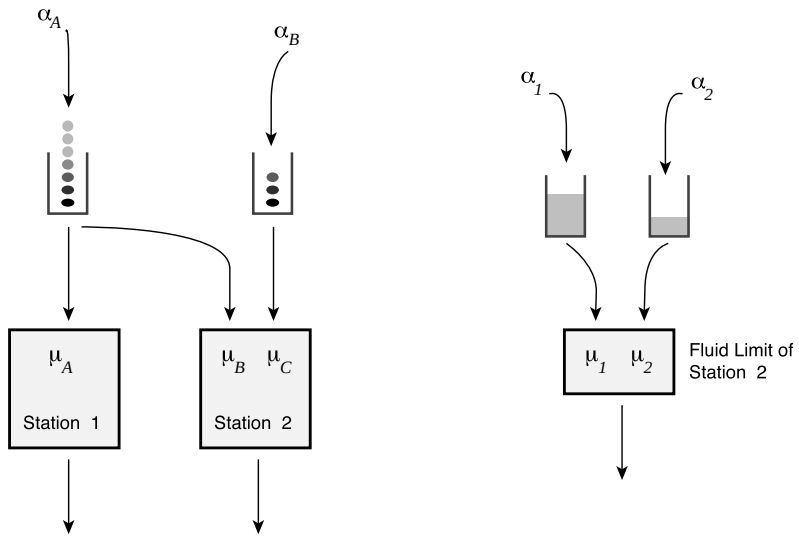


FIG. 8. *On the left is the processor-sharing network of* [27]. *On the right is its fluid limit model.*

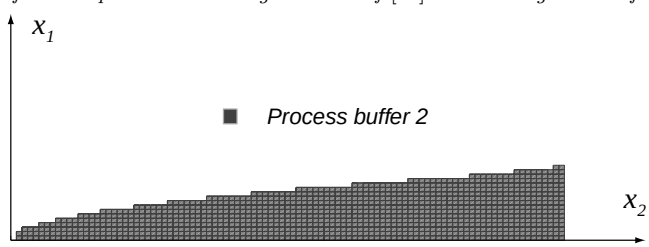


FIG. 9. *Optimal policy for the processor sharing model with ℓ_1 cost, $\alpha = (1, 1)$, and $\mu = (1, 3, 2)$. This closely resembles the optimal fluid policy which gives strict priority to the first server at station two since $\mu_B > \mu_C$.*

where the minimum is with respect to all policies. This gives rise to a finite dimensional optimization problem which can be solved exactly for each x.

The results shown in Figure 11 indicate that the optimal discrete policy is again similar to the optimal fluid policy.

A scheduling-model with no pathwise optimal solution. Consider the network given in Figure 12 with c taken to be the ℓ_1-norm. One policy that minimizes the total cost for the fluid model is defined as follows, where γ is a positive constant defined by the parameters of the network.

(i) Serve $q_3(t)$ exclusively ($\zeta_3(t) = 1$) whenever $q_2(t) > 0$ and $q_3(t) > 0$.
(ii) Serve $q_3(t)$ exclusively whenever $q_2(t) = 0$ and $q_3(t)/q_1(t) > \gamma$.
(iii) Give $q_1(t)$ partial service with $\zeta_1(t) = \mu_2/\mu_1$ whenever $q_2(t) = 0$, and

$$0 < q_3(t)/q_1(t) \le \gamma.$$

This model is most interesting when station 2 is the bottleneck, since one must then make a tradeoff between draining the system and avoiding starvation at the bottleneck. Taking $\rho_2 = \alpha/\mu_2 = 9/10$ and $\rho_1 = \alpha/\mu_1 + \alpha/\mu_3 = 9/11$, the constant γ is equal to one, and hence the optimal policy is of the form illustrated in Figure 13.

Optimal policies are computed numerically in [45] for versions of this model with truncated buffers. Results from one experiment are shown in Figure 14. As in the

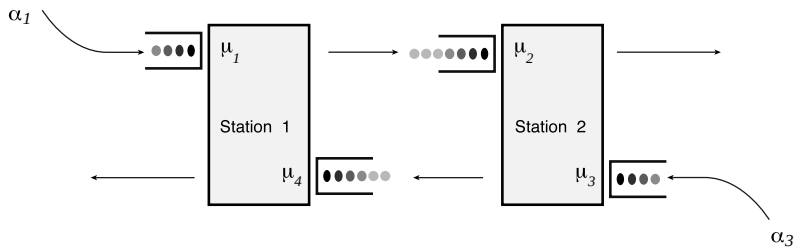


FIG. 10. *A multiclass network with* $\ell_m = 2$ *and* $\ell = \ell_u = 4$.

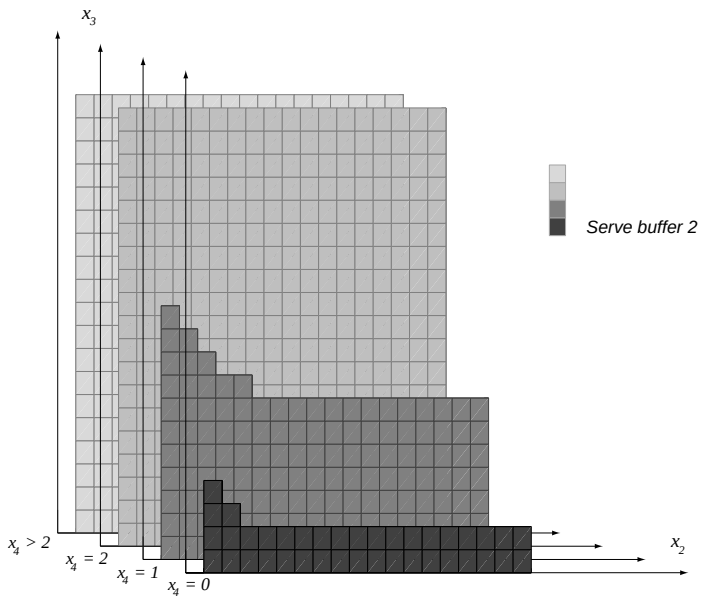


FIG. 11. *Optimal policy for the four-buffer scheduling model shown in Figure* 10 *under the total cost criterion with c equal to the* ℓ_1*-norm. The arrival streams are null, and* $\mu = (2, 1, 2, 1)$. *The figure shows the policy when* $x_1 = 3$, $x_4 = 0, 1, 2, 3$, *with* x_2 *and* x_3 *arbitrary. This optimal policy is of the same form as the fluid policy: It gives strict priority to the exit buffer at station* 2, *unless buffer* 4 *is starved of work, in which case buffer* 3 *releases parts to feed buffer* 4.

previous examples we see that the discrete optimal policy is easily interpreted. It regulates the work waiting at buffer 2, and does so in such a way that buffer 2 is rarely starved of work when the network is congested.

The policy shown in Figure 14 is also very similar to the fluid policy. Performing some curve fitting, we can approximate this discrete policy as follows: serve buffer one at time t if and only if either buffer three is equal to zero or

$$Q_1(k) - \bar{x}_1 > Q_3(k) - \bar{x}_3 \qquad \text{and} \qquad Q_2(k) \le \bar{x}_2, \tag{3.4}$$

where the translation $\bar{x}$ positive. The most accurate approximation is obtained when $\bar{x}$ depends upon the current state $Q(k) = x$, say,

$$\bar{x} = \bar{x}_0 \log(\|x\| + 1), \qquad x \in \mathsf{X}, \tag{3.5}$$

with $\bar{x}_0 > 0$ and constant. Moreover, with this choice, the fluid limit obtained using the policy (3.4) is precisely the optimal policy minimizing the total fluid cost, which is illustrated in Figure 13.

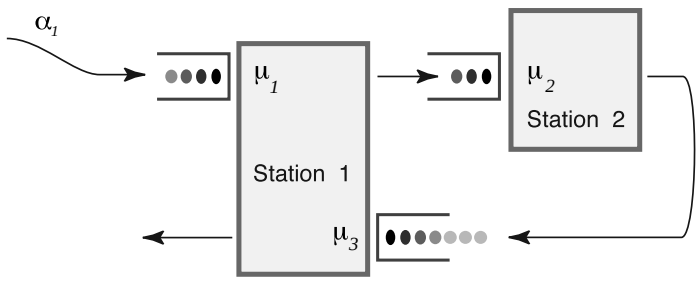


FIG. 12. *A multiclass network with* $\ell_m = 2$ *and* $\ell = \ell_u = 3$.

4. Feedback regulation. The results and examples of the previous section all suggest that the fluid model should play a useful role in control synthesis for network models. Theorem 7 establishes a connection between two optimization problems: one is deterministic and relatively transparent, and the other is stochastic, discrete, and apparently hopeless.

Even if one can find a feedback law $\frac{d}{dt}z(t) = \zeta(t) = f^*(q(t))$ which is optimal for the fluid model, it is not obvious how to use this information. A direct translation such as $F(x) = f^*(x)$ is not appropriate. It is shown in [45] that this policy may have a fluid limit model which differs grossly from the desirable optimal fluid process. However, the numerical results given above all show that, at least for simple models, an optimal policy for a discrete network is approximated by an affine shift of the form

$$F(x) = f^*((x - \bar{x})^+), \qquad x \in \mathsf{X}. \tag{4.1}$$

Moreover, one can show that a properly defined shift of this form ensures that the resulting fluid limit model for the network controlled using F approximates the optimized fluid model (see [42, 2, 43] and Proposition 8 below).

One might arrive at the policy (4.1) without any consideration of optimization. When the feedback law f^* is chosen appropriately, this policy will attempt to regulate the state $Q(k)$ about the value $\bar{x}$. If this regulation is accomplished successfully, then

- provided $\bar{x}$ is not too big, the cost $c(Q(k))$ will not be too large;
- if the target $\bar{x}$ is not too small then this policy will avoid starvation of any resource; and
- regulation to a constant should provide reduced variance at each station, as has been argued for the class of fluctuation smoothing policies [37].

4.1. Discrete review structure. In this section we adapt the approach of [28] to define policies for the physical network based on an idealized allocation derived from the fluid model. Related approaches are described in [22].

In practice one will rarely use a stationary policy F since one is forced to make scheduling decisions at every discrete sampling instance. This is undesirable since it results in high computational overhead and, more importantly, excessive switch-overs. The proposed policies consist of three components:

(a) For each initial condition x, a well-designed fluid trajectory $q(t; x)$ satisfying (2.2) for some allocation process $\boldsymbol{U}$. This will typically be defined through a feedback law f so that

$$q(t; x) = x + \alpha t + B \int_0^t f(q(s; x))\, ds, \qquad t \in \mathbb{R}_+,\ x \in \mathbb{R}_+^\ell.$$

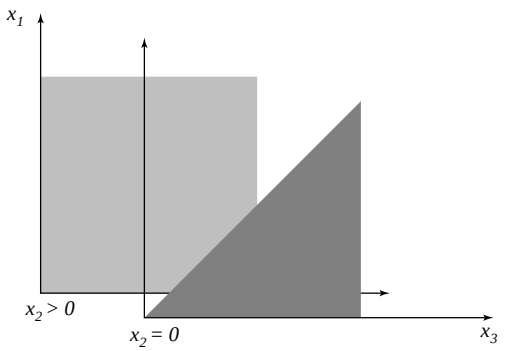


FIG. 13. *The optimal fluid policy for the three buffer re-entrant line with* $\rho_2 = 9/10$ *and* $\rho_1 = 9/11$. *In this illustration, the grey regions indicate those states for which buffer three is given exclusive service.*

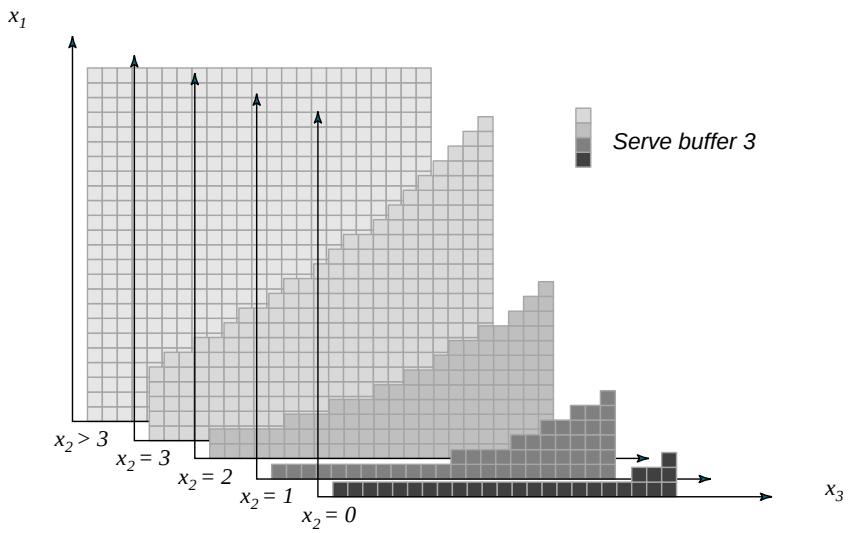


FIG. 14. *Optimal discrete policy for three buffer re-entrant line in the balanced case:* $\alpha/\mu_1 = \alpha/\mu_3 = \frac{1}{2}\rho_1$.

(b) A target vector $\bar{x}$;

(c) A time horizon N over which the policy is fixed.

The target $\bar{x}$ may be a "moving target," in which case it is assumed to be a function of the state. In general, we may also take N as a function of the state, but we will always assume that N and $\bar{x}$ are "roughly constant" for large x.

Given these, we set $N_0 = 0$, and, given $Q(N_0) = x$, we determine the time horizon $N_1 = N(x)$. The values $U(k)$, $N_0 \le t < N_1$ are chosen so that

$$\mathsf{E}[Q(N_1 - N_0; x)] - x \approx \delta(N_1 - N_0; (x - \bar{x})^+), \tag{4.2}$$

where $\delta(T; y) = q(T; y) - y$ for any $y \in \mathbb{R}_+^\ell$, $T \ge 0$.

This final choice is far from unique but will be dictated by considerations such as minimizing switch-over times and avoiding starvation at any station. Once one arrives at time N_1, the choice of $U(k)$ on the time interval $[N_1, N_2)$ proceeds exactly as before, where $N_2 = N_1 + N(Q(N_1))$. Successive review times $\{N_i : i \ge 0\}$ and

actions $\{U(k) : k \geq 0\}$ can then be found by induction.

We shall call any policy of this form a *feedback regulation policy* since it is similar to a state feedback approach to regulation as covered in a second-year control systems course. Below we list some of the issues in design:

- The most basic question is the design of the fluid state trajectories $\{q(\,\cdot\,;x) : x \in \mathbb{R}_+^\ell\}$. A first requirement is stability, and the theory suggests that good performance for the fluid model with respect to the total cost is highly desirable. These design issues will be discussed in depth in section 4.2.
- How do we choose $\bar{x}$?
- How do we choose the time horizon N? This will be dictated by such issues as batch or packet sizes and switch-over costs.
- How do we choose a sequence of actions on $[N_k, N_{k+1})$ so that (4.2) holds? Again, one must consider switch-over costs—two approaches are described below.
- In many models one must also consider idleness avoidance. For routing models this can be treated as in [32] by enforcing routing to a buffer whenever its level falls below a threshold. Scheduling models can be treated similarly.
- Machine failures and maintenance: How should the policy change during a failure? One can again address this problem by considering a fluid model, but one should consider the *delayed* fluid model which includes the residual-life of each service process.

Much has been written on the choice of safety-stock levels. In some simple examples a constant threshold is optimal (see, e.g., [18, 39]). A value of zero is optimal in the single-machine scheduling problem with linear cost since the $c\mu$-rule is optimal for both the fluid and stochastic models.

A general approach is suggested by a rich literature on networks in heavy traffic. In [32] and many other references one considers the case where the system load ρ is close to unity, and

$$(1-\rho)\sqrt{n} \to L, \qquad n \to \infty,$$

where n is a parameter which is sent to ∞ for the purpose of analysis, $\rho = \rho(n)$, and L is a nonzero, finite number. The steady-state number of customers in the system is typically of order $(1-\rho)^{-1}$. (Consider a $G/G/1$ queue or the functional bounds obtained in [31].) Hence the assumptions commonly used in the literature imply that

$$\sqrt{n} = O\left(\frac{1}{1-\rho}\right) = O(\mathsf{E}_\pi[\|Q(k)\|]),$$

where $\mathsf{E}_\pi[\|Q(k)\|]$ denotes the steady-state mean. The thresholds $\bar{x}$ determined in [4, 32] are of order $\log(n)$, so that $\|\bar{x}\| = O\big(\log(\mathsf{E}_\pi[\|Q(k)\|])\big)$. By replacing the steady-state mean with the current value $Q(k) = x$, we arrive at $\|\bar{x}(x)\| = O(\log(\|x\|))$.

However, the results of [43] show that such a small offset may be overly optimistic in general. We are currently exploring parameterizations of the form (3.5), where $\bar{x}_0$ can be tuned, perhaps on-line.

Two major issues are stability and performance of the network under a feedback regulation policy. A stability proof and some approaches to estimating performance through simulation are given in section 4.4. Stability requires some assumptions.

(A1) There exists a function $V\colon \mathbb{R}_+^\ell \to \mathbb{R}_+$ which is Lipschitz continuous and satisfies

$$V(q(t;x)) - V(x) \leq -t, \qquad x \in \mathbb{R}_+^\ell, \qquad t \leq \tau_\theta.$$

(A2) There exist an i.i.d. process $\{\Gamma(k) : k \geq 0\}$ evolving $\mathbb{R}^\ell$ and a family of functions $\{\boldsymbol{F}(\cdot, k, N) : 0 \leq k < N < \infty\}$ such that for each $k < N$,

$$\boldsymbol{F}(\cdot, k, N) \colon \mathsf{X}^{k+1} \times \mathbb{R}^{\ell(k+1)} \to \{0,1\}^{\ell_u},$$

and when $N(Q(0)) = N$,

$$U(k) = \boldsymbol{F}(Q(0), \ldots, Q(k), \Gamma(0), \ldots, \Gamma(k), k, N).$$

(A3) The convergence to the fluid model is uniform in the sense of (U1): For any fixed N, there are a $\Theta(N) > 0$ and a $b_0(N) < \infty$ such that when $U(k) = \boldsymbol{F}(Q(0), \ldots, Q(k), \Gamma(0), \ldots, \Gamma(k), k, N)$ for $k < N$,

$$\frac{1}{N}\mathsf{E}_x\Big[\|Q(N;x) - q(N, (x-\bar{x})^+)\|\Big] \leq \Theta(N), \qquad \|x\| \geq b_0(N),$$

where $\Theta(N) \to 0$ as $N \to \infty$.

Assumption (A1) ensures that the fluid process $q(t;x)$ is stable. The Lipschitz assumption is an important aspect of these policies since it is what allows us to establish robustness with respect to perturbations in system parameters such as arrival and service rates.

The proof of Theorem 13 below is based on a Markovian description of the controlled network, which is possible by assumption (A2). Under this assumption, we can define the Markov state process,

$$\Phi(k)^{\scriptscriptstyle T} = [Q(k), \ldots, Q(n(k))], \qquad t \in \mathbb{Z}_+,$$

where $n(k)$ is the last switch-over time: $n(k) = \min(s \leq t : s = N_k \text{ for some } k)$.

Assumption (A3) requires that one faithfully follow the fluid model. Consider for simplicity the network scheduling problem where $\ell = \ell_u$. Perhaps the most natural approach is to define a processing plan on $[N(k), N(k+1))$ in a generalized round-robin fashion: Take $k = 0$, without loss of generality, and, given $Q(0) = x$, set

$$y = q(T; (x-\bar{x})^+) = (x-\bar{x})^+ + Bz(T; (x-\bar{x})^+) + T\alpha.$$

The vector $a = z(T; (x-\bar{x})^+)/T$ satisfies $a \geq 0$, $Ca \leq \mathbf{1}$, and since $y = (x-\bar{x})^+ + T(Ba+\alpha)$, we obviously have

$$(Ba+\alpha)_i \geq 0 \text{ whenever } x_i \leq \bar{x}_i. \tag{4.3}$$

At machine s suppose we have ℓ_s buffers $i_1, \ldots, i_{\ell_s}$. Given this value of a and given a constant $m > 0$, we perform $a_{i_1} m$ consecutive services at buffer i_1 and then $a_{i_2} m$ consecutive services at buffer i_2, continuing until buffer i_{ℓ_s} has received $a_{i_{\ell_s}} m$ services and then returning to buffer i_1. This cycle repeats itself until time $N(1)$. We again must take the integer parts of $a_k m$, and this will lead to some error. This is insignificant for large N.

An approach based on randomization is particularly straightforward. Suppose that the function $\boldsymbol{F}$ is formed in a stationary, randomized fashion as follows:

$$\begin{aligned} &\mathsf{P}(U_i(k) = 1 \mid Q(0), \ldots, Q(k), \Gamma(0), \ldots, \Gamma(k-1)) \\ &= \mathsf{P}(U_i(t) = 1 \mid Q(k)) \\ &= a_i \mathbb{I}(Q_i(k) > 0). \end{aligned} \tag{4.4}$$

This construction of $U(k)$ for $t < N$ can be equivalently expressed through a feedback function of the form required in (A2), where, for some fixed function $F\colon \mathsf{X}\times\mathbb{R}^{\ell} \to \mathbb{R}^{\ell}_{+}$,

$$\boldsymbol{F}(Q(0),\dots,Q(k),\Gamma(0),\dots,\Gamma(k),t,N) = F(Q(k),\Gamma(k)).$$

PROPOSITION 8. *Consider the network scheduling problem. The randomized policy given in* (4.4) *defines a function* $\boldsymbol{F}$ *satisfying* (A2) *and* (A3).

Proof. We have already seen that (A2) holds.

For any $a \in \mathbb{R}^{\ell}_{+}$ we let $\boldsymbol{Q}^a = \{Q(k;x,a) : k \ge 0\}$ denote the state process for a Jackson network with arrival and service rates given, respectively, by $(\alpha_i, a_i\mu_i)$, $1 \le i \le \ell$. We always assume that $a \in \mathsf{U}$ ($a \in \mathbb{R}^{\ell}_{+}$ with $Ca \le \mathbf{1}$). We can construct all of the processes $\{Q(\,\cdot\,;x,a) : x \in \mathsf{X}, a \in \mathsf{U}\}$ on the same probability space as follows: We are given two mutually independent ℓ-dimensional Poisson processes $M(t), N(t)$ of rate one. The length of the service times, either real or virtual, at buffer i are defined as the interjump times of $N_i(\mu_i a_i t)$; the exogenous arrivals to buffer i occur at the jump times of $M_i(\alpha_i t)$, $t \in \mathbb{R}_{+}$. The kth component of $\boldsymbol{U}^a$ is given by $U_k(k;x,a) = \mathbb{I}(Q_k(k;x,a) > 0)$.

For each n, we let $\{q^n(t;x,a), z^n(t;x,a)\}$ denote the nth scaled queue length and cummulative allocation process. We then set

$$\mathcal{L}_n = \overline{\overline{\bigcup_{k=n}^{\infty} \Big\{(q^k(\,\cdot\,;x,a), z^k(\,\cdot\,;x,a)) : \|x\| = b_0; a \in \mathsf{U}\Big\}}}.$$

The double bar indicates *strong closure* in the function space $C([0,T],\mathbb{R}^{\ell+\ell_v})$, in the uniform norm. The set $\mathcal{L}_n \subset C([0,T],\mathbb{R}^{\ell+\ell_v})$ is compact for any n, and so is its intersection over all n: $\mathcal{L} = \bigcap_n \mathcal{L}_n$.

The set $\mathcal{L}$ is defined for almost every (a.e.) sample path of (M,N). If $(q,u) \in \mathcal{L}$, then there exist $x^i \to x$, $a^i \to a$, and a subsequence $\{n^i\}$ of $\mathbb{Z}_{+}$ such that

$$q^{n^i}(\,\cdot\,;x^i,a^i) \Longrightarrow q, \quad z^{n^i}(\,\cdot\,;x^i,a^i) \Longrightarrow z, \qquad i \to \infty,$$

where the convergence is in $C([0,T],\mathbb{R}^{\ell+\ell_v})$. We then have $q(0) = x$, and for any time t at which q and $\boldsymbol{U}$ are differentiable,

$$\begin{aligned} &\frac{d}{dt}q_i(t) = 0 \qquad \text{if } q_i(t) = 0,\\ &\frac{d}{dt}z_i(t) = 1 \qquad \text{if } q_i(t) > 0, \qquad 1 \le i \le \ell. \end{aligned}$$

This is enough to completely determine the limit set: For any (x,a) there is a unique $q(\,\cdot\,;x,a) \in \mathcal{L}$.

It follows that we have uniformity in the sense of (U2): For any $\varepsilon > 0$,

$$\sup_{\substack{\|x\|=b_0\\ a\in\mathsf{U}}} \mathsf{P}\left(\sup_{0\le t\le T} \|z^n(t;x,a) - z(t;x,a)\| > \varepsilon\right) \to 0, \qquad n \to \infty, \tag{4.5}$$

and the analogous limit holds for $\{q^n\}$.

For $x \in \mathbb{R}^{\ell}_{+}$, $T > 0$, we denote

$$\mathcal{A}(x,T) = \{a \in \mathbb{R}^{\ell}_{+} : Ca \le \mathbf{1}, \quad x + T(Ba + \alpha) \in \mathbb{R}^{\ell}_{+}\}.$$

For the randomized policies considered in Proposition 8 we always have $a \in \mathcal{A}(x,T)$ (see the discussion surrounding (4.3)). Note also that for such a we have $z_i(t) = t$ for any i, and any $0 \le t \le T$. Hence, from (4.5),

$$\sup_{\substack{\|x\|=b_0 \\ a\in\mathcal{A}(x,T)}} \mathsf{P}\left(\sup_{0\le t\le T} \|z_i^n(t;x,a) - t\| > \varepsilon\right) \to 0, \qquad n \to \infty.$$

It is well known that Jackson networks are *monotone* in the sense that if $y \ge x$, then $U_i(k;y,a) \ge U_i(k;x,a)$ for any i,t, and a (see [53]). Hence the above bound can be improved:

$$\sup_{\substack{\|x\|\ge b_0 \\ a\in\mathcal{A}(x,T)}} \mathsf{P}\left(\sup_{0\le t\le T} \|z_i^n(t;x,a) - t\| > \varepsilon\right) \to 0, \qquad n \to \infty, \tag{4.6}$$

and this easily implies that (A3) holds. □

4.2. Design of the fluid trajectory. There are many control strategies for a fluid model which have desirable stability characteristics. Here we describe four approaches which always lead to a stabilizing solution: In each case we construct a Lipschitz continuous Lyapunov function. We have already seen in Theorem 4 that this can imply a strong form of stochastic stability for the network. These results will be generalized to feedback regulation policies in Theorem 13.

The first three classes of policies considered below are based on optimal control: the optimal fluid policies, time-optimal fluid policies, and constrained complexity optimal fluid policies. The latter have fixed complexity which can be chosen in advance by the user. The fourth class that we consider consists of greedy policies. Such policies can be computed easily for large networks by solving an ℓ-dimensional linear program.

Optimal fluid policies. The policy f^* which optimizes the fluid model under the total cost criterion is a natural candidate for application in a feedback regulation policy. The computation of f^* may be posed as an infinite dimensional linear program. Because of the specific structure of the linear program, it is frequently feasible to compute f^* numerically, even though such problems are, in general, intractable (see [41, 49]).

Time-optimal allocations can be computed with only trivial calculation: Proposition 1 implies that a linear time-optimal policy can be constructed for any stabilizable network, which is the basis of the main result of [17]. Time-optimality is used as a *constraint* in the construction of the policies described in [43].

Optimal fluid policies are stabilizing for the fluid model. Moreover, they satisfy assumption (A1) and are hence guaranteed to be stabilizing for the stochastic model when used in a feedback regulation policy.

PROPOSITION 9. *Suppose that the fluid model is stabilizable.*

(i) *Suppose that $q(\,\cdot\,;x)$ is optimal with respect to the total cost* (3.1) *for each x. Then there exists a Lipschitz Lyapunov function so that* (A1) *holds.*

(ii) *Suppose that $q(\,\cdot\,;x)$ is time-optimal in the sense that $\tau_\theta(x)$ is minimized over all fluid policies. Then $V(x) = \tau_\theta(x)$, $x \in \mathbb{R}_+^\ell$, is a Lipschitz Lyapunov function so that* (A1) *holds.*

Proof. Let $V = \beta\sqrt{V^*}$, where $\beta > 0$. From Proposition 6 we can conclude that V is radially homogeneous, and each sublevel set $S_\eta = \{x : V(x) \le \eta\}$ is a convex subset of $\mathbb{R}_+^\ell$ for any $\eta > 0$. It follows that V itself is convex and continuous, which implies Lipschitz continuity. The negative drift required in (A1) holds for β sufficiently large, which establishes (i).

Another Lyapunov function is $V = \beta c$. This satisfies the required drift for sufficiently large β since, as we have already observed, $c(q(t))$ is a convex, decreasing function of t under an optimal policy.

The proof of (ii) is identical since the function $V(x) = \tau_\theta(x)$ is radially homogeneous and convex. □

Although Proposition 9 shows that optimal fluid policies are stabilizing, these policies can be highly complex, even when they are computable (see [41, 49]). We turn next to a simpler class of policies.

The constrained-complexity optimal fluid policy. The difficulty with using an optimal state trajectory $q^*(\,\cdot\,;\,\cdot\,)$ is that complexity, as measured by the number of discontinuities in $\frac{d}{dt}q^*(t;x)$, i.e., the number of switches in the control $z^*(t;x)$, can grow exponentially with ℓ.

To bound complexity, suppose that we take a number κ and demand that q be piecewise linear, with at most κ pieces, so that the control can change no more than κ times. Any $q(\,\cdot\,;\,\cdot\,)$ which is optimal with respect to the total cost (3.1) subject to this constraint will be called a κ-constrained optimal fluid process.

Any such policy can be computed by solving a $\kappa\cdot(\ell+1)$-dimensional quadratic program when the cost is linear [41]. The variables can be taken as the switch-over times $\{0 = T_0, T_1, \ldots, T_\kappa\}$ and the control increments $\{z(T_{i+1}) - z(T_i) : 0 \le i < \kappa\}$.

PROPOSITION 10. *Suppose that the fluid model is stabilizable. Then any κ-constrained optimal fluid process possesses a Lipschitz continuous Lyapunov function so that* (A1) *holds.*

Proof. One can again show that $c(q(t))$ is a convex, decreasing function of t for any κ-constrained optimal fluid process. Hence one can take $V = \beta c$ for $\beta > 0$ sufficiently large. □

The greedy fluid policy. The greedy policy determines the allocation rate $\zeta(t)$ that minimizes $\frac{d}{dt}c(q(t))$ at each t. One motivation for this class of policies comes from considering the dynamic programing equations for the infinite-horizon optimal control problem. The optimal policy is the solution to (4.7) with c replaced by the value function V^*. Greedy heuristics are the most popular in queueing theory. The papers [8, 28] consider greedy policies for state-based cost functions as developed here. The shortest expected delay policy [32] and the least slack policy [37] are based on greedy heuristics for delay minimization.

Suppose that the cost function c is continuously differentiable (C^1). The greedy feedback law $f(x)$ is computed by solving the following ℓ-dimensional linear program: For any x, let ∇c denote the gradient of c evaluated at x, and solve

$$
\begin{aligned}
&\min \langle \nabla c, B\zeta \rangle \\
&\text{subject to} \qquad (B\zeta + \alpha)_i \;\ge\; 0 \quad \text{for all } i \text{ such that } x_i = 0, \\
&\qquad\qquad\qquad\quad \zeta \;\in\; \mathsf{U}.
\end{aligned} \tag{4.7}
$$

Then $f(x)$ is defined to be any ζ which optimizes this linear program. The linear program depends only upon $\operatorname{sign}(x)$ when the cost is linear. Given the feedback law f, we then set $\frac{d}{dt}z(t;x) = f(q(t;x))$. As was seen in the previous examples, in many cases the greedy policy leads to a pathwise optimal solution—geometric conditions ensuring this are developed in [43].

The following is a generalization of a result of [10].

PROPOSITION 11. *Suppose that the fluid model is stabilizable, and the cost function c is C^1. In this case any greedy fluid policy f is stabilizing for the fluid model, and it is pathwise optimal if a unique pathwise optimal solution exists.*

Moreover, assumption (A1) *holds with the Lyapunov function* $V(x) = \beta c(x)$ *for* $\beta > 0$ *sufficiently large.*

Proof. As in the construction of a solution to (2.16), we can use stabilizability to ensure the existence of an $\varepsilon > 0$ such that the equation

$$B\zeta + \alpha = -\varepsilon \frac{x}{\|x\|}$$

has a solution $\zeta^x \in \mathsf{U}$ for any $x \neq \theta$. Hence, under the greedy policy we have, when $q(t) = x \neq \theta$,

$$\begin{aligned} \frac{d}{dt} c(q(t)) &\leq \langle \nabla c, B\zeta^x + \alpha \rangle \\ &\leq -\frac{\varepsilon}{\|x\|} \langle \nabla c, x \rangle = -\varepsilon \frac{c(x)}{\|x\|}. \end{aligned}$$

The last equality follows from radial homogeneity of the norm c. This implies the result with $\beta = \varepsilon^{-1} \max_{x \neq \theta} \{\|x\|/c(x)\}$. ∎

For the first four models shown above in Figures 4, 6, 8, and 10, the greedy fluid policy is pathwise optimal. Hence it attains the minimal cost $V^*(x, T)$ for any x and any $T > 0$. The model shown in Figure 12 is a re-entrant line for which the greedy policy is the LBFS priority policy. In this example the LBFS policy is *not* optimal for the fluid model since it results in excessive starvation at the second machine whenever the second machine is the bottleneck (see Figure 3).

Note that the greedy policy for a *discrete network* is typically defined to be the policy which, at time t, minimizes over all admissible actions a the value $\mathsf{E}[c(Q(k+1)) \mid Q(k), a]$. For any network scheduling problem this policy gives strict priority to exit buffers when $c(\,\cdot\,) = |\,\cdot\,|$. In general, such a policy may perform extremely poorly: For the example given in Figure 10 the greedy fluid policy is pathwise optimal, but we saw in section 2 that the priority policy is destabilizing for some parameter values even under (2.13).

4.3. Information. The policies considered thus far require the following information for successful design:

- the arrival rates α,
- service rates and routing information as coded in the matrix B,
- bounds on variability of $(\boldsymbol{A}, \boldsymbol{R}, \boldsymbol{S})$ so that appropriate safety-stocks can be defined,
- global state information $q(t; x)$ for each time t.

Relaxing this information structure is of interest in various applications.

We consider here only the first issue: In telecommunications applications we may know little about arrival rates to the system, and in a manufacturing application *demand* may be uncertain. Sensitivity with respect to service and arrival rates may be large when the load is close to unity (see [19]).

To obtain a design without knowledge of arrival rates we define a set of *generalized Klimov indices* which assign priorities to buffers, subject to positivity constraints, and buffer level constraints. In this way one can define the policy in terms of observed buffer levels without knowledge of the value of α. One can address (ii) in a similar manner.

Define the permutation $(i_1, \ldots, i_\ell)$ of $\{1, \ldots, \ell\}$ so that for any resource j

$$\langle \nabla c, Be^{i_n} \rangle \leq \langle \nabla c, Be^{i_m} \rangle \qquad \text{if } n < m \text{ and } i_n, i_m \in \mathcal{R}_j.$$

An allocation rate $\zeta = f(x)$ can then be defined as follows: for any $n \ge 1$,

$$\sum\{\zeta_{i_k} : k \le n, i_k \in \mathcal{R}_j\} = 1 \qquad \text{whenever} \quad \sum\{q_{i_k} : k \le n, i_k \in \mathcal{R}_j\} > 0. \tag{4.8}$$

Proposition 12 follows from Proposition 11 since (4.8) may be viewed as an alternative representation of the greedy allocation.

PROPOSITION 12. *Suppose that the network is stabilizable. Then the policy* (4.8) *is stabilizing for the fluid model. It is pathwise optimal if a unique pathwise optimal solution exists.*

4.4. Stability and performance. We are assured of stability under assumptions (A1)–(A3).

THEOREM 13. *Suppose that assumptions* (A1) – (A3) *hold, and suppose that for some* $\varepsilon_1 \ge 0$, $\underline{N} > 0$,

(a) $\limsup_{\|x\|\to\infty} \frac{N(x)}{\|x\|} \le \varepsilon_1$,

(b) $\limsup_{\|x\|\to\infty} \frac{\|\bar{x}(x)\|}{N(x)} \le \varepsilon_1$,

(c) $\liminf_{\|x\|\to\infty} N(x) \ge \underline{N}$.

Then, for all $\underline{N}$ *sufficiently large, the following hold.*

(i) *The state process* $\boldsymbol{Q}$ *is ergodic in the sense that for any initial condition* $Q(0) = x$ *and any function* g *which is bounded by some polynomial function of* x, *there exists a finite* $\nu(g)$ *such that as* $T \to \infty$,

$$\frac{1}{T}\sum_{0}^{T-1} g(Q(k)) \to \nu(g) \qquad a.s.,$$
$$\mathsf{E}_x[g(Q(T))] \to \nu(g).$$

(ii) *There exists* $\Delta > 0$ *such that if* $\boldsymbol{Q}^\Delta$ *is the state process for a new network satisfying* $\|B^\Delta - B\| \le \Delta$, $\|\alpha^\Delta - \alpha\| \le \Delta$, *then the policy will continue to stabilize the perturbed system in the sense of* (i).

Proof. The idea of the proof is to construct a constant b and $\varepsilon > 0$ such that

$$P^{N(x)}V(x) = \mathsf{E}[V(Q(N(x);x))] \le V(x) - \varepsilon N(x), \qquad \|x\| \ge b. \tag{4.9}$$

From the Lipschitz continuity of the model we can then find for each p a $b_p < \infty$ and $\varepsilon_p > 0$ such that

$$P^{N(x)}V^p(x) \le V^p(x) - \varepsilon_p N(x) V^{p-1}(x), \qquad \|x\| \ge b_p.$$

Since V is equivalent to a norm on $\mathbb{R}^\ell$, we can argue as in [15] that

$$P^{N(x)}V^p(x) \le V^p(x) - \varepsilon_p \mathsf{E}_x\left[\sum_{0}^{N(x)-1} \|Q(k)\|^{p-2}\right], \qquad \|x\| \ge b_p,$$

where the constants b_p, ε_p may have to be adjusted but remain finite and nonzero. It follows that the process $\boldsymbol{\Phi}$ is g-regular, with $g(x) = \|x\|^q$, for any $q \ge 1$. The ergodic theorems then follow, and, in fact, the ergodic limit in (i) converges faster than any polynomial function of time (see [15]).

How then do we establish (4.9)? For $x \in \mathsf{X}$ let $T = N(x)$, and write

$$\begin{aligned} V(Q(T;x)) &= V\Big(q(T;(x-\bar{x})^+) + (Q(T;x) - q(T;(x-\bar{x})^+)\Big) \\ &\le V(q(T;(x-\bar{x})^+) + b_0\|Q(T;x) - q(T;(x-\bar{x})^+)\|, \end{aligned}$$

where the inequality follows from Lipschitz continuity. The desired bound easily follows since $V(q(T;(x-\bar{x})^+) \le V((x-\bar{x})^+) - T$ for all x sufficiently large.

This proves (i), and (ii) follows since the drift inequality is preserved under perturbations in the model when V is Lipschitz.

Note that if N and $\bar{x}$ are bounded, then, by following the proof of Theorem 4 and the arguments here, one can show that the state process $\mathbf{\Phi}$ is V_ε-uniformly ergodic. □

Given this large class of policies, how can we compare one to another? If our goal is to estimate η, the steady-state mean of $c(Q(k))$ under a given policy, and if the one step cost c is linear, then bounds on η can be obtained by solving certain linear programs [36, 12, 34] or through comparison methods with a simpler model for which performance is readily computed [48]. If these bounds are not useful, then one can resort to simulation.

The standard estimator of η is given by $\widehat{\eta}(k) := k^{-1}\sum_0^{k-1} c(Q(i))$, and this estimator is strongly consistent for the policies considered here. From g-regularity with $g(\,\cdot\,) = \|\,\cdot\,\|^4$ we can also establish a central limit theorem of the form $\sqrt{t}(\widehat{\eta}(k)-\eta) \Rightarrow \sigma N(0,1)$, where $\Rightarrow$ denotes weak convergence, and $N(0,1)$ is a standard normal random variable [47]. The constant σ^2 is known as the *time-average variance constant* (TAVC) and provides a measure of the effectiveness of $\widehat{\eta}(k)$.

The problem with simulation is that the TAVC is large in heavy traffic. It is known that the TAVC is of order $(1-\rho)^{-4}$ for the M/M/1 queue [1, 57], and similar bounds hold for other network control problems [30]. With such a large variance, long run-lengths will be required to estimate η effectively.

One method of reducing variance is through control variates: For any function $h\colon \mathsf{X} \to \mathbb{R}$ let $\Delta_h = h - Ph$. Here the transition function P may define the statistics of the process $\mathbf{\Phi}$, in which case h is interpreted as a function of the first component of $\mathbf{\Phi}$ only. If the function h is π-integrable, then $\pi(\Delta_h) = 0$, and so one might use the consistent estimator

$$\widehat{\eta}_c(n) = \widehat{\eta}(n) + \frac{1}{n}\sum_{i=0}^{n-1} \Delta_h(Q(i)). \tag{4.10}$$

This approach can lead to substantial variance reductions, especially in heavy traffic, when applied to the GI/G/1 queue [29].

In [30] these ideas are extended to network models. First note that the estimator (4.10) will have a variance of *zero* when h solves the Poisson equation $\Delta_h = -c + \eta$. While this choice is not computable in general, we can approximate h by the fluid value function

$$V(y) := \int_0^\infty c(q(t;x))\,dt.$$

The *fluid estimator* of η is then given by (4.10) with $h = V$,

$$\widehat{\eta}_f(n) = \widehat{\eta}(n) + \frac{1}{n}\sum_{k=0}^{n-1} \Delta_V(\Phi(k)). \tag{4.11}$$

Why should V provide an approximation to the solution of Poisson's equation? This actually follows from the construction of the feedback regulation policy which requires that the increments of $\boldsymbol{Q}$ approximate the increments of the fluid trajectories. See [30] for details.

REFERENCES

[1] S. Asmussen, *Queueing simulation in heavy traffic*, Math. Oper. Res., 17 (1992), pp. 84–111.

[2] N. Baëuerle, *Asymptotic optimality of tracking-policies in stochastic networks*, Ann. Appl. Probab., to appear.

[3] S. Balaji and S. P. Meyn, *Multiplicative ergodic theorems and large deviations for an irreducible Markov chain*, Stochastic Process. Appl., 90 (2000), pp. 123–144.

[4] S. L. Bell and R. J. Williams, *Dynamic scheduling of a system with two parallel servers: Asymptotic optimality of a continuous review threshold policy in heavy traffic*, in Proceedings of the 38th IEEE Conference on Decision and Control, Pheonix, AZ, 1999, pp. 1743–1748.

[5] D. Bertsekas and R. Gallager, *Data Networks: Upper Saddle River*, Prentice-Hall, Englewood Cliffs, NJ, 1992.

[6] M. Bramson, *Instability of FIFO queueing networks*, Ann. Appl. Probab., (1994), pp. 414–431.

[7] H. Chen and A. Mandelbaum, *Discrete flow networks: Bottlenecks analysis and fluid approximations*, Math. Oper. Res., 16 (1991), pp. 408–446.

[8] H. Chen and D. D. Yao, *Dynamic scheduling of a multiclass fluid network*, Oper. Res., 41 (1993), pp. 1104–1115.

[9] R-R. Chen and S. P. Meyn, *Value iteration and optimization of multiclass queueing networks*, Queueing Systems Theory Appl., 32 (1999), pp. 65–97.

[10] D. P. Connors, G. Feigin, and D. Yao, *Scheduling semiconductor lines using a fluid network model*, IEEE Trans. Robotics Automation, 10 (1994), pp. 88–98.

[11] R. L. Cruz, *A calculus for network delay, part* I: *Network elements in isolation*, IEEE Trans. Inform. Theory, 37 (1991), pp. 114–131.

[12] I. Paschalidis, D. Bertsimas, and J. N. Tsitsiklis, *Optimization of multiclass queueing networks: Polyhedral and nonlinear characterizations of achievable performance*, Ann. Appl. Probab., 4 (1994), pp. 43–75.

[13] J. Humphrey, D. Eng, and S. P. Meyn, *Fluid network models: Linear programs for control and performance bounds*, in Proceedings of the 13th IFAC World Congress, Vol. B, J. Cruz, J. Gertler, and M. Peshkin, eds., San Francisco, CA, 1996, pp. 19–24.

[14] J. G. Dai, *On the positive Harris recurrence for multiclass queueing networks: A unified approach via fluid limit models*, Ann. Appl. Probab., 5 (1995), pp. 49–77.

[15] J. G. Dai and S. P. Meyn, *Stability and convergence of moments for multiclass queueing networks via fluid limit models*, IEEE Trans. Automat. Control, 40 (1995), pp. 1889–1904.

[16] J. G. Dai and G. E. Weiss, *Stability and instability of fluid models for reentrant lines*, Math. Oper. Res., 21 (1996), pp. 115–134.

[17] J. G. Dai and G. E. Weiss, *A fluid heuristic for minimizing makespan in job-shops*, Oper. Res., to appear.

[18] B. T. Doshi, *Optimal control of the service rate in an $M/G/1$ queueing system*, Adv. Appl. Probab., 10 (1978), pp. 682–701.

[19] P. Dupuis and K. Ramanan, *A Multiclass Feedback Queueing Network with a Regular Skorokhod Problem*, Preprint, Brown University, Providence, RI, 1999.

[20] P. Dupuis and R. J. Williams, *Lyapunov functions for semimartingale reflecting Brownian motions*, Ann. Appl. Probab., 22 (1994), pp. 680–702.

[21] S. Floyd and V. Jacobson, *Random early detection gateways for congestion avoidance*, IEEE/ACM Trans. Networking, 1 (1993), pp. 397–413.

[22] S. B. Gershwin, *Manufacturing Systems Engineering*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[23] B. Hajek, *Optimal control of two interacting service stations*, IEEE Trans. Automat. Control, 29 (1984), pp. 491–499.

[24] J. M. Harrison, *Brownian models of heterogeneous customer populations*, in Stochastic Differential Systems, Stochastic Control Theory and Applications, IMA Vol. Math. Appl. 10, W. Fleming and P. L. Lions, eds., Springer-Verlag, New York, 1988, pp. 147–186.

[25] J. M. Harrison, *Brownian models of open processing networks: Canonical representations of workload*, Ann. Appl. Probab., 10 (2000), pp. 75–103.

[26] J. M. Harrison and M. J. López, *Heavy traffic resource pooling in parallel-server systems*, Queueing Systems, 33 (1999), pp. 339–368.

[27] J. M. Harrison and L. M. Wein, *Scheduling networks of queues: Heavy traffic analysis of a simple open network*, Queueing Systems Theory Appl., 5 (1989), pp. 265–279.

[28] J. M. Harrison, *The BIGSTEP approach to flow management in stochastic processing networks*, in Stochastic Networks Theory and Applications, F. P. Kelly, S. Zachary, and I. Ziedins, eds., Clarendon Press, Oxford, UK, 1996, pp. 57–89.

[29] S. G. Henderson, *Variance Reduction Via an Approximating Markov Process*, Ph.D. thesis, Stanford University, Stanford, CA, 1997.
[30] S. G. Henderson and S. P. Meyn, *Variance reduction for simulation in multiclass queueing networks*, IIE Transactions on Operations Engineering, 1999, submitted.
[31] C. Humes, Jr., J. Ou, and P. R. Kumar, *The delay of open Markovian queueing networks: Uniform functional bounds, heavy traffic pole multiplicities, and stability*, Math. Oper. Res., 22 (1997), pp. 921–954.
[32] F. C. Kelly and C. N. Laws, *Dynamic routing in open queueing networks: Brownian models, cut constraints and resource pooling*, Queueing Systems Theory Appl., 13 (1993), pp. 47–86.
[33] Y. Kontoyiannis and S. Meyn, *Spectral theory and limit theorems for geometrically ergodic Markov processes*, Ann. Appl. Probab., submitted, 2000.
[34] P. R. Kumar and S. P. Meyn, *Duality and linear programs for stability and performance analysis queueing networks and scheduling policies*, IEEE Trans. Automat. Control, 41 (1996), pp. 4–17.
[35] P. R. Kumar and T. I. Seidman, *Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems*, IEEE Trans. Automat. Control, 35 (1990), pp. 289–298.
[36] S. Kumar and P. R. Kumar, *Performance bounds for queueing networks and scheduling policies*, IEEE Trans. Automat. Control, 39 (1994), pp. 1600–1611.
[37] S. Kumar and P. R. Kumar, *Fluctuation smoothing policies are stable for stochastic re-entrant lines*, J. Discrete Event Dynamic Systems: Theory and Applications, 6 (1996), pp. 361–370.
[38] S. Kumar and M. Muthuraman, *A numerical method for solving singular controls*, in Proceedings of the 39th IEEE Conference on Decision and Control, Volume 1, IEEE Control Systems Society, Piscataway, NJ, 2000, pp. 522–527.
[39] W. Lin and P. R. Kumar, *Optimal control of a queueing system with two heterogeneous servers*, IEEE Trans. Automat. Control, 29 (1984), pp. 696–703.
[40] S. Lippman, *Applying a new device in the optimization of exponential queueing systems*, Oper. Res., 23 (1975), pp. 687–710.
[41] X. Luo and D. Bertsimas, *A new algorithm for state-constrained separated continuous linear programs*, SIAM J. Control Optim., 37 (1998), pp. 177–210.
[42] C. Maglaras, *Design of dynamic control policies for stochastic processing networks via fluid models*, in Proceedings of the 38th IEEE Conference on Decision and Control, Phoenix, AZ, 1997, pp. 1208–1213.
[43] S. P. Meyn, *Sequencing and routing in multiclass queueing networks. Part* II: *Workload relaxations*, SIAM J. Control Optim., submitted, 2000.
[44] S. P. Meyn, *The policy improvement algorithm for Markov decision processes with general state space*, IEEE Trans. Automat. Control, 42 (1997), pp. 1663–1680.
[45] S. P. Meyn, *Stability and optimization of multiclass queueing networks and their fluid models*, in Mathematics of Stochastic Manufacturing Systems: AMS–SIAM Summer Seminar in Applied Mathematics, Lectures in Applied Mathematics 33, AMS, Providence, RI, 1997, pp. 175–199.
[46] S. P. Meyn and D. Down, *Stability of generalized Jackson networks*, Ann. Appl. Probab., 4 (1994), pp. 124–148.
[47] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, London, 1993.
[48] J. Ou and L. M. Wein, *Performance bounds for scheduling queueing networks*, Ann. Appl. Probab., 2 (1992), pp. 460–480.
[49] J. R. Perkins and P. R. Kumar, *Optimal control of pull manufacturing systems*, IEEE Trans. Automat. Control, 40 (1995), pp. 2040–2051.
[50] M. C. Pullan, *Forms of optimal solutions for separated continuous linear programs*, SIAM J. Control Optim., 33 (1995), pp. 1952–1977.
[51] C. Okino, R. Agrawal, R. L. Cruz, and R. Rajan, *Performance bounds for flow control protocols*, IEEE Trans. Networking, 7 (1999), pp. 310–323.
[52] A. N. Rybko and A. L. Stolyar, *On the ergodicity of stochastic processes describing the operation of open queueing networks*, Problemy Peredachi Informatsii, 28 (1992), pp. 3–26.
[53] J. G. Shanthikumar and D. D. Yao, *Stochastic monotonicity in general queueing networks*, J. Appl. Probab., 26 (1989), pp. 413–417.
[54] A. Shwartz and A. Weiss, *Large deviations for performance analysis: Queues, communication and computing*, Chapman-Hall, London, 1995.

[55] S. Lavenberg, ed., *Computer Performance Modeling Handbook*, Academic Press, New York, 1983.
[56] G. Weiss, *Optimal draining of fluid re-entrant lines: Some solved examples*, in Stochastic Networks: Theory and Applications, Roy. Statist. Soc. Lecture Note Ser. 4, F. P. Kelly, S. Zachary, and I. Ziedins, eds., Oxford University Press, Oxford, UK, 1996, pp. 19–34.
[57] W. Whitt, *Planning queueing simulations*, Management Sci., 35 (1994), pp. 1341–1366.
[58] S. H. Xu and H. Chen, *A note on the optimal control of two interacting service stations*, IEEE Trans. Automat. Control, 38 (1993), pp. 187–189.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 3, pp. 777–800

A ROTATED MULTIPLIER APPLIED TO THE CONTROLLABILITY OF WAVES, ELASTICITY, AND TANGENTIAL STOKES CONTROL*

AXEL OSSES†

Abstract. A new family of multipliers with rotated direction is introduced. This technique is applied to obtain new results concerning controllability of waves, elasticity, and Stokes equations. The boundary exact controllability for the wave equation and the dynamic elasticity system is reviewed generalizing the classical exit condition in the case of explicit observability constants. Approximate controllability for the Stokes system is also studied using a boundary control acting only on the tangential component of the velocity. A geometric sufficient condition of exit generalized type is deduced.

Key words. multiplier method, exact controllability, approximate controllability, unique continuation, wave equation, elasticity, Stokes system

AMS subject classifications. 93B05, 35B37, 35B60, 35Q30, 73K50

PII. S0363012998345615

1. Introduction.

In 1940, Rellich [41] introduced a multiplier technique in order to obtain direct a priori estimates in linear partial differential equations (PDEs). This method was called multiplier method since it consists in multiplying the equation by the gradient of the solution following some convenient vector field and then integrating by parts in the domain. This technique was widely used in the classical PDE development [32], [11]. Later on, in the 70s and 80s, this technique was used to derive inverse estimates: the asymptotic estimates in scattering theory for unbounded domains [30], [31] and the direct study of uniform stabilization in bounded domains [5], [6], [17], [21].

In 1986, Ho [10] used this multiplier technique to prove an inverse inequality for the linear wave equation implying its exact controllability. Ho arrived to a geometrical condition called *exit condition*: The control region must contain a subset of the boundary where the scalar product between the outward normal and the vector pointing from some origin towards the normal is positive. By varying the origin, a family of control boundaries satisfying the condition is found. In a square, for instance, the condition gives control boundaries consisting of four, three, or two adjacent sides. Ho's result was improved [25], [24] and adapted to other systems like vibrating plates and the elasticity system [24], [19]. Afterwards, the method also gave similar results for Maxwell [18], [16] and Schrödinger [29] equations.

Several authors have used multiplier techniques for control or stabilization of mathematical models: viscoelastic or thermoelastic beams [23], [12]; semilinear wave equations [44]; wave equation with mixed boundary conditions [8], [7] or in domains with corners or cracks [9], [33], [34]; Euler–Bernoulli equations [13]; hybrid systems in elasticity [40]; networks of membranes or beams with discontinuous coefficients [24], [20]; coupled Schrödinger equations [14]; Korteweg–de Vries equations [42]. See also

*Received by the editors October 7, 1998; accepted for publication (in revised form) October 11, 2000; published electronically September 28, 2001. This work was partially supported by FONDAP through its Program on Mathematical–Mechanics.

http://www.siam.org/journals/sicon/40-3/34561.html

†Departamento de Ingeniería Matemática, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Casilla 170/3 - Correo 3, Santiago, Chile, and Centro de Modelamiento Matemático, UMR 2071 CNRS-Uchile (axosses@dim.uchile.cl).

[15] for other references.

In recent years, microlocal techniques and geometric optics analysis allow us to find geometrical characterization of control location and minimal control time in the exact controllability of waves. After eventual reflection, diffraction, or sliding on the boundary, every optic ray issue from the observation domain has to reach the control zone. This is a necessary and sufficient condition [4] called the *Bardos–Lebeau–Rauch (BLR) condition* [1]. This technique has also been applied to vibrating plates [22], the elasticity system [2], and Maxwell equations [39].

Exit condition turns to be a particular case of the BLR condition. But there is a certain balance: In the BLR condition, control time is optimal but the observability constants are not explicit. In the exit condition, time is not optimal, but observability, constants can be explicit and this fact is very useful in theoretical and numerical estimations. In general [3], the BLR condition assumes more regularity on coefficients and boundaries than exit condition.

In this article we introduce a family of multipliers with rotated direction as a new approach in the multiplier method. More precisely, we propose to multiply the equation by the gradient of the solution following not only a radial but also a rotated direction. This takes advantage of invariances under rotations for the differential operators considered here and leads to derive a *generalized exit condition*. For instance, in two dimensions, the condition is that the control region must contain a subset of the boundary where the scalar product between the outward normal rotated in an angle and the vector pointing from some origin towards the normal is positive. A family of control boundaries is obtained by varying the origin and the angle. The minimal time of control results to be proportional to the inverse of the cosine of the angle of the rotation.

To show the particularities of this method we have chosen some controllability problems, but the technique could be a useful tool in other areas. We revisit the exact or approximate controllability of some linear classical models in PDEs: the wave equation, the elasticity system, and the Stokes system.

The paper is organized as follows. In section 2, rotated multipliers are used to derive a generalization of classical inverse inequality for the linear wave equation conserving explicit observability constants (see Theorem 2.2 and Theorem 2.3). New boundary control geometries are found (Figures 2.1 and 2.2) which are particular cases of the BLR condition and satisfy a generalized exit condition. In section 3, the method is extended to the study of the exact controllability of the elasticity system. Beside the choice of a rotated direction in a natural manner, a second multiplier formula is needed in this case. The classical inverse inequality with explicit constants is also generalized (see Theorem 3.3 and Theorem 3.4). The same geometrical conditions as for the wave equation are found. In section 4, a different application in fluid control is developed: the study of the approximate controllability of the Stokes system with a boundary control acting only on the tangential part of the velocity. As far as we know, this is an almost untreated topic (except for references [35] and [36]). A sufficient geometric condition is found similar to one deduced for the wave equation and the elasticity system. The final results are presented in two and three dimensions, but the technique is actually not limited by dimension.

In the case of controlling all the velocity trace on an arbitrarily small nonempty open part of the boundary, approximate controllability is easily obtained by using a unique continuation property of Holmgren's type. Second, approximate controllability using the normal component of the velocity is studied in [27] and [28], where the result

is proved in a real analytic connected domain with a simple spectrum for the Laplacian with a control acting on an arbitrary small nonempty open part of the boundary and a counterexample in a ball is given. It is amazing to observe that the normal boundary approximate controllability also holds if the boundary has at least a rectangular corner [38]. In the tangential case that we treat in this paper, we begin by following the idea of [27] introducing a spectral decomposition to characterize the unique continuation property of the time dependent system as a unique continuation property on each frequency. Then we use rotated multipliers to obtain an inverse inequality for each eigenfrequency and a sufficient geometrical condition to have the unique continuation (see Theorem 4.3). We prove that this condition is not necessary for two dimensional connected domains with analytic boundary (see Theorem 4.4). But, as far as all the other cases are concerned, there is a lack of counterexamples for which tangential boundary approximate controllability could not hold.

In summary, multipliers with rotated direction generalize the standard multipliers in a natural way. For second order hyperbolic systems, the application of this technique provides a wider class of geometric examples with explicit observability constants which are particular cases of the BLR condition. For partially controlled Petrovskii systems, the technique reveals to be useful to find results of approximate controllability.

2. Wave equation.[1]

2.1. Control problem. Let Ω be a bounded domain of $\mathbb{R}^N$ $(N \geq 2)$ with a regular[2] boundary Γ of class C^2. Let ν be the unit exterior normal to Ω. Let $T > 0$ be given, and define $Q = \Omega \times (0,T)$ and $\Sigma = \Gamma \times (0,T)$. We consider the following classical control problem. Let $\Gamma_0 \subseteq \Gamma$ and $\Sigma_0 = \Gamma_0 \times (0,T)$. Our problem consists of finding T_0 such that for each $T > T_0$ and for every $(y_0, y_1) \in L^2(\Omega)^N \times H^{-1}(\Omega)^N$, there exists $v \in L^2(\Sigma_0)^N$ in such a way that the solution of the wave equation

$$y'' - \Delta y = 0 \quad \text{in} \quad Q, \tag{2.1a}$$

$$y = v \quad \text{on} \quad \Sigma_0, \tag{2.1b}$$

$$y = 0 \quad \text{on} \quad \Sigma \setminus \Sigma_0, \tag{2.1c}$$

$$y(0) = y_0,\ y'(0) = y_1 \quad \text{in} \quad \Omega \tag{2.1d}$$

satisfies

$$y(T) = 0,\ y'(T) = 0 \quad \text{in} \quad \Omega, \tag{2.2}$$

where the prime symbol $'$ stands for derivation with respect to time.

Following the Hilbert uniqueness method (HUM) [24], the solution to this problem is equivalent to studying the observability properties of the adjoint problem. For each pair of initial conditions $(\varphi_0, \varphi_1) \in H_0^1(\Omega)^N \times L^2(\Omega)^N$, let us consider the solution φ of the wave equation as follows:

$$\varphi'' - \Delta \varphi = 0 \quad \text{in} \quad Q, \tag{2.3a}$$

$$\varphi = 0 \quad \text{on} \quad \Sigma, \tag{2.3b}$$

$$\varphi(0) = \varphi_0,\ \varphi'(0) = \varphi_1 \quad \text{in} \quad \Omega. \tag{2.3c}$$

[1] A note about the results of this section was published in [37].

[2] The results of this section are also valid if we suppose that Ω is either a bounded polygonal of $\mathbb{R}^2$ or a bounded polyhedral of $\mathbb{R}^3$. (It suffices to apply the methods of Grisvard [8].)

More precisely, exact controllability is equivalent to demonstrate that for $T > T_0$, the inverse inequality

$$E_0 \leq C(\Omega, T) \int_{\Sigma_0} \left| \frac{\partial \varphi}{\partial \nu} \right|^2 \, d\sigma dt$$

holds, where

$$E_0 = \frac{1}{2} \left(\int_\Omega |\nabla \varphi_0|^2 \, dx + \int_\Omega |\varphi_1|^2 \, dx \right) \tag{2.4}$$

is the initial energy of system (2.3), and $C(\Omega, T)$ is a constant depending only on geometry and final time. Multiplier methods [24] can give explicit constants, but only for large Γ_0 and T. Microlocal techniques [1], [4] characterize all Γ_0 and T for which we obtain such a result, but in this case the constant is not explicit.

Using a new choice in the classical multiplier method, we will enlarge the set of geometric examples with explicit knowledge of constants.

2.2. Inverse inequality and exact controllability.

DEFINITION 2.1. *Let $A \in \mathbb{R}^{N\times N}$ be such that $A = -A^t$ (skew-symmetric). Let $d > 0$ be a positive real number and I the identity matrix in $\mathbb{R}^{N\times N}$. We define for each $x^0 \in \mathbb{R}^N$ the set*

$$\Gamma(x^0, d, A) = \{x \in \Gamma \text{ such that } (x - x^0) \cdot (dI + A)\nu > 0\}. \tag{2.5}$$

Without loss of generality we introduce the following normalizing condition:

$$d^2 + \|A\|_2^2 = 1, \tag{2.6}$$

where $\|A\|_2 = \sup\{|Ax|\,,\ |x| = 1\}$ and $|\cdot|$ is the Euclidean norm in $\mathbb{R}^N$. We also define

$$r(x^0, d, A) = \max\{(x - x^0) \cdot (dI + A)\nu \text{ with } x \in \Gamma(x^0, d, A)\}, \tag{2.7a}$$

$$R(x^0) = \max\{\left|x - x^0\right| \text{ with } x \in \overline{\Omega}\}. \tag{2.7b}$$

THEOREM 2.2 (inverse inequality). *Given $x^0 \in \mathbb{R}^N$, $d > 0$, and a skew-symmetric matrix A normalized as in* (2.6)*, for each $T > 2d^{-1}R(x^0)$ and for each weak solution φ of* (2.3) *the following inequality holds:*

$$E_0 \leq \frac{r(x^0, d, A)}{2\,(dT - 2R(x^0))} \int_0^T \int_{\Gamma(x^0,d,A)} \left| \frac{\partial \varphi}{\partial \nu} \right|^2 \, d\sigma dt. \tag{2.8}$$

THEOREM 2.3 (exact controllability). *Suppose that we can find $x^0 \in \mathbb{R}^N$, $d > 0$, and a skew-symmetric matrix A normalized as in* (2.6) *such that $\Gamma(x^0, d, A)$ is not empty and $\Gamma(x^0, d, A) \subset \Gamma_0$; then for each $T > 2d^{-1}R(x^0)$ there exists a control $v \in L^2(\Sigma_0)^N$ such that the corresponding solution of* (2.1) *satisfies the final condition* (2.2).

Remark 1. In the case $d = 1$ and $A = 0$ we recover classical results (see [24]).

Remark 2. For $N = 2$, introducing $\theta \in]-\pi/2, \pi/2[$, taking $d = \cos\theta$ and $A_{21} = A_{12} = \sin\theta$, definition (2.5) can be replaced by

$$\Gamma(x^0, \theta) = \{x \in \Gamma \text{ such that } (x - x^0) \cdot M(\theta)\nu > 0\}, \tag{2.9}$$

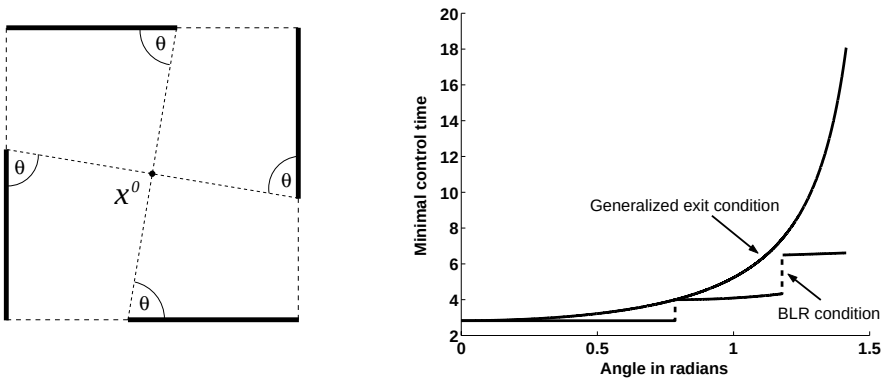


FIG. 2.1. *Left: Control region* $\Gamma(x^0, \theta)$ *(bold line) in the square* $]-1,1[^2$ *for* x^0 *centered and* $\theta < \pi/2$. *Theorem* 2.3 *gives a control time* $T > 2\sqrt{2}/\cos\theta$. *Right: Comparison between BLR minimal control time and the minimal time given by Theorem* 2.3 *for this example.*

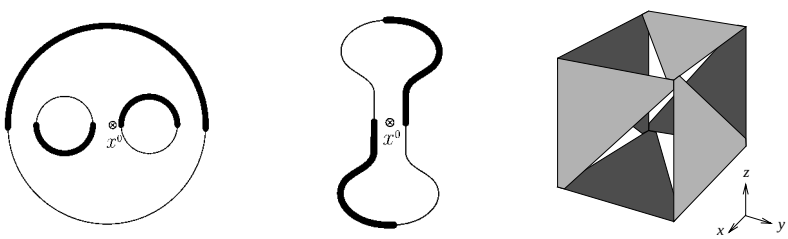


FIG. 2.2. *Other regions of control obtained by applying a rotated multiplier technique. Left: Control region (bold line) for the Ikawa's bowling ball and a bone-shape region for* θ *near* $\pi/2$. *Right: Control region* $\Gamma(x^0, d, \alpha)$ *(in gray) in the cube* $]-1,1[^3$ *for a centered* x^0, $d = 0.1$ *and* α *in the direction* $-(1,1,1)$.

where $M(\theta)$ is a rotation matrix of angle θ anticlockwise (see Figures 2.1 and 2.2).

Remark 3. For $N = 3$, if $\alpha \in \mathbb{R}^3$ and $d^2 + |\alpha|^2 = 1$ we take $A_{12} = -\alpha_3$, $A_{13} = \alpha_2$, and $A_{23} = -\alpha_1$, and the definition (2.5) can be written using the exterior product in $\mathbb{R}^3$ as (see Figure 2.2, right)

$$\Gamma(x^0, d, \alpha) = \{x \in \Gamma \text{ such that } (x - x^0) \cdot (d\nu + \alpha \times \nu) > 0\}. \tag{2.10}$$

2.3. Rotated multiplier: Proof of Theorems 2.2 and 2.3. Let φ be a weak solution of (2.3). Multiplying (2.3) by $\nabla\varphi \cdot q$, where $q \in W^{1,\infty}(\overline{\Omega})^N$, and by φ and integrating by parts, the following classical formulas [24, Chapter I], are deduced:

$$\begin{aligned}(\varphi'(t)\,,\, q \cdot \nabla\varphi)_{0,\Omega}\Big|_{t=0}^{T} + \frac{1}{2}\int_Q \operatorname{div} q(|\varphi'|^2 - |\nabla\varphi|^2)\,dxdt \\ + \int_Q (\nabla q)\nabla\varphi \cdot \nabla\varphi\,dxdt = \frac{1}{2}\int_\Sigma q \cdot \nu \left|\frac{\partial\varphi}{\partial\nu}\right|^2 d\sigma dt\end{aligned} \tag{2.11}$$

and

$$(\varphi'(t)\,,\, \varphi(t))_{0,\Omega}\Big|_{t=0}^{T} = \int_Q (|\varphi'|^2 - |\nabla\varphi|^2)\,dxdt, \tag{2.12}$$

where $(\cdot\,,\,\cdot)_{0,\Omega}$ and $\|\cdot\|_{0,\Omega}$ denote the usual inner product and norm in $L^2(\Omega)^N$, respectively.

We consider now in (2.11) a direction of type

$$q = (dI - A)(x - x^0), \tag{2.13}$$

where $d > 0$ and $A = -A^t$ verify the normalizing condition (2.6). Note that $\operatorname{div} q = dN$, $\nabla q = dI - A$, and $(A\nabla\varphi\,,\,\nabla\varphi)_{0,\Omega} = 0$. With this choice (2.11) becomes

$$\begin{aligned}(\varphi'\,,\,q\cdot\nabla\varphi)_{0,\Omega}\Big|_{t=0}^{T} + \frac{dN}{2}\int_Q (|\varphi'|^2 - |\nabla\varphi|^2)\,dxdt \\ + d\int_Q |\nabla\varphi|^2\,dxdt = \frac{1}{2}\int_\Sigma q\cdot\nu\left|\frac{\partial\varphi}{\partial\nu}\right|^2 d\sigma dt.\end{aligned} \tag{2.14}$$

If we add up this last identity to (2.12) multiplied by $d(N-1)/2$, we obtain

$$\left(\varphi'\,,\,q\cdot\nabla\varphi + \frac{d(N-1)}{2}\varphi\right)_{0,\Omega}\Big|_{t=0}^{T} + \frac{d}{2}\int_Q (|\varphi'|^2 + |\nabla\varphi|^2)\,dxdt = \frac{1}{2}\int_\Sigma q\cdot\nu\left|\frac{\partial\varphi}{\partial\nu}\right|^2 d\sigma dt.$$

In virtue of the energy conservation principle it follows that

$$\left(\varphi'\,,\,q\cdot\nabla\varphi + \frac{d(N-1)}{2}\varphi\right)_{0,\Omega}\Big|_{t=0}^{T} + dTE_0 = \frac{1}{2}\int_\Sigma q\cdot\nu\left|\frac{\partial\varphi}{\partial\nu}\right|^2 d\sigma dt. \tag{2.15}$$

Now, from

$$(q\cdot\nabla\varphi\,,\,\varphi)_{0,\Omega} = -\frac{1}{2}\int_\Omega \operatorname{div} q\,|\varphi|^2\,dx = -\frac{dN}{2}\,\|\varphi\|_{0,\Omega}^2\,,$$

we can see that

$$\begin{aligned}\left\|q\cdot\nabla\varphi + \frac{d(N-1)}{2}\varphi\right\|_{0,\Omega}^2 &= \|q\cdot\nabla\varphi\|_{0,\Omega}^2 - \frac{d^2N(N-1)}{2}\,\|\varphi\|_{0,\Omega}^2 + \frac{d^2(N-1)^2}{4}\,\|\varphi\|_{0,\Omega}^2 \\ &\le \|q\cdot\nabla\varphi\|_{0,\Omega}^2 + \frac{d^2(1-N^2)}{4}\,\|\varphi\|_{0,\Omega}^2 \\ &\le \|q\cdot\nabla\varphi\|_{0,\Omega}^2\,.\end{aligned}$$

The above inequality implies that the first term in the left hand side of (2.15) is bounded by

$$2\left(\frac{R(x^0)}{2}\,\|\varphi'\|_{0,\Omega}^2 + \frac{1}{2R(x^0)}\,\|q\cdot\nabla\varphi\|_{0,\Omega}^2\right),$$

where $R(x^0)$ was defined in (2.7). Using the normalization condition (2.6), we obtain

$$\|q\cdot\nabla\varphi\|_{0,\Omega} \le (d^2 + \|A\|_2^2)^{1/2} R(x^0)\,\|\nabla\varphi\|_{0,\Omega} = R(x^0)\,\|\nabla\varphi\|_{0,\Omega}\,.$$

Therefore, from (2.15) we deduce that

$$-2R(x^0)E_0 + dTE_0 \le \frac{1}{2}\int_\Sigma (dI - A)(x - x^0)\cdot\nu\left|\frac{\partial\varphi}{\partial\nu}\right|^2 d\sigma dt.$$

If we note that $(dI-A)(x-x^0)\cdot\nu = (x-x^0)\cdot(dI+A)\nu$, we have only to use definitions (2.5) of $\Gamma(x^0,d,A)$ and (2.7a) of $r(x^0,d,A)$ in order to conclude the inverse inequality (2.8) and Theorem 2.2.

The exact controllability result of Theorem 2.3 follows directly from Theorem 2.2 applying HUM method (see [24, Chapter IV]). □

3. Elasticity system.

3.1. Control problem. We consider an isotropic homogeneous elastic body occupying a bounded open subset Ω of $\mathbb{R}^N$. We keep the same regularity assumptions and notations of section 2. We introduce the boundary Γ of Ω, the control boundary Γ_0, and given a final time $T > 0$, the associated cilinders $Q = \Omega \times (0,T)$, $\Sigma = \Gamma \times (0,T)$, and $\Sigma_0 = \Gamma_0 \times (0,T)$. We study the exact controllability of the system of linear elasticity with a control acting on a part of the boundary. More precisely, given $\mathbf{f} \in L^2(Q)^N$, a control $\mathbf{v} \in L^2(\Sigma_0)^N$, and initial conditions $\mathbf{u}^0 \in L^2(\Omega)^N$ and $\mathbf{u}^1 \in H^{-1}(\Omega)^N$, let $\mathbf{u}$ be the solution of

$$\mathbf{u}'' - \mu\Delta\mathbf{u} - (\lambda+\mu)\nabla\operatorname{div}\mathbf{u} = \mathbf{f} \quad \text{in} \quad Q, \tag{3.1a}$$

$$\mathbf{u} = \mathbf{v} \quad \text{on} \quad \Sigma_0, \tag{3.1b}$$

$$\mathbf{u} = \mathbf{0} \quad \text{on} \quad \Sigma \setminus \Sigma_0, \tag{3.1c}$$

$$\mathbf{u}(0) = \mathbf{u}^0,\ \mathbf{u}'(0) = \mathbf{u}^1 \quad \text{in} \quad \Omega, \tag{3.1d}$$

where μ and λ are Lamé's constants with $\lambda + 2\mu > 0$. The symbol $'$ (prime) means derivation with respect to time. We take the notation $(\Delta\mathbf{u})_i = \partial^2 u_i/\partial x_j \partial x_j$ and the convention that a repeated index in some expression means implicit sum on this index.

Under the conditions described above, system (3.1) has a solution in a transposition sense. It can be shown that $\mathbf{u} \in C([0,T]; L^2(\Omega)^N)$ and also that $\mathbf{u}' \in C([0,T]; H^{-1}(\Omega)^N)$; hence the conditions (3.1d) have a sense.

We seek for a control function $\mathbf{v}$ such that

$$\mathbf{u}(T) = \mathbf{0} \ \text{ and } \ \mathbf{u}'(T) = \mathbf{0}. \tag{3.2}$$

Now, let us consider the solution φ of the adjoint system:

$$\varphi'' - \mu\Delta\varphi - (\lambda+\mu)\nabla\operatorname{div}\varphi = \mathbf{0} \quad \text{in} \quad Q, \tag{3.3a}$$

$$\varphi = \mathbf{0} \quad \text{on} \quad \Sigma, \tag{3.3b}$$

$$\varphi(0) = \varphi^0,\ \varphi'(0) = \varphi^1 \quad \text{in} \quad \Omega \tag{3.3c}$$

for each $\varphi^0 \in H_0^1(\Omega)^N$ and $\varphi^1 \in L^2(\Omega)^N$. From classical regularity results, we know that $\varphi \in C([0,T]; H_0^1(\Omega)^N)$ and $\varphi' \in C([0,T]; L^2(\Omega)^N)$.

If we define the initial energy by

$$E_0 = \frac{1}{2}\int_\Omega \left(\left|\varphi^1\right|^2 + \mu \left|\nabla\varphi^0\right|^2 + (\lambda+\mu)\left|\operatorname{div}\varphi^0\right|^2 \right) dx, \tag{3.4}$$

multiplying (3.3) by φ' we obtain the conservation of energy

$$E(t) = \frac{1}{2}\int_\Omega \left(|\varphi'(t)|^2 + \mu\,|\nabla\varphi(t)|^2 + (\lambda+\mu)\,|\operatorname{div}\varphi(t)|^2 \right) dx = E_0 \text{ for all } t \in [0,T]. \tag{3.5}$$

3.2. Two multiplier formulas. The following geometric property will be useful in this section and in the next section.

PROPOSITION 3.1. *Let Γ_0 be a subset of Γ with positive measure. Let $\varphi \in H^2(\Omega)^N$ be such that the trace of φ on Γ_0 is a constant vector of $\mathbb{R}^N$. Then*

$$\frac{\partial\varphi_i}{\partial x_j}\nu_k = \frac{\partial\varphi_i}{\partial x_k}\nu_j \quad \text{on} \quad \Gamma_0 \quad \forall \text{ different} \quad i,j,k \in \{1,\dots,N\}. \tag{3.6}$$

Proof. We assume that $\varphi \in C^1(\overline{\Omega})^N$ and we can deduce the general case thanks to a density argument. If the symbol $\times$ stands for the exterior product in $\mathbb{R}^N$, the condition imposed to φ on Γ_0 is equivalent to $\nabla\varphi_i \times \nu = 0$ on Γ_0 for each $i = 1, \dots, N$, and this corresponds exactly to (3.6). ∎

We introduce the well-known tensorial product

$$e : f = e_{ij}\, f_{ij},$$

where $e = \{e_{ij}\}_{j=1}^N$ and $f = \{f_{ij}\}_{j=1}^N$ are tensorial fields defined in $\overline{\Omega}$ onto $\mathbb{R}^{N\times N}$. We suppose that this tensorial product has lower precedence than the usual matrix product in $\mathbb{R}^{N\times N}$.

Let q be a vector field defined in $\overline{\Omega}$ with $q \in W^{2,\infty}(\overline{\Omega})$.

Taking the multiplier $(\nabla\varphi)q$ for each term of the left-hand side in (3.3a) the following identities are deduced:

$$\int_\Omega \nabla \operatorname{div}\varphi \cdot (\nabla\varphi)q\, dx = \int_\Omega \frac{\partial^2\varphi_i}{\partial x_i \partial x_j}\frac{\partial\varphi_j}{\partial x_k} q_k\, dx \tag{3.7}$$
$$= -\int_\Omega \frac{\partial\varphi_i}{\partial x_i}\frac{\partial^2\varphi_j}{\partial x_j \partial x_k} q_k\, dx - \int_\Omega \frac{\partial\varphi_i}{\partial x_i}\frac{\partial\varphi_j}{\partial x_k}\frac{\partial q_k}{\partial x_j}\, dx + \int_\Gamma \frac{\partial\varphi_i}{\partial x_i}\frac{\partial\varphi_j}{\partial x_k}\, q_k\nu_j\, d\sigma,$$

but

$$-\int_\Omega \frac{\partial\varphi_i}{\partial x_i}\frac{\partial^2\varphi_j}{\partial x_j \partial x_k} q_k\, dx = \frac{1}{2}\int_\Omega \frac{\partial\varphi_i}{\partial x_i}\frac{\partial\varphi_j}{\partial x_j}\frac{\partial q_k}{\partial x_k}\, dx - \frac{1}{2}\int_\Gamma \frac{\partial\varphi_i}{\partial x_i}\frac{\partial\varphi_j}{\partial x_j}\, q_k\nu_k\, d\sigma$$

and, taking into account Proposition 3.1,

$$\int_\Gamma \frac{\partial\varphi_i}{\partial x_i}\frac{\partial\varphi_j}{\partial x_k}\, q_k\nu_j\, d\sigma = \int_\Gamma \frac{\partial\varphi_i}{\partial x_i}\frac{\partial\varphi_j}{\partial x_j}\, q_k\nu_k\, d\sigma.$$

Hence it follows that

$$\int_\Omega \nabla\operatorname{div}\varphi\cdot(\nabla\varphi)q\, dx = \frac{1}{2}\int_\Omega |\operatorname{div}\varphi|^2 \operatorname{div} q\, dx - \int_\Omega \operatorname{div}\varphi\, \nabla\varphi : \nabla^t q\, dx + \frac{1}{2}\int_\Gamma |\operatorname{div}\varphi|^2\, q\cdot\nu\, d\sigma.$$

The other term gives

$$\int_\Omega \Delta\varphi \cdot (\nabla\varphi)q\, dx = \int_\Omega \frac{\partial^2\varphi_i}{\partial x_j \partial x_j}\frac{\partial\varphi_i}{\partial x_k} q_k\, dx \tag{3.8}$$
$$= -\int_\Omega \frac{\partial\varphi_i}{\partial x_j}\frac{\partial^2\varphi_i}{\partial x_j \partial x_k} q_k\, dx - \int_\Omega \frac{\partial\varphi_i}{\partial x_j}\frac{\partial\varphi_i}{\partial x_k}\frac{\partial q_k}{\partial x_j}\, dx + \int_\Gamma \frac{\partial\varphi_i}{\partial x_j}\frac{\partial\varphi_i}{\partial x_k}\, q_k\nu_j\, d\sigma,$$

but

$$-\int_\Omega \frac{\partial\varphi_i}{\partial x_j}\frac{\partial^2\varphi_i}{\partial x_j \partial x_k} q_k\, dx = \frac{1}{2}\int_\Omega \frac{\partial\varphi_i}{\partial x_j}\frac{\partial\varphi_i}{\partial x_j}\frac{\partial q_k}{\partial x_k}\, dx - \frac{1}{2}\int_\Gamma \frac{\partial\varphi_i}{\partial x_j}\frac{\partial\varphi_i}{\partial x_j}\, q_k\nu_k\, d\sigma,$$

and, from Proposition 3.1,

$$\int_\Gamma \frac{\partial\varphi_i}{\partial x_j}\frac{\partial\varphi_i}{\partial x_k}\, q_k\nu_j\, d\sigma = \int_\Gamma \frac{\partial\varphi_i}{\partial x_j}\frac{\partial\varphi_i}{\partial x_j}\, q_k\nu_k\, d\sigma;$$

then

$$\int_\Omega \Delta\varphi \cdot (\nabla\varphi)q\, dx = \frac{1}{2}\int_\Omega |\nabla\varphi|^2 \operatorname{div} q\, dx - \int_\Omega \nabla\varphi : \nabla\varphi\nabla q\, dx + \frac{1}{2}\int_\Gamma |\nabla\varphi|^2\, q\cdot\nu\, d\sigma.$$

Always with the multiplier $(\nabla\varphi)q$ the last term in (3.3a) gives in a classical manner

$$\int_Q \varphi'' \cdot (\nabla\varphi)q\,dxdt = \frac{1}{2}\int_Q |\varphi'|^2 \operatorname{div} q\,dxdt - \frac{1}{2}\int_\Sigma |\varphi'|^2\, q\cdot\nu\,d\sigma dt + (\varphi'\,,\,(\nabla\varphi)q)_{0,\Omega}\Big|_{t=0}^{T}.$$

Using the other multiplier $(\nabla^t q)\varphi$ for each term of the left-hand side in (3.3a) it follows that

$$\int_\Omega \nabla\operatorname{div}\varphi\cdot(\nabla^t q)\varphi\,dx = \int_\Omega \frac{\partial^2\varphi_i}{\partial x_i\partial x_j}\frac{\partial q_k}{\partial x_j}\varphi_k\,dx \tag{3.9}$$

$$= -\int_\Omega \frac{\partial\varphi_i}{\partial x_i}\frac{\partial^2 q_k}{\partial x_j\partial x_j}\varphi_k\,dx - \int_\Omega \frac{\partial\varphi_i}{\partial x_i}\frac{\partial q_k}{\partial x_j}\frac{\partial\varphi_k}{\partial x_j}\,dx + \int_\Omega \frac{\partial\varphi_i}{\partial x_i}\frac{\partial q_k}{\partial x_j}\,\varphi_k\nu_j\,dx$$

$$= -\int_\Omega \operatorname{div}\varphi\Delta q\cdot\varphi\,dx - \int_\Omega \operatorname{div}\varphi\nabla\varphi : \nabla q\,dx,$$

since $\varphi = 0$ on Γ. For the second term,

$$\int_\Omega \Delta\varphi\cdot(\nabla^t q)\varphi\,dx = \int_\Omega \frac{\partial^2\varphi_i}{\partial x_j\partial x_j}\frac{\partial q_k}{\partial x_i}\varphi_k\,dx \tag{3.10}$$

$$= -\int_\Omega \frac{\partial\varphi_i}{\partial x_j}\frac{\partial^2 q_k}{\partial x_i\partial x_j}\varphi_k\,dx - \int_\Omega \frac{\partial\varphi_i}{\partial x_j}\frac{\partial q_k}{\partial x_i}\frac{\partial\varphi_k}{\partial x_j}\,dx + \int_\Gamma \frac{\partial\varphi_i}{\partial x_j}\frac{\partial q_k}{\partial x_i}\,\varphi_k\nu_j\,d\sigma$$

$$= -\int_\Omega \frac{\partial\varphi_i}{\partial x_j}\frac{\partial^2 q_k}{\partial x_i\partial x_j}\varphi_k\,dx - \int_\Omega \nabla\varphi : (\nabla^t q)\nabla\varphi\,dx.$$

For the last term, always with the second multiplier$(\nabla^t q)\varphi$, integration by parts in Q gives

$$\int_Q \varphi''\cdot(\nabla^t q)\varphi\,dxdt = -\int_Q (\nabla q)\varphi'\cdot\varphi'\,dxdt + \big(\varphi'\,,\,(\nabla^t q)\varphi\big)_{0,\Omega}\Big|_{t=0}^{T}.$$

Combining the identities above, a multiplier formula appears for each multiplier $(\nabla\varphi)q$ and $(\nabla^t q)\varphi$.

LEMMA 3.2. *Let φ be the solution of* (3.3). *For all $q\in W^{2,\infty}(\overline{\Omega})^N$ we have*

$$(\varphi'\,,\,(\nabla\varphi)q)_{0,\Omega}\Big|_{t=0}^{T} + \frac{1}{2}\int_Q \operatorname{div} q(|\varphi'|^2 - \mu\,|\nabla\varphi|^2 - (\lambda{+}\mu)\,|\operatorname{div}\varphi|^2)\,dxdt \tag{3.11}$$

$$+\mu\int_Q \nabla\varphi : \nabla\varphi\nabla q\,dxdt + (\lambda{+}\mu)\int_Q \operatorname{div}\varphi\,\nabla\varphi : \nabla^t q\;dxdt$$

$$= \frac{1}{2}\int_\Sigma q\cdot\nu(\mu\,|\nabla\varphi|^2 + (\lambda{+}\mu)\,|\operatorname{div}\varphi|^2)\,d\sigma dt,$$

$$\big(\varphi'\,,\,(\nabla^t q)\varphi\big)_{0,\Omega}\Big|_{t=0}^{T} = \int_Q (\nabla q)\varphi'\cdot\varphi'\,dxdt - \mu\int_Q \nabla\varphi : \nabla^t q\nabla\varphi\,dxdt \tag{3.12}$$

$$-(\lambda{+}\mu)\int_Q \operatorname{div}\varphi\,\nabla\varphi : \nabla q\,dxdt - \mu\int_Q \frac{\partial\varphi_i}{\partial x_j}\frac{\partial^2 q_k}{\partial x_i\partial x_j}\varphi_k\,dxdt$$

$$-(\lambda{+}\mu)\int_Q \operatorname{div}\varphi\,\Delta q\cdot\varphi\,dxdt.$$

3.3. Choice of the rotated direction. The classical choice $q = x - x^0$, $x^0 \in \mathbb{R}^N$, in (3.11) and (3.12) gives (see [24])

$$(3.13)\quad \big(\varphi'\,,\,\nabla\varphi(x-x^0)\big)_{0,\Omega}\Big|_{t=0}^{T} + \frac{N}{2}\int_Q (|\varphi'|^2 - \mu\,|\nabla\varphi|^2 - (\lambda+\mu)\,|\operatorname{div}\varphi|^2)\,dxdt$$
$$+ \int_Q (\mu\,|\nabla\varphi|^2 + (\lambda+\mu)\,|\operatorname{div}\varphi|^2)\,dxdt = \frac{1}{2}\int_\Sigma (\mu\,|\nabla\varphi|^2 + (\lambda+\mu)\,|\operatorname{div}\varphi|^2)(x-x^0)\cdot\nu\,d\sigma dt$$

and

$$(3.14)\qquad (\varphi'\,,\,\varphi)_{0,\Omega}\Big|_{t=0}^{T} = \int_Q (|\varphi'|^2 - \mu\,|\nabla\varphi|^2 - (\lambda+\mu)\,|\operatorname{div}\varphi|^2)\,dxdt.$$

Now, a rotated direction $q = A(x-x^0)$, $A = -A^t$ (skew-symmetric) in (3.11) and (3.12) gives the following new identities:

$$(3.15)\big(\varphi'\,,\,\nabla\varphi A(x-x^0)\big)_{0,\Omega}\Big|_{t=0}^{T} - (\lambda+\mu)\int_Q \operatorname{div}\varphi\nabla\varphi : A\,dxdt$$
$$= \frac{1}{2}\int_\Sigma (\mu\,|\nabla\varphi|^2 + (\lambda+\mu)\,|\operatorname{div}\varphi|^2)A(x-x^0)\cdot\nu\,d\sigma dt,$$

$$(3.16)\qquad (\varphi'\,,\,A\varphi)_{0,\Omega}\Big|_{t=0}^{T} = (\lambda+\mu)\int_Q \operatorname{div}\varphi\,\nabla\varphi : A\;dxdt.$$

Remark 4. For $N = 2$ or $N = 3$, the tensorial product $\nabla\varphi : A$ in (3.15) and (3.16) can be written in terms of the curl and **curl** operators, respectively. Indeed, let $\mathbf{e}_{ij} = \mathbf{e}_i \otimes \mathbf{e}_j = \mathbf{e}_j\,\mathbf{e}_i^t$, where $\{\mathbf{e}_i\}_{i=1}^N$ is the canonical basis in $\mathbb{R}^N$. If $N = 2$ and $A = \alpha(\mathbf{e}_{21} - \mathbf{e}_{12})$, where $\alpha \in \mathbb{R}$, then

$$\nabla\varphi : A = \alpha\left(\frac{\partial\varphi_2}{\partial x_1} - \frac{\partial\varphi_1}{\partial x_2}\right) = \alpha\operatorname{curl}\varphi.$$

In $N = 3$ and $A = \alpha_1(\mathbf{e}_{32}-\mathbf{e}_{23})+\alpha_2(\mathbf{e}_{13}-\mathbf{e}_{31})+\alpha_3(\mathbf{e}_{21}-\mathbf{e}_{12})$ where $\alpha = (\alpha_1,\alpha_2,\alpha_3) \in \mathbb{R}^3$, then

$$\nabla\varphi : A = \alpha_1\left(\frac{\partial\varphi_3}{\partial x_2} - \frac{\partial\varphi_2}{\partial x_3}\right) + \alpha_2\left(\frac{\partial\varphi_1}{\partial x_3} - \frac{\partial\varphi_3}{\partial x_1}\right) + \alpha_3\left(\frac{\partial\varphi_2}{\partial x_1} - \frac{\partial\varphi_1}{\partial x_2}\right) = \alpha\cdot\mathbf{curl}\,\varphi.$$

One can compare with the analogous properties of Corollary 4.12 in section 4.

3.4. Inverse inequality and exact controllability. Given the same notations as in section 2, if we introduce the subset $\Gamma(x^0,d,A)$ of Γ as in (2.5) and the quantities $r(x^0,d,A)$ and $R(x^0)$ as in (2.7), we obtain the following observability inequality and exact controllability result.

THEOREM 3.3 (inverse inequality). *Given $x^0 \in \mathbb{R}^N$, $d > 0$, and a skew-symmetric matrix A normalized as in* (2.6)*, if $\lambda_0^2 = \inf\{\|\nabla\varphi\|_{0,\Omega}^2 \,/\, \|\varphi\|_{0,\Omega}^2; \varphi \in H_0^1(\Omega)^N\}$ and if we define*

$$(3.17)\qquad T(x^0,d,A) = \frac{2}{\sqrt{\mu}}\left(R(x^0) + \frac{\|A\|_2}{\lambda_0}\right),$$

then for each $T > d^{-1}T(x^0, d, A)$ and for each weak solution φ of (3.3) *the following inequality holds:*

$$E_0 \le \frac{r(x^0,d,A)}{2(dT - T(x^0,d,A))} \int_0^T \int_{\Gamma(x^0,d,A)} (\mu\,|\nabla\varphi|^2 + (\lambda{+}\mu)\,|\operatorname{div}\varphi|^2)\,d\sigma dt, \tag{3.18}$$

where the initial energy E_0 was defined in (3.4).

THEOREM 3.4 (exact controllability). *Suppose that there exist $x^0 \in \mathbb{R}^N$, $d > 0$, and a skew-symmetric matrix A normalized as in* (2.6) *such that $\Gamma(x^0, d, A)$ is not empty and $\Gamma(x^0, d, A) \subset \Gamma_0$, then for each $T > d^{-1}T(x^0, d, A)$ there exists a control $v \in L^2(\Sigma_0)^N$ such that the corresponding solution of* (3.1) *satisfies the final time condition* (3.2).

Proof. Adding up the classical formulas (3.13) multiplied by b with (3.14) multiplied by $d(N-1)/2$ let us to obtain

$$\left(\varphi'\,,\,\nabla\varphi(x{-}x^0) + \frac{d(N-1)}{2}\varphi\right)_{0,\Omega}\Bigg|_{t=0}^{T} + \frac{d}{2}\int_Q (|\varphi'|^2 - \mu\,|\nabla\varphi|^2 - (\lambda{+}\mu)\,|\operatorname{div}\varphi|^2)\,dxdt$$
$$+ d\int_Q (\mu\,|\nabla\varphi|^2 + (\lambda{+}\mu)\,|\operatorname{div}\varphi|^2)\,dxdt = \frac{d}{2}\int_\Sigma (\mu\,|\nabla\varphi|^2 + (\lambda{+}\mu)\,|\operatorname{div}\varphi|^2)(x{-}x^0)\cdot\nu\,d\sigma dt.$$

Now, the new formula (3.16) replaced into the new identity (3.15) gives

$$\left(\varphi'\,,\,\nabla\varphi A(x{-}x^0) - A\varphi\right)_{0,\Omega}\Big|_{t=0}^{T} = \frac{1}{2}\int_\Sigma (\mu\,|\nabla\varphi|^2 + (\lambda{+}\mu)\,|\operatorname{div}\varphi|^2)A(x{-}x^0)\cdot\nu\,d\sigma dt.$$

By subtracting the last two identities we establish that

$$\begin{aligned}&(X_1(t) + X_2(t))\big|_{t=0}^{T} + dTE_0\\ &= \frac{1}{2}\int_\Sigma (\mu\,|\nabla\varphi|^2 + (\lambda{+}\mu)\,|\operatorname{div}\varphi|^2)(x - x^0)\cdot(dI + A)\nu\,d\sigma dt,\end{aligned} \tag{3.19}$$

where we have defined the quantities

$$X_1(t) = (\varphi'\,,\,A\varphi)_{0,\Omega} \quad\text{and}\quad X_2(t) = \left(\varphi'\,,\,\nabla\varphi(dI - A)(x{-}x^0) + d\frac{N-1}{2}\varphi\right)_{0,\Omega}.$$

We will prove the inverse inequality (3.18). On one hand, we deduce, from the Cauchy–Schwarz inequality, the inequality $ab \le a^2/(4\varepsilon) + \varepsilon b^2$ with $\varepsilon = \lambda_0\sqrt{\mu}/(2\,\|A\|_2)$ and from the definition of λ_0 and E_0 that

$$\left|X_1(t)\big|_{t=0}^{T}\right| \le 2\frac{\|A\|_2}{2\lambda_0\sqrt{\mu}}\,\|\varphi'\|_{0,\Omega}^2 + 2\frac{\lambda_0\sqrt{\mu}}{2\,\|A\|_2}\,\|A\|_2^2\,\frac{1}{\lambda_0^2}\,\|\nabla\varphi\|_{0,\Omega}^2 \le 2\frac{\|A\|_2}{\lambda_0\sqrt{\mu}}E_0. \tag{3.20}$$

Note that if $\|A\|_2 = 0$ then $X_1 = 0$. On the other hand, using the Cauchy–Schwarz inequality and the same inequality as before with $\varepsilon = \sqrt{\mu}/(2R(x^0))$ we obtain

$$\begin{aligned}|X_2(t)| \quad &\le \frac{R(x^0)}{2\sqrt{\mu}}\,\|\varphi'\|_{0,\Omega}^2 + \frac{\mu}{2R(x^0)\sqrt{\mu}}\left(\left\|\nabla\varphi(dI - A)(x{-}x^0)\right\|_{0,\Omega}^2\right.\\ &\left. + d^2\frac{(N-1)^2}{4}\,\|\varphi\|_{0,\Omega}^2 + d(N-1)\left(\nabla\varphi(dI - A)(x{-}x^0)\,,\,\varphi\right)_{0,\Omega}\right),\end{aligned} \tag{3.21}$$

where $R(x^0) = \max_{x\in\overline{\Omega}} \left|x - x^0\right| > 0$ was introduced in (2.7). We note that

$$\left(\nabla\varphi(dI - A)(x-x^0)\,,\,\varphi\right)_{0,\Omega} = -\frac{Nd}{2}\,\|\varphi\|^2_{0,\Omega}\,,$$

hence the last two terms in (3.21) have a negative sum $-d^2(N^2-1)\,\|\varphi\|^2_{0,\Omega}/4$. One also remarks that

$$\left\|\nabla\varphi(dI - A)(x-x^0)\right\|^2_{0,\Omega} \le (d^2 + \|A\|_2^2)R(x^0)^2\,\|\nabla\varphi\|^2_{0,\Omega} \le R(x^0)^2\,\|\nabla\varphi\|^2_{0,\Omega}\,,$$

since it has been assumed that $d^2 + \|A\|_2^2 = 1$. Finally we obtain

$$\left|X_2(t)|_{t=0}^T\right| \le 2\frac{R(x^0)}{2\sqrt{\mu}}\,\|\varphi'\|^2_{0,\Omega} + 2\frac{1}{2R(x^0)\sqrt{\mu}}R(x^0)^2\mu\,\|\nabla\varphi\|^2_{0,\Omega} \le 2\frac{R(x^0)}{\sqrt{\mu}}E_0. \tag{3.22}$$

The inequality (3.18) follows from (3.19), (3.20), (3.22), and definitions (2.5) and (2.7).

Theorem 3.4 follows immediately from Theorem 3.3 applying the HUM method (see [24, Chapter IV]). □

4. Stokes system.

4.1. Control problem. Let Ω be an open bounded subset of $\mathbb{R}^N$ ($N = 2$ or 3) with boundary Γ of class C^2. Let ν be the unit exterior normal to Ω. If $N = 2$ we refer to the tangent vector on Γ as $\tau = (-\nu_2, \nu_1)$.

DEFINITION 4.1. *For a vector field* $\mathbf{z}$ *defined on the boundary* Γ*, we introduce the operators* γ_n *and* γ_τ *defined by*

$$\gamma_n\mathbf{z} = \mathbf{z}\cdot\nu, \tag{4.1}$$

$$\gamma_\tau\mathbf{z} = \begin{cases} \mathbf{z}\cdot\tau & \text{if } N = 2,\\ \nu\times\mathbf{z} & \text{if } N = 3. \end{cases} \tag{4.2}$$

Remark 5. It should be noticed that on Γ

$$\mathbf{z} = \gamma_n(\mathbf{z})\nu + \gamma_\tau(\mathbf{z})\tau \quad \text{if} \quad N = 2, \tag{4.3a}$$

$$\mathbf{z} = \gamma_n(\mathbf{z})\nu + \gamma_\tau(\mathbf{z})\times\nu \quad \text{if} \quad N = 3. \tag{4.3b}$$

Therefore, γ_n corresponds to the normal trace for $N = 2$ and $N = 3$ and γ_τ corresponds to the tangential trace for $N = 2$. In case $N = 3$, $\gamma_\tau\times\nu$ corresponds to the tangential components.

Now, let us introduce the following classical functional spaces (see [43]) with their usual topologies:

$$H = \{\mathbf{v}\in L^2(\Omega)^N \mid \operatorname{div}\mathbf{v} = 0\ ,\gamma_n\mathbf{v} = 0 \text{ on } \Gamma\},\ V = \{\mathbf{v}\in H_0^1(\Omega)^N \mid \operatorname{div}\mathbf{v} = 0\}$$

in the standard embedding scheme $V'\subset H'\equiv H\subset V$.

The following control problem is considered. Let $T > 0$ and let Γ_0 be a subset of Γ with positive measure. Given $\mathbf{y}_0\in H$ and $\mathbf{f}\in L^2(0,T;V')$, for each control v scalar field if $N = 2$ or $\mathbf{v}$ vector field in $\mathbb{R}^3$ if $N = 3$, we consider (formally at the moment) the solution $(\mathbf{y}, p)$ of the following evolution Stokes problem:

$$\mathbf{y}' - \Delta\mathbf{y} + \nabla p = \mathbf{f} \quad \text{in} \quad \Omega\times(0,T), \tag{4.4a}$$

$$\operatorname{div} \mathbf{y} = 0 \quad \text{in} \quad \Omega \times (0,T), \tag{4.4b}$$

$$\gamma_n \mathbf{y} = 0 \quad \text{on} \quad \Gamma_0 \times (0,T), \tag{4.4c}$$

$$\gamma_\tau \mathbf{y} = \begin{cases} v & \text{if } N = 2 \\ \mathbf{v} & \text{if } N = 3 \end{cases} \quad \text{on} \quad \Gamma_0 \times (0,T), \tag{4.4d}$$

$$\mathbf{y} = 0 \quad \text{on} \quad (\Gamma \setminus \Gamma_0) \times (0,T), \tag{4.4e}$$

$$\mathbf{y}(0) = \mathbf{y}_0 \quad \text{in} \quad \Omega. \tag{4.4f}$$

Here $\mathbf{y}$ is the vector velocity field in $\mathbb{R}^N$ and p is the pressure, defined up to an additive constant. The symbol $'$ (prime) means derivation with respect to time and we use the usual notation $(\Delta \mathbf{y})_i = \partial^2 y_i / \partial x_j \partial x_j$ here (where repeated index means sum). In case $N = 3$, the control $\mathbf{v}$ must satisfy the following compatibility condition:

$$\gamma_n \mathbf{v} = 0 \quad \text{on} \quad \Gamma_0 \quad \text{if} \quad N = 3. \tag{4.5}$$

Our main goal is to find geometric conditions over Γ_0 in such a way that the space $\{\mathbf{y}(T)\}$ is dense in a suitable space when the control function v or $\mathbf{v}$ varies in a space also to be determined. In other words, we seek for conditions to have the tangential boundary approximate controllability.

Without loss of generality, for the approximate control problem, the initial datum $\mathbf{y}_0$ and $\mathbf{f}$ may be taken to be zero.

4.2. Tangential boundary approximate controllability. In Theorem 4.3 we obtain approximate controllability by using a rotated direction multiplier technique, with a geometric condition similar to that required in sections 2 and 3. In Theorem 4.4 we state the result for analytic boundaries and, in this particular case, the geometric condition of Theorem 4.3 is not necessary.

In order to describe the condition appearing in Theorem 4.3, we set down an analogous to Definition 2.1 in two and three dimensions including the case $d = 0$ and other signs of the inner product appearing in the main condition.

DEFINITION 4.2. *Let x^0 be a vector in $\mathbb{R}^N$, $d \geq 0$, and $\alpha \in \mathbb{R}$ when $N = 2$ or $\alpha \in \mathbb{R}^3$ when $N = 3$. We define the following subset of Γ:*

$$\Gamma^+(x^0, d, \alpha) = \begin{cases} \{x \in \Gamma \text{ such that } (x - x^0) \cdot (d\nu + \alpha\tau) > 0\} & \text{if } N = 2, \\ \{x \in \Gamma \text{ such that } (x - x^0) \cdot (d\nu + \alpha \times \nu) > 0\} & \text{if } N = 3, \end{cases} \tag{4.6}$$

and analogously we define $\Gamma^-(x^0, d, \alpha)$ and $\Gamma^0(x^0, d, \alpha)$ if the sign of the inner product in (4.6) is negative or zero, respectively.

In order to shorten notations let us define

$$\Sigma_0 = \Gamma_0 \times (0,T).$$

THEOREM 4.3 (approximate controllability). *We suppose that there exist $x^0 \in \mathbb{R}^N$, $d \geq 0$, and $\alpha \in \mathbb{R}$ when $N = 2$ or $\alpha \in \mathbb{R}^3$ when $N = 3$ such that Γ_0 satisfies the conditions*

$$\Gamma_0 \supseteq \Gamma^+(x^0, d, \alpha) \quad \text{if} \quad d > 0, \tag{4.7a}$$

$$\Gamma_0 \supseteq \Gamma^+(x^0, 0, \alpha) \cup \Gamma^0(x^0, 0, \alpha) \quad \text{if} \quad d = 0. \tag{4.7b}$$

For each $T > 0$ and for each $v \in L^2(\Sigma_0)$ when $N = 2$ or $\mathbf{v} \in L^2(\Sigma_0)^3$ satisfying (4.5) when $N = 3$, let us consider a solution $(\mathbf{y}, \mathbf{y}^T)$ of (4.4) in the weak sense of Definition 4.6. Then for all $T > 0$ the following sets are dense in V':

$$\begin{array}{ll} \{\mathbf{y}^T \text{ such that } v \in L^2(\Sigma_0)\} & \text{if} \quad N = 2, \\ \{\mathbf{y}^T \text{ such that } \mathbf{v} \in L^2(\Sigma_0)^3 \text{ satisfying (4.5)}\} & \text{if} \quad N = 3. \end{array} \tag{4.8}$$

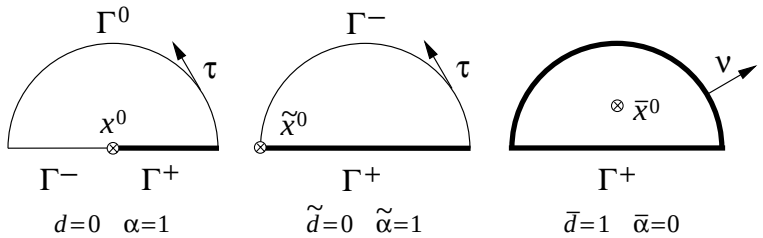


FIG. 4.1. *Example for Remark* 7 *in a half circle. Left: Take* $\Gamma_0 = \Gamma^+(x^0, 0, 1)$ *(bold line).* Γ_0 *can be extended to the whole diameter* $\Gamma_0 = \Gamma^+(x^0, 0, 1) \cup \Gamma^-(x^0, 0, 1)$. *Center: For a new* $\tilde{x}^0$, Γ_0 *can be extended to the whole boundary* $\Gamma_0 = \Gamma^+(\tilde{x}^0, 0, 1) \cup \Gamma^-(\tilde{x}^0, 0, 1)$. *Right: Another choice of parameters e.g.,* $\overline{x}^0$, $\overline{d} = 1$, $\overline{\alpha} = 0$ *leads to apply Theorem* 4.3 *to conclude the approximate controllability.*

Remark 6. With a boundary control in $L^2(\Sigma_0)$, we will give a sense to $\mathbf{y}(T)$ only in V'. More precisely, if $(\mathbf{y}, \mathbf{y}^T)$ is a weak solution of (4.4) in the sense of Definition 4.6, we will show that $\mathbf{y} \in C^0([0,T]; V')$.

Remark 7. If Γ_0 verifies $\Gamma_0 \supseteq \Gamma^+(x^0, 0, \alpha)$ for a choice of the parameters x^0, $d = 0$ and α, but the condition (4.7b) is not fulfilled, then it is showed that Γ_0 can be replaced by $\Gamma_0 \cup \Gamma^-(x^0, 0, \alpha)$ and Theorem 4.3 can be applied again with a new choice $\overline{x}^0$, $\overline{d}$, $\overline{\alpha}$ of the parameters (see the example provided by Figure 4.1).

The following result shows that condition (4.7) of Theorem 4.3 is not necessary for analytic boundaries if $N = 2$.

THEOREM 4.4 (approximate controllability for analytic boundaries). *Assume that* $N = 2$ *and* Ω *connected. Let* $\{\Gamma^i\}_{i=1}^K$ *be the connected components of* Γ *and define* $\Gamma_0^i = \Gamma_0 \cap \Gamma^i$. *If for each* $i = 1, \ldots, K$ *we have* $\Gamma_0^i = \Gamma^i$ *or* Γ^i *analytic and* Γ_0^i *an arbitrary nonempty open subset of* Γ^i, *then for each* $T > 0$ *the sets* (4.8) *are dense in* V'.

Remark 8. Condition (4.7) is not necessary in a generic sense: For all N and for a regular boundary Γ, if Γ_0 is an arbitrary nonempty subset of Γ, we could always slightly modify Γ_0 in order to have the result of Theorem 4.3 (see [36]).

Remark 9. There is no counterexample in order to decide if condition (4.7) is too restrictive, for instance, in the case of nonanalytic boundaries for $N = 2$.

4.3. A trace property. Here, we recall the definitions

$$(\nabla \mathbf{z})_{ij} = \frac{\partial z_i}{\partial x_j} \quad \text{and} \quad \operatorname{curl} \mathbf{z} = \frac{\partial z_2}{\partial x_1} - \frac{\partial z_1}{\partial x_2} \quad \text{if} \quad N = 2, \tag{4.9a}$$

$$\mathbf{curl}\, \mathbf{z} = \left(\frac{\partial z_3}{\partial x_2} - \frac{\partial z_2}{\partial x_3}, \frac{\partial z_1}{\partial x_3} - \frac{\partial z_3}{\partial x_1}, \frac{\partial z_2}{\partial x_1} - \frac{\partial z_1}{\partial x_2} \right) \quad \text{if} \quad N = 3. \tag{4.9b}$$

PROPOSITION 4.5. *Let* $\mathbf{z} \in H^2(\Omega)^N$ *with* $\mathbf{z} = \mathbf{c}$ *on* Γ_0, *where* $\mathbf{c}$ *is a constant vector of* $\mathbb{R}^N$ *and* Γ_0 *is a subset of* Γ *of positive measure. Then, on* Γ_0, *we have*

$$\gamma_n((\nabla \mathbf{z})\nu) = \operatorname{div} \mathbf{z}, \tag{4.10a}$$

$$\gamma_\tau((\nabla \mathbf{z})\nu) = \begin{cases} \operatorname{curl} \mathbf{z} & \text{if } N = 2, \\ \mathbf{curl}\, \mathbf{z} & \text{if } N = 3. \end{cases} \tag{4.10b}$$

Proof. We will prove the result for $\mathbf{z} \in C^1(\overline{\Omega})$. The condition $\mathbf{z} = \mathbf{c}$ on Γ_0 allows us to use Proposition 3.1, that is to say, that the derivative and normal indexes can

be permuted on Γ_0. Thanks to this property we easily see that

$$(\nabla \mathbf{z})\nu \cdot \nu = \frac{\partial z_i}{\partial x_j}\nu_j\nu_i = \frac{\partial z_i}{\partial x_i}\nu_j\nu_j = \operatorname{div} \mathbf{z}.$$

We recall the condensed formulas:

$$\operatorname{curl} \mathbf{z} = \varepsilon_{ij}\frac{\partial z_j}{\partial x_i} \text{ if } N = 2 \quad \text{and} \quad (\mathbf{curl}\, \mathbf{z})_i = \varepsilon_{ijk}\frac{\partial z_k}{\partial x_j} \text{ if } N = 3,$$

where ε_{ij} and ε_{ijk} are the signs of index permutations, i.e., if $\mathbf{e}_i$ is the ith canonical base vector of $\mathbb{R}^N$, we have $\varepsilon_{ij} = \det(\mathbf{e}_i\, \mathbf{e}_j)$ and $\varepsilon_{ijk} = \det(\mathbf{e}_i\, \mathbf{e}_j\, \mathbf{e}_k)$. In case $N = 2$, if we remember that $\tau_i = -\varepsilon_{ik}\nu_k$ and $\varepsilon_{ik} = -\varepsilon_{ki}$, then

$$(\nabla \mathbf{z})\nu \cdot \tau = -\frac{\partial z_i}{\partial x_j}\nu_j\varepsilon_{ik}\nu_k = \varepsilon_{ki}\frac{\partial z_i}{\partial x_k}\nu_j\nu_j = \operatorname{curl} \mathbf{z}.$$

For $N = 3$, if we use the fact that $(\mathbf{a} \times \mathbf{b})_i = \varepsilon_{ijk}a_j\, b_k$, we obtain

$$(\nu \times (\nabla \mathbf{z})\nu)_i = \varepsilon_{ijk}\nu_j\frac{\partial z_k}{\partial x_l}\nu_l = \varepsilon_{ijk}\frac{\partial z_k}{\partial x_j}\nu_l\nu_l = (\mathbf{curl}\, \mathbf{z})_i.$$

The general case $\mathbf{z} \in H^2(\Omega)$ can be deduced by a density argument from the regular case. □

4.4. Weak formulation of the nonsmooth data problem (4.4). Without loss of generality, we consider the case $\mathbf{f} = \mathbf{0}$ and $\mathbf{y}_0 = \mathbf{0}$. We can always do this by choosing new variables $\mathbf{y} - \overline{\mathbf{y}}$ and $p - \overline{p}$ where $(\overline{\mathbf{y}}, \overline{p})$ is the solution of (4.4) for $v = 0$ (or $\mathbf{v} = \mathbf{0}$). We recall that [43, Theorem 1.1, p. 254],

$$\overline{\mathbf{y}} \in L^2(0,T;V) \cap C([0,T];H),\ \overline{\mathbf{y}}' \in L^2(0,T;V'),$$
$$\overline{p} = \overline{P}' \text{ with } \overline{P} \in C([0,T];L^2(\Omega)).$$

DEFINITION 4.6. *For each $v \in L^2(\Sigma_0)$ if $N = 2$ or $\mathbf{v} \in L^2(\Sigma_0)^3$ verifying* (4.5) *if $N = 3$, we say that $(\mathbf{y}, \mathbf{y}^T)$ is a weak solution to Problem* (4.4) *if $\mathbf{y} \in L^2(0,T;H)$, $\mathbf{y}^T \in V'$, and*

$$\int_Q \mathbf{y} \cdot \mathbf{h}\, dxdt + \langle \mathbf{y}^T, \mathbf{z}^T \rangle_{V',V} = \begin{cases} -\int_{\Sigma_0} v \operatorname{curl} \mathbf{z}\, d\sigma dt & \text{if } N = 2, \\ -\int_{\Sigma_0} \mathbf{v} \cdot \mathbf{curl}\, \mathbf{z}\, d\sigma dt & \text{if } N = 3 \end{cases} \tag{4.11}$$

for each $\mathbf{h} \in L^2(0,T;H)$ and $\mathbf{z}^T \in V$, where $\mathbf{z}$ is the solution of

$$-\mathbf{z}' - \Delta \mathbf{z} + \nabla q = \mathbf{h} \quad \text{in} \quad \Omega \times (0,T), \tag{4.12a}$$
$$\operatorname{div} \mathbf{z} = 0 \quad \text{in} \quad \Omega \times (0,T), \tag{4.12b}$$
$$\mathbf{z} = \mathbf{0} \quad \text{on} \quad \Gamma \times (0,T), \tag{4.12c}$$
$$\mathbf{z}(T) = \mathbf{z}^T \quad \text{on} \quad \Omega. \tag{4.12d}$$

We know [43, Proposition 1.2, p. 267], that problem (4.12) has a unique solution

$$\mathbf{z} \in L^2(0,T;H^2(\Omega)^N \cap V) \text{ and } q \in L^2(0,T;H^1(\Omega)/\mathbb{R}),$$

with continuous dependence with respect to the data. From (4.12a) we also see that $\mathbf{z}' \in L^2(0,T;H)$ hence $\mathbf{z} \in C([0,T];V)$ (see [43, Chapter 1, Proposition 2.1] or [26, Chapter 3, Lemma 1.1]), and then (4.12d) is meaningful.

LEMMA 4.7. *For each $v \in L^2(\Sigma_0)$ when $N = 2$ or $\mathbf{v} \in L^2(\Sigma_0)^3$ verifying (4.5) when $N = 3$, there exists a unique solution $(\mathbf{y}, \mathbf{y}^T)$ of (4.11) – (4.12) in $L^2(0,T;H) \times V'$. Moreover, $\mathbf{y} \in C([0,T];V')$ and $\mathbf{y}(T) = \mathbf{y}^T$ in V'.*

Proof. The hardest part is to show that $\mathbf{y} \in C([0,T];V')$. Let $\mathcal{A}$ be the Stokes operator and let $D(\mathcal{A})$ be its domain,

$$D(\mathcal{A}) = \{\mathbf{v} \in V \text{ such that } \mathcal{A}\mathbf{v} \in H\} = H^2(\Omega)^N \cap V.$$

If $\mathbf{h} \in L^2(0,T;D(\mathcal{A}))$ and $\mathbf{z}^T = \mathbf{0}$, it can be shown that the solution of (4.12) satisfies $\mathbf{z}' \in L^2(0,T;D(\mathcal{A}))$ with a continuous dependence. Note that for each $\mathbf{h} \in \mathcal{D}((0,T);D(\mathcal{A}))$ we have in the sense of distributions

$$\langle \mathbf{y}', \mathbf{h} \rangle = -\langle \mathbf{y}, \mathbf{h}' \rangle = -\int_Q \mathbf{y} \cdot \mathbf{h}' \, dxdt,$$

and then using problem (4.11)–(4.12) with $\mathbf{z}^T = \mathbf{0}$, it follows that

$$|\langle \mathbf{y}', \mathbf{h} \rangle| = \left| \int_Q \mathbf{y} \cdot \mathbf{h}' \, dxdt \right| \leq \|v\|_{L^2(\Sigma_0)} \|\operatorname{curl} \mathbf{z}'\|_{L^2(\Sigma_0)} \leq C \|\mathbf{h}\|_{L^2(0,T;D(\mathcal{A}))},$$

and then by density $\mathbf{y}' \in L^2(0,T;D(\mathcal{A})')$. Since $\mathbf{y} \in L^2(0,T;H)$ we have $\mathbf{y} \in C([0,T];X)$ (see [26]) with

$$X = [H, D(\mathcal{A})']_{\frac{1}{2}} = [D(\mathcal{A}), H]'_{\frac{1}{2}} = V',$$

where $[X_1, X_2]_{\frac{1}{2}}$ denotes the interpolated space between X_1 and X_2 (see [26, Chapter 1, Proposition 2.1]). Once we know that $\mathbf{y} \in C([0,T];V')$ a density argument can be used to prove $\mathbf{y}(T) = \mathbf{y}^T$. Indeed, by taking $\mathbf{z} \in D((0,T);\mathcal{V})$, where $\mathcal{V} = \{\phi \in D(\Omega) \mid \operatorname{div} \phi = 0 \text{ in } \Omega\}$ in (4.11)–(4.12) we obtain the weak solution $\mathbf{y}$ satisfying (4.4a) in the sense of distributions. Then taking a test function $\mathbf{z} \in L^2(0,T;\mathcal{V})$ in this equation and observing that the Green formula

$$\int_0^T \langle \mathbf{y}', \mathbf{z} \rangle_{D(\mathcal{A})', D(\mathcal{A})} \, dt = -\int_0^T \langle \mathbf{y}, \mathbf{z}' \rangle_H \, dt + \langle \mathbf{y}(T), \mathbf{y}^T \rangle_{V',V}$$

is valid, after comparing with (4.11) we obtain that

$$\langle \mathbf{y}(T) - \mathbf{y}^T, \mathbf{z}^T \rangle_{V',V} = 0 \quad \forall \, \mathbf{z}^T \in \mathcal{V},$$

that is, for all $\mathbf{z}^T \in V$ and this implies that $\mathbf{y}(T) = \mathbf{y}^T$ in V'. ∎

LEMMA 4.8. *If v (or $\mathbf{v}$) is a regular function, then problem (4.4) is equivalent to problems (4.11)–(4.12).*

Proof. If we multiply (4.4) by the solution of (4.12) and if we integrate by parts, we obtain

$$\int_Q \mathbf{y} \cdot \mathbf{h} \, dxdt + \int_\Omega \mathbf{y}(T) \cdot \mathbf{z}^T \, dx = -\int_{\Sigma_0} (\nabla \mathbf{z})\nu \cdot \mathbf{y} \, d\sigma dt.$$

But we know that $\mathbf{z}$ satisfies (4.12c), thus it is constant on Γ_0. By using Proposition 4.5, conditions (4.12b), (4.4c), and (4.4d), we infer that on Σ_0

$$(\nabla \mathbf{z})\nu \cdot \mathbf{y} = \begin{cases} \gamma_n((\nabla \mathbf{z})\nu)\gamma_n \mathbf{y} + \gamma_\tau((\nabla \mathbf{z})\nu)\gamma_\tau \mathbf{y} = v \operatorname{curl} \mathbf{z} & \text{if } N = 2, \\ \gamma_n((\nabla \mathbf{z})\nu)\gamma_n \mathbf{y} + \gamma_\tau((\nabla \mathbf{z})\nu) \cdot \gamma_\tau \mathbf{y} = \mathbf{v} \cdot \mathbf{curl}\, \mathbf{z} & \text{if } N = 3. \end{cases}$$

Conversely, in the previous lemma we have proved that the solution of Problems (4.11)–(4.12) satisfies (4.4) except for (4.4d). But this can be easily shown in the regular case by comparison. □

4.5. Spectral decomposition. We want to prove that, under the hypothesis of Theorem 4.3 or Theorem 4.4, the set $\{\mathbf{y}^T\}$ of solutions to problems (4.11)–(4.12) is dense in V' as v (or $\mathbf{v}$) varies in the control set. Given $\mathbf{z}^T \in V$, we suppose that for each $v \in L^2(\Sigma_0)$ if $N = 2$ or $\mathbf{v} \in L^2(\Sigma_0)^3$ verifying (4.5) if $N = 3$, we have

$$\langle \mathbf{y}^T, \mathbf{z}^T \rangle_{V',V} = 0.$$

We will show that $\mathbf{z}^T = 0$. We take $\mathbf{h} = 0$ in Problem (4.11)–(4.12); therefore we have

$$\begin{cases} \int_{\Sigma_0} v \operatorname{curl} \mathbf{z}\, d\sigma dt = 0 & \text{if } N = 2, \\ \int_{\Sigma_0} \mathbf{v} \cdot \mathbf{curl}\, \mathbf{z}\, d\sigma dt = 0 & \text{if } N = 3 \end{cases}$$

for each $v \in L^2(\Sigma_0)$ when $N = 2$ or $\mathbf{v} \in L^2(\Sigma_0)^3$ verifying (4.5) when $N = 3$. Thus we have $\operatorname{curl} \mathbf{z} = 0$ on Σ_0 when $N = 2$. In case $N = 3$, Proposition 4.5 implies that $\mathbf{curl}\, \mathbf{z} \cdot \nu = 0$ on Σ_0 and then we also have $\mathbf{curl}\, \mathbf{z} = \mathbf{0}$ on Σ_0.

After reversing time and changing notations by $\mathbf{z}^0 = \mathbf{z}^T$, we see that in order to prove Theorem 4.3 or Theorem 4.4 we need to show the following unique continuation property: Let $(\mathbf{z}, q)$ be a solution of

$$\mathbf{z}' - \Delta \mathbf{z} + \nabla q = \mathbf{0} \quad \text{in} \quad \Omega \times (0,T), \tag{4.13a}$$

$$\operatorname{div} \mathbf{z} = 0 \quad \text{in} \quad \Omega \times (0,T), \tag{4.13b}$$

$$\mathbf{z} = \mathbf{0} \quad \text{on} \quad \Gamma \times (0,T), \tag{4.13c}$$

$$\mathbf{z}(0) = \mathbf{z}^0 \quad \text{in} \quad \Omega \tag{4.13d}$$

under the condition

$$\begin{cases} \operatorname{curl} \mathbf{z} = 0 & \text{if } N = 2 \\ \mathbf{curl}\, \mathbf{z} = \mathbf{0} & \text{if } N = 3 \end{cases} \quad \text{on} \quad \Sigma_0; \tag{4.14a}$$

then necessarily

$$\mathbf{z} = \mathbf{0} \ \text{ and } \ q = \textit{constant } (ct) \quad \text{in} \quad \Omega \times (0,T).$$

In order to study this property, we use an spectral decomposition method as in [27]. In this way, we first extend the solutions of (4.13) analytically for $t > 0$ and we introduce the spectrum of the Stokes operator ordered as

$$0 < \lambda_1 < \lambda_2 < \cdots \to \infty.$$

For each eigenvalue λ_i, $i \geq 1$, with multiplicity l_i, the associated eigenfunctions are designated by (φ_i^j, π_i^j), $j = 1, \ldots, l_i$, and they form an orthonormal basis. Thus we have

$$-\Delta \varphi_i^j + \nabla \pi_i^j = \lambda_i \varphi_i^j \quad \text{in} \quad \Omega, \tag{4.15a}$$

$$\operatorname{div} \varphi_i^j = 0 \quad \text{in} \quad \Omega, \tag{4.15b}$$

$$\varphi_i^j = \mathbf{0} \quad \text{on} \quad \Gamma. \tag{4.15c}$$

If we decompose $\mathbf{z}^0 = \sum_{i\geq 1}\sum_{j=1,\dots,l_i} a_i^j\varphi_i^j$, the solution of (4.13) can be written as follows: for every $t>0$:

$$\mathbf{z} = \sum_{i\geq 1}\sum_{j=1}^{l_i} a_i^j \exp(-\lambda_i t)\varphi_i^j.$$

Then, from condition (4.14), we have on Γ_0 for each $t>0$

$$\operatorname{curl}\mathbf{z} = 0 = \sum_{i\geq 1}\sum_{j=1}^{l_i} a_i^j \exp(-\lambda_i t)\operatorname{curl}\varphi_i^j.$$

Finally, instead of condition (4.14), from the strictly increasing ordering of the eigenvalues, we have the following condition for every $i\geq 1$:

$$\sum_{j=1}^{l_i} a_i^j \operatorname{curl}\varphi_i^j = 0 \quad \text{on} \quad \Sigma_0.$$

If now, for a fixed $i\geq 1$, we define

$$\varphi = \sum_{j=1}^{l_i} a_i^j\varphi_i^j \quad \text{and} \quad \nabla\pi = \sum_{j=1}^{l_i} a_i^j\nabla\pi_i^j,$$

we deduce that, for proving Theorem 4.3 or Theorem 4.4, we only have to show that the following unique continuation property on each frequency holds.

LEMMA 4.9. *Under the hypothesis of Theorem* 4.3 *or Theorem* 4.4, *if* (φ,π,λ) *is the solution of*

$$-\Delta\varphi + \nabla\pi = \lambda\varphi \quad \textit{in} \quad \Omega, \tag{4.16a}$$

$$\operatorname{div}\varphi = 0 \quad \textit{in} \quad \Omega, \tag{4.16b}$$

$$\varphi = \mathbf{0} \quad \textit{on} \quad \Gamma, \tag{4.16c}$$

with the additional condition that

$$\begin{cases} \operatorname{curl}\varphi = 0 & \textit{if } N=2 \\ \mathbf{curl}\,\varphi = \mathbf{0} & \textit{if } N=3 \end{cases} \quad \textit{on} \quad \Gamma_0, \tag{4.17}$$

then

$$\varphi = \mathbf{0} \quad \textit{and} \quad \pi = ct \quad \textit{in} \quad \Omega. \tag{4.18}$$

4.6. Proof of Lemma 4.9 under the hypothesis of Theorem 4.4. We take $\varphi = \mathbf{curl}\, w$. Then $-\Delta^2 w = \lambda\Delta w$ in Ω, $w=0$ on Γ^1, w constant on Γ^i, $i=2,\dots,K$, and $\frac{\partial w}{\partial\nu}=0$ on $\Gamma=\Gamma^1\cup\cdots\cup\Gamma^K$. From (4.16a) we also see that on Γ

$$\frac{\partial\pi}{\partial\nu} = \frac{\partial(\Delta w+\lambda w)}{\partial\tau} \quad \text{and} \quad \frac{\partial\pi}{\partial\tau} = -\frac{\partial(\Delta w+\lambda w)}{\partial\nu}. \tag{4.19}$$

Thanks to the assumed hypothesis we have $\Delta w = 0$ on Γ; then (4.19) implies that $\frac{\partial\pi}{\partial\nu}=0$ on Γ. Since $\Delta\pi=0$ in Ω then $\pi = ct$ in Ω. Using again (4.19) we obtain that $\frac{\partial\Delta w}{\partial\nu}=0$ on Γ and then

$$w = \frac{\partial w}{\partial\nu} = \Delta w = \frac{\partial\Delta w}{\partial\nu} = 0 \quad \text{on} \quad \Gamma^1,$$

and this implies that $w=0$, then $\varphi=\mathbf{0}$ by Holmgren's uniqueness property. This concludes the proof of Theorem 4.4.

4.7. Proof of Lemma 4.9 under the hypothesis of Theorem 4.3.

4.7.1. Step 1. Two multiplier formulas. In order to prove Lemma 4.9 under the hypothesis of Theorem 4.3, we have to introduce two new multiplier identities for the eigenvalue problem (4.16). We state the results in a more general manner than is required in this paper, since the results have an independent interest. We recall the notation $e : f = e_{ij} f_{ij}$ for e, f tensorial fields.

LEMMA 4.10. *Let (φ, π) and (ϕ, ρ) be solutions to the eigenvalue problem* (4.16) *for the same λ. Then for all $m \in W^{1,\infty}(\overline{\Omega})^N$ we have the following two formulas:*

$$\begin{aligned}(4.20)\quad \int_\Gamma (\nabla\varphi)\nu \cdot (\nabla\phi)\nu \ (m\cdot\nu)\, d\sigma &= \lambda \int_\Omega \varphi\cdot\phi \ \operatorname{div} m\, dx - \int_\Omega \nabla\varphi : \nabla\phi \ \operatorname{div} m\, dx \\ &+ \int_\Omega \nabla\varphi : \nabla\phi(\nabla m + \nabla m^t)\, dx - \int_\Omega \pi\, \nabla\phi : \nabla m^t\, dx - \int_\Omega \rho\, \nabla\varphi : \nabla m^t\, dx,\end{aligned}$$

$$\begin{aligned}(4.21)\quad &\int_\Omega \nabla\pi\cdot(\nabla m)^t\phi\, dx + \int_\Omega \nabla\rho\cdot(\nabla m)^t\varphi\, dx = \lambda\int_\Omega \varphi\cdot(\nabla m + \nabla m^t)\phi\, dx \\ &- \int_\Omega \nabla\varphi : (\nabla m + \nabla m^t)\nabla\phi\, dx + \int_\Omega \nabla\varphi : (\nabla\phi^t\nabla m + \nabla m\nabla\phi^t)\, dx.\end{aligned}$$

Proof. Briefly, we deduce the identities by using the multipliers $(\nabla\phi)m$ and $(\nabla\varphi)m$ in (4.16) in order to obtain (4.20) and the multipliers $(\nabla m)^t\phi$ and $(\nabla m)^t\varphi$ again in (4.16) to deduce (4.21). We now give the details. We multiply the pressure term in (4.16a) by $(\nabla\phi)m$ and we integrate the result by parts in Ω to obtain

$$\begin{aligned}(4.22)\quad \int_\Omega \nabla\pi\cdot(\nabla\phi)m\, dx &= \int_\Omega \frac{\partial\pi}{\partial x_i}\frac{\partial\phi_i}{\partial x_j} m_j\, dx \\ &= -\int_\Omega \pi\frac{\partial^2\phi_i}{\partial x_i\partial x_j} m_j\, dx - \int_\Omega \pi\frac{\partial\phi_i}{\partial x_j}\frac{\partial m_j}{\partial x_i}\, dx + \int_\Gamma \pi\frac{\partial\phi_i}{\partial x_j} m_j\,\nu_i\, d\sigma \\ &= -\int_\Omega \pi\nabla\phi : \nabla m^t\, dx + \int_\Gamma \pi\nu\cdot(\nabla\phi)m\, d\sigma = -\int_\Omega \pi\nabla\phi : \nabla m^t\, dx,\end{aligned}$$

since $\operatorname{div}\phi = 0$ and thanks to index change property (3.1)

$$(4.23)\qquad (\nabla\phi)m = \frac{\partial\phi_i}{\partial x_j} m_j = \frac{\partial\phi_i}{\partial x_j}\,\nu_k\,\nu_k\, m_j = \frac{\partial\phi_i}{\partial x_k}\,\nu_j\,\nu_k\, m_j = (\nabla\phi)\nu\,(m\cdot\nu)$$

and then from Proposition 4.5

$$(\nabla\phi)m\cdot\nu = (\nabla\phi)\nu\cdot\nu\,(m\cdot\nu) = \operatorname{div}\phi\,(m\cdot\nu) = 0.$$

Now, if we multiply the diffusion term in (4.16a) by $(\nabla\phi)m$ we have

$$\begin{aligned}(4.24)\quad \int_\Omega \Delta\varphi\cdot(\nabla\phi)m\, dx &= \int_\Omega \frac{\partial^2\varphi_i}{\partial x_j\partial x_j}\frac{\partial\phi_i}{\partial x_k} m_k\, dx \\ &= -\int_\Omega \frac{\partial\varphi_i}{\partial x_j}\frac{\partial^2\phi_i}{\partial x_k\partial x_j} m_k\, dx - \int_\Omega \frac{\partial\varphi_i}{\partial x_j}\frac{\partial\phi_i}{\partial x_k}\frac{\partial m_k}{\partial x_j}\, dx + \int_\Gamma \frac{\partial\varphi_i}{\partial x_j}\frac{\partial\phi_i}{\partial x_k} m_k\,\nu_j\, d\sigma \\ &= -\int_\Omega \nabla\varphi : (m\cdot\nabla)\nabla\phi\, dx - \int_\Omega \nabla\varphi : \nabla\phi\nabla m\, dx + \int_\Gamma (\nabla\varphi)\nu\cdot(\nabla\phi)m\, d\sigma,\end{aligned}$$

with the notation $(m\cdot\nabla)\nabla\phi \equiv m_k \frac{\partial}{\partial x_k}\nabla\phi$. To sum up, (4.16a) multiplied by $(\nabla\phi)m$ gives

$$\int_\Omega \nabla\varphi : (m\cdot\nabla)\nabla\phi\,dx + \int_\Omega \nabla\varphi : \nabla\phi\nabla m\,dx \tag{4.25}$$
$$-\int_\Gamma (\nabla\varphi)\nu\cdot(\nabla\phi)\nu\,(m\cdot\nu)\,d\sigma - \int_\Omega \pi\nabla\phi : \nabla m^t\,dx = \lambda\int_\Omega \varphi\cdot(\nabla\phi)m\,dx.$$

If we add the identity (4.25) to the same one obtained by interchanging the roles of φ and ϕ we obtain the identity (4.20). We have only to remark that

$$\nabla\phi : \nabla\varphi\nabla m = \nabla\varphi : \nabla\phi\nabla m^t$$

to observe that

$$\int_\Omega \varphi\cdot(\nabla\phi)m\,dx + \int_\Omega (\nabla\varphi)m\cdot\phi\,dx = -\int_\Omega \varphi\cdot\phi\ \operatorname{div} m\,dx,$$
$$\int_\Omega \nabla\varphi : (m\cdot\nabla)\nabla\phi\,dx + \int_\Omega (m\cdot\nabla)\nabla\varphi : \nabla\phi\,dx$$
$$= -\int_\Omega \nabla\varphi : \nabla\phi\ \operatorname{div} m\,dx + \int_\Gamma \nabla\varphi : \nabla\phi\,(m\cdot\nu)\,d\sigma$$

and to transform the last boundary integral by proving the following relation on Γ

$$\nabla\varphi : \nabla\phi = \frac{\partial\varphi_i}{\partial x_j}\frac{\partial\phi_i}{\partial x_j} = \frac{\partial\varphi_i}{\partial x_j}\,\nu_k\,\nu_k\,\frac{\partial\phi_i}{\partial x_j} = \frac{\partial\varphi_i}{\partial x_k}\,\nu_j\,\nu_k\,\frac{\partial\phi_i}{\partial x_j} = (\nabla\varphi)\nu\cdot(\nabla\phi)\nu,$$

valid thanks to the Proposition 3.1.

To prove (4.21), we notice that if we multiply the diffusion term in (4.16a) by $(\nabla m)^t\phi$ and if we integrate by parts, then we obtain

$$\int_\Omega \Delta\varphi\cdot(\nabla m)^t\phi\,dx = \int_\Omega \frac{\partial^2\varphi_i}{\partial x_k\partial x_k}\frac{\partial m_j}{\partial x_i}\phi_j\,dx \tag{4.26}$$
$$= -\int_\Omega \frac{\partial\varphi_i}{\partial x_k}\frac{\partial^2 m_j}{\partial x_k\partial x_i}\phi_j\,dx - \int_\Omega \frac{\partial\varphi_i}{\partial x_k}\frac{\partial m_j}{\partial x_i}\frac{\partial\phi_j}{\partial x_k}\,dx$$
$$= \int_\Omega \frac{\partial^2\varphi_i}{\partial x_k\partial x_i}\frac{\partial m_j}{\partial x_k}\phi_j\,dx + \int_\Omega \frac{\partial\varphi_i}{\partial x_k}\frac{\partial m_j}{\partial x_k}\frac{\partial\phi_j}{\partial x_i}\,dx - \int_\Omega \frac{\partial\varphi_i}{\partial x_k}\frac{\partial m_j}{\partial x_i}\frac{\partial\phi_j}{\partial x_k}\,dx$$
$$= \int_\Omega \nabla\varphi : \nabla\phi^t\nabla m\,dx - \int_\Omega \nabla\varphi : \nabla m^t\nabla\phi\,dx.$$

Therefore, the multiplier $(\nabla m)^t\phi$ in (4.16) gives

$$-\int_\Omega \nabla\varphi : \nabla\phi^t\nabla m\,dx + \int_\Omega \nabla\varphi : \nabla m^t\nabla\phi\,dx$$
$$+\int_\Omega \nabla\pi\cdot(\nabla m)^t\phi\,dx = \lambda\int_\Omega \varphi\cdot(\nabla m)^t\phi\,dx. \tag{4.27}$$

By interchanging the roles of φ and ϕ we obtain an analogous formula to (4.27). If we add up these two formulas, by using that (note the easy rules $F : GH = H : G^tF = G : FH^t$ and that $F : G = F^t : G^t$)

$$\nabla\varphi : \nabla\phi^t\nabla m = \nabla\phi : \nabla m\nabla\varphi^t,$$
$$\nabla\phi : \nabla m^t\nabla\varphi = \nabla\varphi : \nabla m\nabla\phi,$$

we obtain the identity (4.21). □

4.7.2. Step 2. A particular case in the formulas. A particular case of formulas (4.20) and (4.21) in which we are interested is the choice $m = B(x - x^0)$, where B is a constant matrix in $\mathbb{R}^{N\times N}$ and x^0 is a constant vector in $\mathbb{R}^N$.

COROLLARY 4.11. *Let (φ, π) and (ϕ, ρ) be solutions to the eigenvalue problem (4.16) for the same λ. For each matrix $B \in \mathbb{R}^{N\times N}$ and for each $x^0 \in \mathbb{R}^N$ we have*

$$\int_\Gamma (\nabla\varphi)\nu \cdot (\nabla\phi)\nu\,((x - x^0) \cdot B\nu)\,d\sigma = \int_\Omega \nabla\varphi : \nabla\phi(B + B^t)\,dx + \lambda \int_\Omega \varphi \cdot (B + B^t)\phi\,dx - \int_\Omega \nabla\varphi : (B + B^t)\nabla\phi\,dx, \tag{4.28}$$

$$\int_\Omega \pi\nabla\phi : B\,dx + \int_\Omega \rho\nabla\varphi : B\,dx = -\lambda \int_\Omega \varphi \cdot (B + B^t)\phi\,dx + \int_\Omega \nabla\varphi : (B + B^t)\nabla\phi\,dx. \tag{4.29}$$

Proof. We take $m = B(x - x^0)$ in (4.21). Since $\nabla m = B$ and

$$\int_\Omega \nabla\varphi : \nabla\phi^t B\,dx = \int_\Omega \frac{\partial\varphi_i}{\partial x_j}\frac{\partial\phi_k}{\partial x_i}b_{kj}\,dx = -\int_\Omega \frac{\partial\varphi_i}{\partial x_j}\phi_k\frac{\partial b_{kj}}{\partial x_i}\,dx = 0,$$
$$\int_\Omega \nabla\varphi : B\nabla\phi^t\,dx = \int_\Omega \frac{\partial\varphi_i}{\partial x_j}b_{ik}\frac{\partial\phi_j}{\partial x_k}\,dx = -\int_\Omega \varphi_i\frac{\partial b_{ik}}{\partial x_j}\frac{\partial\phi_j}{\partial x_k}\,dx = 0,$$

we directly obtain (4.29), after noticing that

$$\int_\Omega \nabla\pi \cdot B^t\phi\,dx + \int_\Omega \nabla\rho \cdot B^t\varphi\,dx = -\int_\Omega \pi\nabla\phi : B\,dx - \int_\Omega \rho\nabla\varphi : B\,dx. \tag{4.30}$$

On the other hand, if we multiply (4.16a) by ϕ and if we integrate by parts, we obtain

$$\int_\Omega \nabla\varphi : \nabla\phi\,dx = \lambda \int_\Omega \varphi \cdot \phi\,dx. \tag{4.31}$$

By taking $m = B^t(x - x^0)$ in (4.20) and using (4.30) and (4.31) we deduce (4.28). □

Remark 10. There are two cases where formulas (4.28) and (4.29) are simpler. The case $B = B^t$ (skew-symmetric) and the case $B = dI$, $d \in \mathbb{R}$, I the identity matrix in $\mathbb{R}^{N\times N}$. The case $B = I$ in (4.28) was also treated in [36] to study the generic simplicity of the Stokes' spectrum.

Now, we will restrict ourselves to the two and three dimensional cases. First we choose $B = A$ where A is a skew-symmetric matrix of the form

$$A = \begin{pmatrix} 0 & -\alpha \\ \alpha & 0 \end{pmatrix} \text{ if } N = 2 \quad \text{and} \quad A = \begin{pmatrix} 0 & -\alpha_3 & \alpha_2 \\ \alpha_3 & 0 & -\alpha_1 \\ -\alpha_2 & \alpha_1 & 0 \end{pmatrix} \text{ if } N = 3, \tag{4.32}$$

where $\alpha \in \mathbb{R}$ if $N = 2$ or $\alpha \in \mathbb{R}^3$ if $N = 3$.

From (4.29), we obtain the following orthogonality property.

COROLLARY 4.12. *Let* (φ,π) *and* (ϕ,ρ) *be solutions to the eigenvalue problem* (4.16) *for the same* λ. *Then*

$$\int_\Omega \pi \operatorname{curl}\phi\,dx + \int_\Omega \rho\operatorname{curl}\varphi\,dx = 0 \quad \textit{if} \quad N=2, \tag{4.33a}$$

$$\int_\Omega \pi\,\mathbf{curl}\,\phi\,dx + \int_\Omega \rho\,\mathbf{curl}\,\varphi\,dx = \mathbf{0} \quad \textit{if} \quad N=3. \tag{4.33b}$$

Now, we choose $B = dI + A$ in (4.28) with $d \geq 0$, I the identity matrix in $\mathbb{R}^{N\times N}$ and A always a skew-symmetric matrix in the form (4.32). The following result is obtained.

COROLLARY 4.13. *Let* (φ,π) *and* (ϕ,ρ) *solutions of* (4.16) *for the same* λ. *Then for each* $x^0\in\mathbb{R}^N$, $d\geq 0$, *and* $\alpha\in\mathbb{R}$ *if* $N=2$ *or* $\alpha\in\mathbb{R}^3$ *if* $N=3$ *we have*

$$\int_\Gamma \operatorname{curl}\varphi\operatorname{curl}\phi\,(x-x^0)\cdot(d\nu+\alpha\tau)\,d\sigma = 2d\int_\Omega \nabla\varphi:\nabla\phi\,dx \ \textit{if}\ N=2, \tag{4.34a}$$

$$\int_\Gamma \mathbf{curl}\,\varphi\cdot\mathbf{curl}\,\phi\,(x-x^0)\cdot(d\nu+\alpha\times\nu)\,d\sigma = 2d\int_\Omega \nabla\varphi:\nabla\phi\,dx \ \textit{if}\ N=3. \tag{4.34b}$$

4.7.3. Step 3. Splitting the boundary. By taking $\varphi=\phi$ in (4.34), we obtain

$$\int_\Gamma |\operatorname{curl}\varphi|^2\,(x-x^0)\cdot(d\nu+\alpha\tau)\,d\sigma = 2d\int_\Omega |\nabla\varphi|^2\,dx \quad \text{if} \quad N=2, \tag{4.35a}$$

$$\int_\Gamma |\mathbf{curl}\,\varphi|^2\,(x-x^0)\cdot(d\nu+\alpha\times\nu)\,d\sigma = 2d\int_\Omega |\nabla\varphi|^2\,dx \quad \text{if} \quad N=3. \tag{4.35b}$$

Now, we split Γ as follows:

$$\Gamma = \Gamma^+(x^0,d,\alpha)\cup\Gamma^0(x^0,d,\alpha)\cup\Gamma^-(x^0,d,\alpha). \tag{4.36}$$

Since Γ_0 satisfies the geometric condition (4.7a) of Theorem 4.3, then

$$\operatorname{curl}\mathbf{z}=0 \ \text{ or } \ \mathbf{curl}\,\mathbf{z}=\mathbf{0} \quad \text{on} \quad \Gamma^+(x^0,d,\alpha). \tag{4.37}$$

If we consider decomposition (4.36) and condition (4.37) in identity (4.35), it follows that $d\int_\Omega|\nabla\varphi|^2\,dx\leq 0$ and therefore we directly obtain $\varphi=0$ if $d>0$.

The case $d=0$ is more complicated. Using decomposition (4.36) and condition (4.37) in (4.35) with $d=0$ let us deduce only that

$$\operatorname{curl}\mathbf{z}=0 \ \text{ or } \ \mathbf{curl}\,\mathbf{z}=\mathbf{0} \quad \text{on} \quad \Gamma^-(x^0,d,\alpha). \tag{4.38}$$

Nevertheless, thanks to the geometric condition (4.7b) of Theorem 4.3, we obtain that $\operatorname{curl}\mathbf{z}=0$ or $\mathbf{curl}\,\mathbf{z}=\mathbf{0}$ on the whole Γ. Since $\Gamma\supseteq\Gamma^+(\overline{x}^0,1,0)$ for some $\overline{x}^0$ we deduce from the previous case ($d>0$) that $\varphi=0$ in all Ω. This concludes the proof of Theorem 4.3. □

Remark 11. Condition (4.37) implies (4.38) if $d=0$. This justifies Remark 7.

Acknowledgments. The author wishes to thank Prof. E. Zuazua for fruitful discussions and suggestions. The author gratefully acknowledges the Chilean and French Governments through its Scientific Committee Ecos-Conicyt.

REFERENCES

[1] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[2] C. Bardos, T. Masrour, and F. Tatout, *Condition nécessaire et suffisante pour la contrôlabilité exacte et la stabilisation du problème de l'élastodynamique*, C. R. Acad. Sci. Paris Sér. I Math., 320 (1995), pp. 1279–1281.

[3] N. Burq, *Contrôlabilité exacte des ondes dans des ouverts peu réguliers*, Asymptotic Anal., 14 (1997), pp. 157–191.

[4] N. Burq and P. Gérard, *Condition nécessaire et suffisante pour la contrôlabilité exacte des ondes*, C. R. Acad. Sci. Paris Sér. I Math., 325 (1997), pp. 749–752.

[5] G. Chen, *Control and stabilization for the wave equation in a bounded domain*, SIAM J. Control Optim., 17 (1979), pp. 66–81.

[6] G. Chen, *Control and stabilization for the wave equation in a bounded domain, part* II, SIAM J. Control Optim., 19 (1981), pp. 114–122.

[7] P. Grisvard, *Contrôlabilité exacte avec conditions mêlées*, C. R. Acad. Sci. Paris Sér. I Math., 305 (1987), pp. 363–366.

[8] P. Grisvard, *Contrôlabilité exacte dans les polygones et polyèdres*, C. R. Acad. Sci. Paris Sér. I Math., 304 (1987), pp. 367–370.

[9] P. Grisvard, *Boundary control of cracked domains*, in Proceedings of the Third International Workshop Conference of Evolution Equations, Control Theory, and Biomathematics, held at the Han-sur-Lesse Conference Center of the Belgian Ministry of Education, Basel, Lecture Notes in Pure and Appl. Math. 155, P. Clement and G. Luner, eds., 1993, Dekker, Basel, pp. 235–240.

[10] L. Ho, *Observabilité frontière de l'équation des ondes*, C. R. Acad. Sci. Paris Sér. I Math., 302 (1986), pp. 443–446.

[11] L. Hormander, *Uniqueness theorems and estimates for normally hyperbolic partial differential equations of the second order*, in Comptes Rendus 12, Congres Mathématiques Scandinaves, Lund, 1953, pp. 105–115.

[12] J. U. Kim, *Exact semi-internal control of an Euler–Bernoulli equation*, SIAM J. Control Optim., 30 (1992), pp. 1001–1023.

[13] J. U. Kim, *On the energy decay of a linear thermoelastic bar and plate*, SIAM J. Math. Anal., 23 (1992), pp. 889–899.

[14] K. Kime, *Control of matter waves in adjacent potential wells*, Math. Methods Appl. Sci., 20 (1997), pp. 369–381.

[15] V. Komornik, *Exact Controllability and Stabilization. The Multiplier Method*, RAM 36, Wiley, Chichester, Masson, Pairs, 1994.

[16] V. Komornik, *Stabilization frontière des équations de Maxwell*, C. R. Acad. Sci. Paris Sér. I Math., 318 (1994), pp. 535–540.

[17] J. Lagnese, *Boundary stabilization of linear elastodynamic systems*, SIAM J. Control Optim., 21 (1983), pp. 968–984.

[18] J. E. Lagnese, *Exact boundary controllability of Maxwell's equations in a general region*, SIAM J. Control Optim., 27 (1989), pp. 374–388.

[19] J. Lagnese, *Uniform asymptotic energy estimates for solutions of the equations of dynamic plane elasticity with nonlinear dissipation at the boundary*, Nonlinear Anal., 16 (1991), pp. 35–54.

[20] J. Lagnese, *Boundary controllability in problems of transmission for a class of second order hyperbolic systems*, ESAIM Control Optim. Calc. Var., 2 (1997), pp. 343–357.

[21] I. Lasiecka and R. Triggiani, *Uniform exponential energy decay of wave equations in a bounded region with* $L_2(0,\infty;L_2(\Gamma))$*-feedback control in the Dirichlet boundary conditions*, J. Differential Equations, 66 (1987), pp. 340–390.

[22] G. Lebeau, *Contrôle de l'équation de Schrödinger*, J. Math. Pures Appl., IX, 71 (1992), pp. 267–291.

[23] G. Leugering, *On boundary feedback stabilisability of a viscoelastic beam*, Proc. Roy. Soc. Edinburgh Sect. A, 114 (1990), pp. 57–69.

[24] J.-L. Lions, *Contrôlabilité exacte, perturbation et stabilisation de systèmes distribués*, vol. 1, Rech. Math. Appl. 8, Masson, Paris, 1988.
[25] J. L. Lions, *Exact controllability, stabilization and perturbations for distributed systems*, SIAM Rev., 30 (1988), pp. 1–68.
[26] J.-L. Lions and E. Magenes, *Problèmes aux Limites non Homogènes et Applications*, Vol. 1, Dunod, Paris, 1968.
[27] J.-L. Lions and E. Zuazua, *Approximate controllability of a hydro-elastic coupled system*, ESAIM Control Optim. Calc. Var., 1 (1995), pp. 1–15.
[28] J.-L. Lions and E. Zuazua, *A generic uniqueness result for the Stokes system and its control theoretical consequences*, in Partial Differential Equations and Applications, Collected Papers in Honor of Carlo Pucci on the Occasion of His 70th Birthday, P. Marcellini, G. Talenti, and E. Vesertini, eds., Lecture Notes in Pure and Appl. Math. 177, Marcel Dekker, New York, 1996, pp. 221–235.
[29] E. Machtyngier, *Exact controllability for the Schroedinger equation*, SIAM J. Control Optim., 32 (1994), pp. 24–34.
[30] C. Morawetz, J. Ralston, and W. Strauss, *Decay of solutions of the wave equation outside nontrapping obstacles*, Comm. Pure Appl. Math., 30 (1977), pp. 447–508.
[31] C. Morawetz, J. Ralston, and W. Strauss, *A correction to: Decay of solutions of the wave equation outside nontrapping obstacles*, Comm. Pure Appl. Math., 31 (1978), p. 795.
[32] J. Nečas, *Les Méthodes Directes en Théorie des Équations Elliptiques*, Masson, Paris, 1967.
[33] M. Niane and O. Seck, *Contrôlabilité exacte frontière de l'équation des ondes en présence de fissures par adjonction de contrôles internes au voisinage des sommets des fissures*, C. R. Acad. Sci. Paris Sér. I Math., 316 (1993), pp. 695–700.
[34] S. Nicaise, *Boundary exact controllability of interface problems with singularities* I*: Addition of the coefficients of singularities*, SIAM J. Control Optim., 34 (1996), pp. 1512–1532.
[35] J. Ortega, *Comportamiento asintótico, control y estabilización de algunos sistemas parabólicos y de placas*, Ph.D. thesis, Universidad Complutense de Madrid, Spain, 1997.
[36] J. Ortega and E. Zuazua, *Generic simplicity of the eigenvalues of the Stokes system in two space dimensions*, Adv. Differential Equations, 6 (2001), pp. 987–1023.
[37] A. Osses, *Une nouvelle famille des multiplicateurs et ses applications à la contrôlabilité exacte de l'équation d'ondes*, C. R. Acad. Sci. Paris Sér. I Math., 326 (1998), pp. 1099–1104.
[38] A. Osses and J.-P. Puel, *Approximate controllability of a linear model in solid-fluid interaction*, ESAIM Control Optim., Calc. Var., 4 (1999), pp. 497–513.
[39] K. Phung, *Contrôlabilité exacte et stabilization interne des équations de Maxwell*, C. R. Acad. Sci. Paris Ser. I Math, 323 (1996), pp. 169–174.
[40] B. Rao, *Contrôlabilité exacte frontière d'un système hybride en élasticité*, C. R. Acad. Sci. Paris Sér. I Math., 324 (1997), pp. 889–894.
[41] F. Rellich, *Darstellung der eigenwerte von $\delta u + \lambda u$ durch ein randintegral*, Math. Z., 46 (1940), pp. 635–646.
[42] L. Rosier, *Exact boundary controllability for the Korteweg-de Vries equation on a bounded domain*, ESAIM Control Optim. Calc. Var., 2 (1997), pp. 32–55.
[43] R. Temam, *Navier–Stokes Equations*, North–Holland, Amsterdam, 1977.
[44] E. Zuazua, *Exponential decay for the semilinear wave equation with localized damping in unbounded domains*, J. Math. Pures Appl., 70 (1991), pp. 513–529.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 3, pp. 801–823

STOCHASTIC LINEAR-QUADRATIC CONTROL VIA SEMIDEFINITE PROGRAMMING*

DAVID D. YAO†, SHUZHONG ZHANG‡, AND XUN YU ZHOU‡

Abstract. We study stochastic linear-quadratic (LQ) optimal control problems over an infinite time horizon, allowing the cost matrices to be indefinite. We develop a systematic approach based on semidefinite programming (SDP). A central issue is the stability of the feedback control; and we show this can be effectively examined through the complementary duality of the SDP. Furthermore, we establish several implication relations among the SDP complementary duality, the (generalized) Riccati equation, and the optimality of the LQ control problem. Based on these relations, we propose a numerical procedure that provides a thorough treatment of the LQ control problem via primal-dual SDP: it identifies a stabilizing feedback control that is optimal or determines that the problem possesses no optimal solution. For the latter case, we develop an ϵ-approximation scheme that is asymptotically optimal.

Key words. stochastic LQ control, semidefinite programming, complementary duality, mean-square stability, generalized Riccati equation

AMS subject classifications. 93E20, 90C25, 93D15

PII. S0363012999355484

1. Introduction.

Consider the following stochastic linear-quadratic (LQ) optimal control problem:

$$\text{(LQ)} \qquad \min \quad J(x_0, u(\cdot)) := \mathsf{E}\int_0^{+\infty} [x(t)^T Q x(t) + u(t)^T R u(t)]dt \tag{1.1}$$

$$\text{s.t. (subject to)} \quad dx(t) = [Ax(t) + Bu(t)]dt + [Cx(t) + Du(t)]dW(t) \tag{1.2}$$
$$x(0) = x_0 \in \Re^n.$$

Here and throughout the paper, A, B, C, D and Q, R are constant matrices, with Q and R being symmetric matrices; the superscript T denotes the transpose of matrices and vectors; $W(\cdot)$ is a one-dimensional standard Brownian motion (with $t \in [0, +\infty)$ and $W(0) = 0$), defined on a filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$; and $u(\cdot)$ denotes the (open-loop) control, which belongs to $L^2_{\mathcal{F}}(\Re^m)$, the space of all $\Re^m$-valued, $\mathcal{F}_t$-adapted measurable processes satisfying

$$\mathsf{E}\int_0^{+\infty} \|u(t)\|^2 dt < +\infty.$$

Note that allowing multidimensional Brownian motion will render no additional difficulty to the results below. Also note that the dynamics in (1.2) involve multiplicative

*Received by the editors April 30, 1999; accepted for publication (in revised form) January 4, 2001; published electronically September 28, 2001.

http://www.siam.org/journals/sicon/40-3/35548.html

†Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027 (yao@ieor.columbia.edu). This research was undertaken while the author was on leave at the Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong. This research was supported in part by NSF grant ECS-97-05392, RGC earmarked grant CUHK 4175/00E, and a direct grant from CUHK.

‡Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, Hong Kong (zhang@se.cuhk.edu.hk, xyzhou@se.cuhk.edu.hk). The research of the second author was supported by RGC earmarked grant CUHK 4175/00E. The research of the third author was supported by RGC earmarked grants CUHK 4125/97E and CUHK 4175/00E.

noise in both the state and the control. A motivating application is given below in section 2.1, where the control of a portfolio affects not only the average return of the portfolio but also its volatility.

Studies on LQ control problems have a long history that can be traced back to the pioneering works of Kalman [12] and Wonham [22] (also refer to [5, 6, 9] and the references therein). A primary tool in solving LQ control problems is the following (stochastic algebraic) Riccati equation, with the symmetric matrix P being the unknown:

$$A^TP+PA+Q+C^TPC-(PB+C^TPD)(R+D^TPD)^{-1}(B^TP+D^TPC)=0. \tag{1.3}$$

Suppose P^* is a solution to the above equation with $R+D^TP^*D \succ 0$ (positive definite). Then, following the classical LQ theory, we know that

$$u^*(t)=-(R+D^TP^*D)^{-1}(B^TP^*+D^TP^*C)x^*(t), \tag{1.4}$$

assuming that it is mean-square stabilizing, is an optimal state feedback control for (LQ).

The limitation of the classical theory, however, lies in the difficulty involved in solving the Riccati equation (1.3), in particular since the matrix inverse term also involves the unknown P. In fact, there is no guarantee for $R+D^TP^*D \succ 0$, except when $Q \succeq 0$ (positive semidefinite) and $R \succ 0$, in which case $P^* \succeq 0$ follows as a solution to (1.3); see [1, Corollary 5.1].

In the deterministic case, $R \succeq 0$ is in fact necessary for the LQ problem to be well-posed (cf. [25, Chapter 6, Proposition 2.4]), whereas the Riccati approach further requires the nonsingularity of R. The positive definiteness of R has been a starting point of most of the stochastic LQ literature; see [5, 6, 9, 22] and the references therein. Recent studies (e.g., [8]) have, however, made a case for studying LQ problems in which R is singular or even indefinite in the stochastic case, with relevant applications ranging from portfolio selection to pollution control [8, 13, 26]. In particular, a singular or indefinite R may naturally arise in a class of problems in which the control affects the diffusion part of the system dynamics (i.e., $D \neq 0$ in (1.2)).

A traditional method for solving the Riccati equation (in the case of $D=0$) is to consider the so-called associated Hamiltonian matrix [7]. In this case it is known that the Riccati equation has a solution if and only if the associated Hamiltonian matrix admits no pure imaginary eigenvalues, a condition that can be verified using a Routh–Hurwitz-type test on a set of polynomial inequalities involving the given matrices. If the associated Hamiltonian matrix passes this test, then the solution to the Riccati equation can be constructed using the eigenvectors of the associated Hamiltonian matrix. This procedure, however, does not apply when R is indefinite, as the matrix inverse in (1.3), and hence the Riccati equation and the associated Hamiltonian matrix themselves, may not be well defined.

To overcome this difficulty, our idea here is to use semidefinite programming (SDP) as a unifying approach to solve the stochastic LQ control problem, generally in the absence of the positive definiteness/semidefiniteness of the cost matrices R and Q. SDP is a newly developed tool in optimization (see [16, 3]). It relates intimately to the so-called linear matrix inequalities (LMI) (see [23]). Pioneered by Yakubovich [24] and Willems [21], a vast literature has since appeared, applying the LMI approach to both deterministic and stochastic systems; refer to [7] for a systematic exposition and detailed literature review. However, the definiteness of R has remained a predominant

assumption in the research literature. In a recent work, [1], the relationship between the Riccati equation in (1.3) and the associated LMI has been examined under the assumption that $R + D^T P D \succ 0$ for some solution P of the Riccati equation (while R itself is allowed to be indefinite). This assumption, however, is hard to verify a priori as P is unknown; and even when this assumption is not satisfied, the corresponding LQ problem may still possess a meaningful optimal control; see Example 6.1 below.

In contrast, our focus here is to develop a direct connection between the LQ control problem and the duality theory of SDP. In particular, we demonstrate that to extend beyond the confines of the classical LQ theory (with positive definite cost matrices) the central issue is *stability*; and stability is intimately related to the *complementary duality* of the SDP associated with the LQ problem. We establish several equivalence relations between the stability/optimality of the LQ problem and the duality of the SDP, and demonstrate that a new class of optimal controls can be constructed based on the dual SDP. Furthermore, exploiting the primal-dual structure of the SDP also leads to powerful and efficient computational means, based on the newly developed primal-dual interior-point technologies (refer to, e.g., [18]), to solving the LQ problem. In short, while the LMI approach is a primal-only method, our primal-dual SDP approach applies to a more general class of problems, generates new theoretical results, and leads to practical computational algorithms.

Briefly, the rest of the paper is organized as follows. In section 2, we start with an application example that motivates the general LQ problem formulated above, introduce several regularity conditions relating to stability, and present the preliminaries of SDP. The main results of the paper are presented in the next two sections: We establish first in section 3, the connection between stability and the dual SDP; and then in section 4, we establish several implication relations among the optimality of LQ control problem, the complementary duality of the SDP, and a generalized version of the Riccati equation (1.3) involving a matrix pseudoinverse. A synthesis of these results is presented in section 5, along with a computational procedure that provides a systematic treatment of the LQ control problem via SDP. Several examples are collected in section 6 to illustrate some of the key technical issues involved in the SDP approach. For problems that do not possess an attainable optimal control, an ϵ-approximation scheme is developed in section 7, which achieves asymptotic optimality. Brief concluding remarks are summarized in section 8.

2. Preliminaries.

2.1. An application example. To motivate the general LQ problem formulated in the last section, consider a market where there is one bond and one stock, with price dynamics governed, respectively, by

$$dP_0(t) = rP_0(t)dt, \quad P_0(0) = p_0,$$

and

$$dP_1(t) = P_1(t)\big[bdt + \sigma dW(t)\big], \quad P_1(0) = p_1,$$

where $W(\cdot)$ denotes the one-dimensional standard Brownian motion. Suppose an agent, with an initial endowment z_0, wants to track a (stochastic) wealth trajectory (e.g., that of an index fund) $I(t)$ determined by the following equation:

$$dI(t) = I(t)[b_I dt + \sigma_I(t)dW(t)], \quad I(0) = i_0.$$

At any time $t \geq 0$ the total wealth of the agent is denoted by $z(t)$, of which the market value of the stock is denoted by $\pi(t)$. If we assume that the stock is traded continuously, and that there is no transaction cost, dividend payment, and withdrawal for consumption, then $z(t)$ must satisfy the following (see, e.g., [25, Chapter 2, section 3.2]):

$$dz(t) = [rz(t) + (b-r)\pi(t)]dt + \sigma\pi(t)dW(t), \quad z(0) = z_0.$$

The objective of the agent is to choose $\pi(\cdot)$ so as to minimize the following objective:

$$J(z(0), \pi(\cdot)) = \mathsf{E}\int_0^{+\infty} e^{-\rho t}|z(t) - I(t)|^2 dt,$$

where ρ is the discount rate.

To transform the above into the formulation in (1.1) and (1.2), define the state and control variables as follows:

$$(x(t), y(t)) = e^{-\frac{1}{2}\rho t}(z(t), I(t)), \qquad u(t) = e^{-\frac{1}{2}\rho t}\pi(t).$$

Then, the state dynamics become

$$\begin{aligned} dx(t) &= \left[\left(r - \frac{1}{2}\rho\right)x(t) + (b-r)u(t)\right]dt + \sigma u(t)dW(t), \quad x(0) = z_0, \\ dy(t) &= \left(b_I - \frac{1}{2}\rho\right)y(t) + \sigma_I y(t)dW(t), \quad y(0) = i_0, \end{aligned}$$

and the objective function can be rewritten as follows:

$$\begin{aligned} J(x(0), y(0), u(\cdot)) &= \mathsf{E}\int_0^{+\infty} |x(t) - y(t)|^2 dt \\ &= \mathsf{E}\int_0^{+\infty} \begin{bmatrix} x(t), & y(t) \end{bmatrix} \begin{bmatrix} 1, & -1 \\ -1, & 1 \end{bmatrix} \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} dt. \end{aligned}$$

This way, we are in the framework of (1.1) and (1.2), with the control cost $R \equiv 0$, and the state cost $Q = \begin{bmatrix} 1, & -1 \\ -1, & 1 \end{bmatrix}$ being singular.

2.2. Regularity conditions.

DEFINITION 2.1.

(i) *An open-loop control $u(\cdot)$ is called (mean-square) stabilizing at x_0, if the corresponding state $x(\cdot)$ of (1.2) with the initial state x_0 satisfies* $\lim_{t\to+\infty} \mathsf{E}[x(t)^T x(t)] = 0$.

(ii) *A feedback control $u(t) = Kx(t)$, where K is a constant matrix, is called stabilizing if for every initial state x_0, the solution to the equation*

$$\begin{cases} dx(t) = (A + BK)x(t)dt + (C + DK)x(t)dW(t), \\ x(0) = x_0, \end{cases}$$

satisfies $\lim_{t\to+\infty} \mathsf{E}[x(t)^T x(t)] = 0$.

(iii) *The system in (1.2) is called (mean-square) stabilizable if there exists a stabilizing feedback control of the form $u(t) = Kx(t)$ where K is a constant matrix.*

DEFINITION 2.2.

(i) *An open-loop control* $u(\cdot) \in L^2_{\mathcal{F}}(\Re^m)$ *is called admissible (at* x_0*) if it is stabilizing at* x_0*. The set of all admissible controls at* x_0 *is denoted as* $U_{ad}^{x_0}$*.*

(ii) *An admissible pair* $(x^*(\cdot), u^*(\cdot))$ *is called* optimal *(at* x_0*) if* $u^*(\cdot)$ *achieves the infimum of* $J(x_0, u(\cdot))$ *over* $u(\cdot) \in U_{ad}^{x_0}$*.*

(iii) *The control problem* (LQ) *is called well-posed (at* x_0*) if*

$$-\infty < \inf_{u(\cdot)\in U_{ad}^{x_0}} J(x_0, u(\cdot)) < +\infty;$$

(LQ) *is called attainable (at* x_0*) if it is well-posed (at* x_0*) and there exists an optimal admissible control.*

Unless explicitly stated otherwise, we shall assume throughout the paper that the system under consideration, (1.2), is mean-square stabilizable. Note that this is a very mild regularity condition; in particular, it is implied by the well-posedness of (LQ) when $Q \succ 0$ and $R \succeq 0$. Indeed, the well-posedness of (LQ) yields that there is at least one control whose cost is finite. As a result, under that control, $\lim_{t\to+\infty} \mathsf{E}[x(t)^T Q x(t)] = 0$. Thus the control must be stabilizing due to the nonsingularity of Q.

On the other hand, to appreciate why the admissible controls have to be stabilizing, consider the case when $Q \succ 0$ and $R \succeq 0$. In order for the cost objective in (1.1) to be finite it is necessary (as shown above) that the corresponding control must be stabilizing. In general, a nonstabilizing control is ill behaved and hence should be excluded.

Attainability of an optimal admissible control is another issue, even when the problem is well-posed. Consider the following simple (deterministic) LQ problem:

$$\begin{array}{ll} \min & \displaystyle\int_0^\infty x(t)^2 dt \\ \text{s.t.} & dx(t) = [-x(t) + u(t)]dt, \\ & x(0) = 1. \end{array}$$

Clearly, this problem has an infimum value zero. However, there is no control that attains the zero cost. In general, deciding whether or not a problem has an attainable optimal control is as hard as solving the (LQ) problem, especially for large problems.

2.3. Semidefinite programming. SDP is a special form of the so-called *conic optimization problem*, which is in essence to optimize a linear function over the intersection of two closed convex sets: one being an affine subspace and the other a cone. In SDP, the cone is formed by positive semidefinite matrices in the linear space of symmetric matrices. Similar to linear programming, an SDP problem can be cast in various ways. In standard form, an SDP is the following convex optimization problem:

$$\begin{array}{lll} (SDP)_p & \min & \langle C, X\rangle \\ & \text{s.t.} & \langle A_i, X\rangle = b_i \text{ for } i = 1, \ldots, m, \\ & & X \succeq 0, \end{array}$$

where C and A_i are $n \times n$ symmetric matrices, $b_i \in \Re^m$ is a vector, $i = 1, \ldots, m$, and $\langle X, Y\rangle := \sum_{i,j} X_{ij} Y_{ij}$ denotes the matrix inner-product.

The above has an associated dual, which is also an SDP problem:

$$\begin{array}{lll} (SDP)_d & \max & b^T y \\ & \text{s.t.} & \sum_{i=1}^m y_i A_i + Z = C, \\ & & Z \succeq 0. \end{array}$$

A conic optimization problem is said to satisfy the *Slater condition* if its feasible region has a nonempty intersection with the interior of the cone. In our standard primal-dual SDP problems stated above, the Slater condition has the following characterization. For the primal problem $(SDP)_p$, the Slater condition is equivalent to the existence of a primal feasible solution X^0 such that $X^0 \succ 0$. Similarly, for the dual problem $(SDP)_d$, the Slater condition is the existence of a dual feasible solution (y^0, Z^0) with $Z^0 \succ 0$.

For conic optimization problems, a well-developed duality theory exists; see e.g., [23, 14, 16]. Key points of the theory can be highlighted as follows:

- The *weak duality* always holds, i.e., any feasible solution to the primal (minimization) problem always possesses an objective value that is no less than the (dual) objective value of any dual feasible solution (the dual being a maximization problem).
- In contrast, the *strong duality*—that the optimal values of the primal and dual problems coincide—holds if there exists a pair of *complementary optimal solutions* X^* and (y^*, Z^*), namely, they satisfy $X^*Z^* = 0$. If, furthermore, it holds that $X^* + Z^* \succ 0$, then this pair of optimal solutions is called *strictly complementary*.
- Unlike linear programming, SDP may fail to satisfy the strong duality, let alone strict complementarity. It is known, however, that if the primal problem is feasible and the dual satisfies the Slater condition, then the primal problem must have a nonempty and compact optimal solution set. Moreover, if both the primal and the dual satisfy the Slater condition, then both must have nonempty and compact optimal solution sets, and the strong duality holds.

In the SDP literature, the Slater type regularity conditions are mostly assumed in order to avoid pathological cases. In Luo, Sturm, and Zhang [14, 15], extensive analysis can be found addressing the issue of regularity in the context of duality theory, and the related issue of how to detect the duality status numerically.

The linkage between the LQ control problem and the SDP is best understood in the deterministic setting. Consider the deterministic version of (LQ) as follows:

$$\begin{aligned} \min & \int_0^\infty [x(t)^T Q x(t) + u(t)^T R u(t)] dt \\ \text{s.t. } & \dot{x}(t) = Ax(t) + Bu(t), \\ & x(0) = x_0 \in \Re^n. \end{aligned}$$

Assume $Q \succeq 0$ and $R \succ 0$. Then, the optimal solution to the above problem is

$$u^*(t) = -R^{-1}B^T P^* x^*(t),$$

where P^* is a nonnegative definite solution to the Riccati equation

$$Q + A^T P + PA - PBR^{-1}B^T P = 0.$$

It turns out that solutions to the Riccati equation can be found through solving the following SDP:

$$\begin{aligned} \max & \langle I, P \rangle \\ \text{s.t. } & \begin{bmatrix} R, & B^T P \\ PB, & Q + A^T P + PA \end{bmatrix} \succeq 0, \\ & P \in \mathcal{S}^{n \times n}, \end{aligned}$$

where $\mathcal{S}^{n\times n}$ denotes the space of n by n symmetric matrices. Note that the above is an SDP problem because the constraint set can be viewed as an intersection between a linear subspace and the cone of positive semidefinite matrices. The dual problem is

$$\begin{aligned}&\min \langle R, Z_b\rangle + \langle Q, Z_n\rangle\\ &\text{s.t. } I + Z_u^T B^T + BZ_u + Z_n A^T + AZ_n = 0,\\ &\qquad Z := \begin{bmatrix} Z_b, & Z_u \\ Z_u^T, & Z_n \end{bmatrix} \succeq 0.\end{aligned}$$

It is interesting to note that both the primal and dual SDPs above are well defined even in the presence of the singularity of R and Q. In particular, there is no matrix inverse involved. Similarly, the primal and dual SDPs for the stochastic problem in (LQ) can be written as follows:

$$\begin{aligned}\text{(P)}\quad &\max \langle I, P\rangle\\ &\text{s.t.}\quad \mathcal{L}(P) \succeq 0,\\ &\qquad P \in \mathcal{S}^{n\times n},\end{aligned}$$

where

$$\mathcal{L}(P) := \begin{bmatrix} R + D^T PD, & B^T P + D^T PC \\ PB + C^T PD, & Q + C^T PC + A^T P + PA \end{bmatrix}; \tag{2.1}$$

and

$$\begin{aligned}\text{(D)}\quad &\min \langle R, Z_B\rangle + \langle Q, Z_N\rangle\\ &\text{s.t.}\quad I + Z_U^T B^T + BZ_U + Z_N A^T + AZ_N\\ &\qquad +CZ_N C^T + DZ_U C^T + CZ_U^T D^T + DZ_B D^T = 0,\\ &\qquad Z := \begin{bmatrix} Z_B, & Z_U \\ Z_U^T, & Z_N \end{bmatrix} \succeq 0.\end{aligned}$$

In the above forms, (P) and (D) are said to satisfy the Slater condition, if there exist primal and dual feasible solutions, P^0 and Z^0, such that $\mathcal{L}(P^0) \succ 0$ and $Z^0 \succ 0$, respectively. On the other hand, a primal optimal solution P^* and a dual optimal solution Z^* are called complementary optimal solutions if $\mathcal{L}(P^*)Z^* = 0$. Furthermore, they are called strictly complementary if $\mathcal{L}(P^*) + Z^* \succ 0$.

As a least condition in order to apply the SDP approach, we assume throughout the paper that the feasible set of (P) is nonempty. This assumption is satisfied automatically if the Riccati equation (1.3) has a solution P^* with $R + D^T P^* D \succ 0$ (which is the key assumption in [1]) by virtue of the well-known Schur lemma. It is also satisfied when $Q \succeq 0$ and $R \succeq 0$. Therefore, without this assumption, the original LQ problem cannot be solved by either the Riccati or the SDP approach.

3. Stability. Since we require admissible controls to be stabilizing, we need first to address the issue of stability, which, as it will become evident below, relates closely to the dual SDP.

The following results from [1, Theorems 2.1, 5.2] will be used later.

PROPOSITION 3.1. *The following conditions are equivalent.*

(i) *System* (1.2) *is mean-square stabilizable.*

(ii) *Problem (D) satisfies the Slater condition.*

(iii) *There exists a matrix K and a symmetric matrix Y such that*

$$(A+BK)Y+Y(A+BK)^T+(C+DK)Y(C+DK)^T \prec 0, Y \succ 0. \tag{3.1}$$

In this case the feedback $u(t)=Kx(t)$ is a stabilizing control.

(iv) *There exists a matrix K such that for any X there exists a unique solution Y to the following equation:*

$$(A+BK)Y+Y(A+BK)^T+(C+DK)Y(C+DK)^T+X=0. \tag{3.2}$$

Moreover, if $X \succ 0$ (resp., $X \succeq 0$) then $Y \succ 0$ (resp., $Y \succeq 0$). Furthermore, in this case the feedback $u(t)=Kx(t)$ is a stabilizing control.

(v) *There exist a matrix X and a positive definite matrix $Y \succ 0$ such that*

$$\begin{bmatrix} AY+YA^T+BX+X^TB^T & CY+DX \\ YC^T+X^TD^T & -Y \end{bmatrix} \prec 0. \tag{3.3}$$

In this case, the feedback $u(t)=XY^{-1}x(t)$ is a stabilizing control.

Note that the last equivalent condition above is an LMI condition, based on which the mean-square stabilizability can be verified numerically (cf. [7, 10]). Moreover, to check whether or not a given feedback control $u(t)=Kx(t)$ is stabilizing, it suffices to check if the LMIs in (3.1) have a feasible solution, which again can be carried out numerically.

Define

$$F(P) := A^TP+PA+Q+C^TPC-(PB+C^TPD)(R+D^TPD)^+(B^TP+D^TPC). \tag{3.4}$$

Here, M^+ stands for the pseudoinverse of a matrix M (refer to [17]). Note that when M is a positive semidefinite matrix, M^+ satisfies the following properties:

$$M^+ \succeq 0, \quad (M^+)^T = M^+, \quad M^+M = MM^+;$$

$$MM^+M = M, \qquad M^+MM^+ = M^+.$$

Clearly, the equation $F(P)=0$ generalizes the classical Riccati equation (1.3). Hence, we shall refer to it below as the *generalized Riccati equation.*

The following extended Schur's lemma [2] plays an important technical role.

LEMMA 3.2. *Let matrices $M=M^T, N$ and $S=S^T$ be given with appropriate dimensions. Then the following three conditions are equivalent:*

(i) $M-NS^+N^T \succeq 0$, $S \succeq 0$, *and* $N(I-SS^+)=0$.

(ii) $\begin{bmatrix} M & N \\ N^T & S \end{bmatrix} \succeq 0$.

(iii) $\begin{bmatrix} S & N^T \\ N & M \end{bmatrix} \succeq 0$.

THEOREM 3.3. *If a feasible solution of* (P), *P^*, is such that $F(P^*)=0$, and the feedback control*

$$u(t) = -(R+D^TP^*D)^+(B^TP^*+D^TP^*C)x(t) \tag{3.5}$$

is stabilizing, then there exist complementary optimal solutions of (P) *and* (D). *In particular, P^* is optimal to* (P); *and there exists a complementary dual optimal solution Z^*, such that $Z_N^* \succ 0$.*

Proof. Denote $K := -(R + D^T P^* D)^+ (B^T P^* + D^T P^* C)$. By the stability assumption of the control (3.5) and Proposition 3.1(iv), the equation

$$(A + BK)Y + Y(A + BK)^T + (C + DK)Y(C + DK)^T + I = 0$$

has a positive solution. Let it be $Y^* \succ 0$. Furthermore, let

$$Z_N^* = Y^*,\ Z_U^* = KZ_N^*, \text{ and } Z_B^* = K(Z_U^*)^T. \tag{3.6}$$

By this construction, we can easily verify that

$$\begin{bmatrix} Z_B^*, & Z_U^* \\ (Z_U^*)^T, & Z_N^* \end{bmatrix} = \begin{bmatrix} I, & K \\ 0, & I \end{bmatrix} \begin{bmatrix} 0, & 0 \\ 0, & Z_N^* \end{bmatrix} \begin{bmatrix} I, & 0 \\ K^T, & I \end{bmatrix} \succeq 0.$$

Moreover, by direct verification, we know

$$I + (Z_U^*)^T B^T + BZ_U^* + Z_N^* A^T + AZ_N^* + CZ_N C^T + DZ_U^* C^T + C(Z_U^*)^T D^T + DZ_B^* D^T = 0.$$

Therefore, Z^* is a feasible solution of (D). Moreover, using extended Schur's lemma (Lemma 3.2), we have

$$\begin{aligned}
&\mathcal{L}(P^*) \begin{bmatrix} Z_B^*, & Z_U^* \\ (Z_U^*)^T, & Z_N^* \end{bmatrix} \\
&= \begin{bmatrix} I, & 0 \\ -K^T, & I \end{bmatrix} \begin{bmatrix} R + D^T P^* D, & 0 \\ 0, & F(P^*) \end{bmatrix} \begin{bmatrix} I, & -K \\ 0, & I \end{bmatrix} \begin{bmatrix} Z_B^*, & Z_U^* \\ (Z_U^*)^T, & Z_N^* \end{bmatrix} \\
&= \begin{bmatrix} I, & 0 \\ -K^T, & I \end{bmatrix} \begin{bmatrix} (R + D^T P^* D)(Z_B^* - K(Z_U^*)^T), & R(Z_U^* - KZ_N^*) \\ F(P^*)(Z_U^*)^T, & F(P^*)Z_N^* \end{bmatrix} \\
&= \begin{bmatrix} 0, & 0 \\ 0, & 0 \end{bmatrix},
\end{aligned}$$

where in the first equation the decomposition of $\mathcal{L}(P^*)$ into the product of three matrices (the Schur decomposition) is possible because $\mathcal{L}(P^*) \succeq 0$, since P^* is a feasible solution to (P); hence, Lemma 3.2 can be invoked.

Therefore, the above establishes that P^* and Z^* are complementary solutions; in particular, P^* is optimal to (P), and Z^* is optimal to (D). The last statement of the theorem follows from the fact that $Z_N^* = Y^* \succ 0$. □

Note that the assumption in the above theorem, that the control in (3.5) is stabilizing, is not automatically satisfied even in the case when $R + D^T P^* D \succ 0$; see Example 6.2. The following result shows, however, that we can possibly obtain a stabilizing feedback control via the *dual* SDP.

THEOREM 3.4. *Let* $Z = \begin{bmatrix} Z_B, & Z_U \\ Z_U^T, & Z_N \end{bmatrix}$ *be a feasible solution of* (D) *with* $Z_N \succ 0$, *then the feedback control* $u(t) = Z_U Z_N^{-1} x(t)$ *is stabilizing.*

Proof. First of all, by feasibility of Z to (D) along with Schur's lemma we have

$$Z_B \succeq Z_U Z_N^{-1} Z_U^T.$$

Then,

$$\begin{aligned}
0 = \ & I + Z_U^T B^T + BZ_U + Z_N A^T + AZ_N + CZ_N C^T + DZ_U C^T + CZ_U^T D^T + DZ_B D^T \\
\succeq \ & I + Z_U^T B^T + BZ_U + Z_N A^T + AZ_N + CZ_N C^T \\
+ \ & DZ_U C^T + CZ_U^T D^T + DZ_U Z_N^{-1} Z_U^T D^T \\
\succ \ & Z_U^T B^T + BZ_U + Z_N A^T + AZ_N + (CZ_N + DZ_U) Z_N^{-1} (Z_N C^T + Z_U^T D^T).
\end{aligned}$$

Applying Schur's lemma again, we conclude that Proposition 3.1(v) holds with $X = Z_U$ and $Y = Z_N \succ 0$. Hence $u(t) = Z_U Z_N^{-1} x(t)$ is stabilizing. □

4. Optimality. Here we establish the relationship among the optimality of the original LQ problem, the primal/dual SDP problems, and the generalized Riccati equation.

THEOREM 4.1. *If* (LQ) *is attainable at any* $x_0 \in \Re^n$, *then* (P) *must have an optimal solution* P^* *satisfying* $F(P^*) = 0$.

Proof. First, note that since (LQ) has a finite optimal value with respect to any initial value, it must have a quadratic representation

$$\inf_{u(\cdot)\in U_{ad}^{x_0}} J(x_0, u(\cdot)) = x_0^T M x_0 \ \ \forall x_0 \in \Re^n;$$

see [4, p. 21]. (Note that the proof there, which is for the deterministic case, readily extends to the stochastic case.)

Suppose for the time being that the matrix M is a feasible solution of (P), the validity of which will be proved later. Let $u^*(\cdot)$ be the optimal control and $x^*(\cdot)$ be the corresponding state from the initial x_0. Then for any feasible solution P of (P), applying Itô's formula (see, e.g., [25, p. 36]) yields

$$\begin{aligned}
&d(x^*(t)^T P x^*(t)) \\
&= \Big[(Ax^*(t) + Bu^*(t))^T P x^*(t) + x^*(t)^T P(Ax^*(t) + Bu^*(t)) \\
&\quad +(Cx^*(t) + Du^*(t))^T P(Cx^*(t) + Du^*(t))\Big] dt + \{\ldots\} dW(t) \\
&= \big[x^*(t)^T (A^T P + PA + C^T PC) x^*(t) \\
&\quad +2u^*(t)^T (B^T P + D^T PC) x^*(t) + u^*(t)^T D^T PD u^*(t)\ \big] dt \\
&\quad +\{\ldots\} dW(t).
\end{aligned}$$

If we integrate the above over $[0, \infty)$, take expectations, and use that $\mathsf{E}[x^*(t)^T P x^*(t)] \to 0$ (as $u^*(\cdot)$ is stabilizing), we obtain

$$\begin{aligned}
0 = \quad & x_0^T P x_0 + \mathsf{E} \int_0^\infty [x^*(t)^T (A^T P + PA + C^T PC) x^*(t) \\
& +2u^*(t)^T (B^T P + D^T PC) x^*(t) \\
& +u^*(t)^T D^T PD u^*(t)] dt.
\end{aligned} \tag{4.1}$$

Completion of square yields

$$\begin{aligned}
&J(x_0, u^*(\cdot)) \\
&= \mathsf{E} \int_0^\infty [x^*(t)^T Q x^*(t) + u^*(t)^T R u^*(t)] dt \\
&= x_0^T P x_0 + \\
&\mathsf{E} \int_0^\infty \left\{ [u^*(t) - Kx^*(t)]^T (R + D^T PD)[u^*(t) - Kx^*(t)] + x^*(t)^T F(P) x^*(t) \right\} dt,
\end{aligned} \tag{4.2}$$

where $K := -(R + D^T PD)^+ (B^T P + D^T PC)$. Since P is feasible to (P), we have $R + D^T PD \succeq 0$ and $F(P) \succeq 0$ by extended Schur's lemma. This means that

$$x_0^T M x_0 \equiv J(x_0, u^*(\cdot)) \geq x_0^T P x_0 \tag{4.3}$$

for any P feasible to (P). Hence M must be optimal to (P). On the other hand, taking $P = M$ in (4.2) and noting $J(x_0, u^*(\cdot)) = x_0^T M x_0$, we conclude that $x(t)^T F(M) x(t) =$

0 for all $t \in [0, \infty)$. Since x_0 can be chosen arbitrarily it follows that $F(M) = 0$. The desired result thus follows.

What remains is to show the primal feasibility of M. To this end we consider a perturbation on the problem (LQ). That is, we keep all the data A, B, C, and D unchanged, and let $R_\epsilon = R + \epsilon I$ and $Q_\epsilon = Q + \epsilon I$ where $\epsilon > 0$ is a small positive number.

Under the perturbation ($\epsilon > 0$), the corresponding SDPs,

$$\begin{aligned}(P_\epsilon)\quad & \max \langle I, P\rangle \\ & \text{s.t.} \begin{bmatrix} R+\epsilon I + D^T P D, & B^T P + D^T P C \\ PB + C^T P D, & Q + \epsilon I + C^T P C + A^T P + PA \end{bmatrix} \succeq 0, \\ & P \in \mathcal{S}^{n\times n},\end{aligned}$$

and

$$\begin{aligned}(D_\epsilon)\quad & \min \langle R+\epsilon I, Z_B\rangle + \langle Q + \epsilon I, Z_N\rangle \\ & \text{s.t. } I + Z_U^T B^T + B Z_U + Z_N A^T + A Z_N \\ & \qquad + C Z_N C^T + D Z_U C^T + C Z_U^T D^T + D Z_B D^T = 0, \\ & \begin{bmatrix} Z_B, & Z_U \\ Z_U^T, & Z_N \end{bmatrix} \succeq 0,\end{aligned}$$

both satisfy the Slater condition (the former does because the feasible set of (P) is assumed to be nonempty, and the latter because of the mean-square stabilizability assumption and Proposition 3.1(ii)), and therefore complementary optimal solutions exist [23]. Observe that the feasible set of (D_ϵ) is independent of ϵ. Take any dual feasible solution Z^0. By weak duality, we have

$$\text{tr } P = \langle I, P\rangle \le \langle R + \epsilon I, Z_B^0\rangle + \langle Q + \epsilon I, Z_N^0\rangle. \tag{4.4}$$

Let $\hat{P}$ be a feasible solution of (P), which exists by our assumption. Certainly, $\hat{P}$ is feasible to (P_ϵ) for all $\epsilon \ge 0$. Theorem 5.5 in [1] asserts that for $\epsilon > 0$, the unique optimal solution for (P_ϵ), denoted by P_ϵ^*, dominates any other feasible solutions. Hence, $\hat{P} \preceq P_\epsilon^*$ for all $\epsilon > 0$.

This, together with (4.4), implies in particular that P_ϵ^* are contained in a compact set, with $0 \le \epsilon \le \epsilon_0$ ($\epsilon_0 > 0$ is a predetermined constant.) Now, take a convergent subsequence such that

$$\lim_{i\to\infty} P_{\epsilon_i}^* = P_0^*$$

with $\epsilon_i \to 0$ as $i \to \infty$.

Clearly, P_0^* is a feasible solution of (P) since the feasible region of (P_ϵ) monotonically shrinks as $\epsilon \downarrow 0$. We now show that $P_0^* = M$. Define the perturbed cost function

$$J_\epsilon(x_0, u(\cdot)) = \mathsf{E}\int_0^\infty [x(t)^T Q_\epsilon x(t) + u(t)^T R_\epsilon u(t)]dt.$$

Similar to (4.2), we can show

$$\begin{aligned}J_\epsilon(x_0, u(\cdot)) = x_0^T P_\epsilon^* x_0 + \mathsf{E}\int_0^\infty \big\{&[u(t) - K_\epsilon x(t)]^T (R_\epsilon + D^T P_\epsilon^* D)[u(t) - K_\epsilon x(t)] \\ &+ x(t)^T F_\epsilon(P_\epsilon^*) x(t)\big\} dt\end{aligned}$$

for any $u(\cdot) \in U_{ad}^{x_0}$, where $K_\epsilon := -(R_\epsilon + D^T P_\epsilon^* D)^+ (B^T P_\epsilon^* + D^T P_\epsilon C)$, and F_ϵ is the "perturbed" Riccati operator with Q and R in (3.4) replaced by Q_ϵ and R_ϵ, respectively. In [1, Theorems 5.4 and 5.5] it is stated that if the problem (P_ϵ) is strictly feasible (which is the case here) with an optimal solution P_ϵ^*, then the corresponding feedback control $u(t) = K_\epsilon x(t)$ must be stabilizing (and hence admissible), and P_ϵ^* must be the unique optimal solution to (P_ϵ) satisfying the corresponding Riccati equation $F_\epsilon(P) = 0$. Thus $u(t) = K_\epsilon x(t)$ must be optimal, and

$$\inf_{u(\cdot)\in U_{ad}^{x_0}} J_\epsilon(x_0, u(\cdot)) = x_0^T P_\epsilon^* x_0,$$

which further yields

$$x_0^T P_{\epsilon_i}^* x_0 = \inf_{u(\cdot)\in U_{ad}^{x_0}} J_{\epsilon_i}(x_0, u(\cdot)) \geq \inf_{u(\cdot)\in U_{ad}^{x_0}} J(x_0, u(\cdot)) = x_0^T M x_0.$$

Letting $\epsilon_i \to 0$, we obtain

$$x_0^T P_0^* x_0 \geq x_0^T M x_0.$$

On the other hand, (4.3) gives rise to the opposite inequality since P_0^* is feasible to (P). Thus we have $M = P_0^*$. This establishes that M is indeed a primal feasible solution. □

An important implication, which is the contrapositive of the above theorem, is this: If (P) has no optimal solution, or if (P) has optimal solutions but none of them satisfies the generalized Riccati equation $F(P) = 0$, then (LQ) has no attainable optimal control; in particular, it does not have any optimal feedback control.

A natural question to ask at this point is whether or not the converse of the above statement is true. Namely, if (P) admits an optimal solution P^* satisfying $F(P^*) = 0$, then does (LQ) have an attainable optimal control? Recall that in the finite horizon case an optimal feedback control is represented as $u^*(t) = -(R+D^T P^* D)^{-1}(B^T P^* + D^T P^* C)x^*(t)$ (cf. [8, Theorem 3.2]). However in the present case $R + D^T P^* D$ may be singular, therefore we naturally expect that an optimal feedback control should be

$$u^*(t) = -(R + D^T P^* D)^+ (B^T P^* + D^T P^* C)x^*(t). \tag{4.5}$$

The following result establishes that this control is indeed optimal if it is stabilizing. (Recall that in the infinite horizon case, stability is essential.)

THEOREM 4.2. *If a feasible solution of* (P)*, P^*, is such that $F(P^*) = 0$, and the feedback control $u^*(t)$ in* (4.5) *is stabilizing, then it must be optimal for* (LQ).

Proof. For any admissible control $u(\cdot) \in U_{ad}^{x_0}$, an argument similar to that in proving (4.2) leads to

$$\begin{aligned}
&J(x_0, u(\cdot)) \\
&= \mathsf{E}\int_0^\infty [x(t)^T Q x(t) + u(t)^T R u(t)]dt \\
&= x_0^T P^* x_0 \\
&\quad + \mathsf{E}\int_0^\infty \left\{[u(t) - Kx(t)]^T (R + D^T P^* D)[u(t) - Kx(t)] + x(t)^T F(P^*) x(t)\right\} dt \\
&= x_0^T P^* x_0 \\
&\quad + \mathsf{E}\int_0^\infty [u(t) - Kx(t)]^T (R + D^T P^* D)[u(t) - Kx(t)]dt,
\end{aligned} \tag{4.6}$$

where $K := -(R + D^T P^* D)^+ (B^T P^* + D^T P^* C)$. Since $u^*(t) = Kx^*(t)$ is stabilizing, the above shows that $u^*(\cdot)$ must be optimal. □

As mentioned earlier, the stability of the control in (4.5) can be examined numerically via the LMIs as stipulated in Proposition 3.1(iii). On the other hand, by Theorem 3.3, in order for this control to be stabilizing, it is *necessary* that there exist complementary solutions P^* and Z^* of (P) and (D), respectively, with $Z_N^* \succ 0$. It is interesting that under these (weaker) conditions, one can prove the existence of an explicitly representable optimal feedback control of (LQ) in (4.9) below (which is not necessarily in the same form as the control in (4.5)).

First we need a lemma.

LEMMA 4.3. *Suppose* (P) *and* (D) *have complementary optimal solutions,* P^* *and* Z^**, respectively. Then,* $F(P^*) = 0$.

Proof. We have the following decomposition:

$$\mathcal{L}(P^*) = \begin{bmatrix} I, & 0 \\ -K^T, & I \end{bmatrix} \begin{bmatrix} R + D^T P^* D, & 0 \\ 0, & F(P^*) \end{bmatrix} \begin{bmatrix} I, & -K \\ 0, & I \end{bmatrix}, \tag{4.7}$$

where $K := -(R + D^T P^* D)^+ (B^T P^* + D^T P^* C)$. From the relation $\mathcal{L}(P^*) Z^* = 0$, it follows that

$$\begin{aligned} &\begin{bmatrix} R + D^T P^* D, & 0 \\ 0, & F(P^*) \end{bmatrix} \begin{bmatrix} I, & -K \\ 0, & I \end{bmatrix} \begin{bmatrix} Z_B^*, & Z_U^* \\ (Z_U^*)^T, & Z_N^* \end{bmatrix} \\ &= \begin{bmatrix} (R + D^T P^* D)(Z_B^* - K(Z_U^*)^T), & (R + D^T P^* D)(Z_U^* - K Z_N^*) \\ F(P^*)(Z_U^*)^T, & F(P^*) Z_N^* \end{bmatrix} \\ &= \begin{bmatrix} 0, & 0 \\ 0, & 0 \end{bmatrix}. \end{aligned} \tag{4.8}$$

Therefore,

$$F(P^*)(Z_U^*)^T = 0, \qquad F(P^*) Z_N^* = 0$$

and

$$Z_U^* F(P^*) = 0, \qquad Z_N^* F(P^*) = 0.$$

On the other hand, the dual nonnegativity constraint, together with the extended Schur's lemma, yields

$$Z_B^* - Z_U^* (Z_N^*)^+ (Z_U^*)^T \succeq 0,$$

and

$$Z_U^* (I - Z_N^* (Z_N^*)^+) = 0.$$

Multiplying $F(P^*)$ on both sides of the dual equality constraint and making use of the above relations, we obtain

$$\begin{aligned} 0 &= F(P^*)(I + C Z_N^* C^T + D Z_U^* C^T + C^T Z_U^* D + D Z_B^* D^T) F(P^*) \\ &\succeq F(P^*)(I + C Z_N^* C^T + D Z_U^* C^T + C^T Z_U^* D + D Z_U^* (Z_N^*)^+ (Z_U^*)^T D^T) F(P^*) \\ &= F(P^*)^2 + F(P^*)[C Z_N^* + D Z_U^*](Z_N^*)^+ [Z_N^* C^T + (Z_U^*)^T D^T] F(P^*) \\ &\succeq F(P^*)^2. \end{aligned}$$

This means, $F(P^*) = 0$. □

THEOREM 4.4. *Assume that solving* (P) *and* (D) *yields complementary optimal solutions* P^* *and* Z^*, *with* $Z_N^* \succ 0$. *Then* $F(P^*) = 0$, *and* (LQ) *has an attainable optimal feedback control given by*

$$u^*(t) = Z_U^*(Z_N^*)^{-1}x^*(t). \tag{4.9}$$

Proof. First, that $F(P^*) = 0$ is seen from Lemma 4.3 (even without the assumption $Z_N^* \succ 0$).

Next, for any feasible solution P of (P) and any (stabilizing) control $u(\cdot) \in U_{ad}^{x_0}$, along with the corresponding state $x(\cdot)$, an argument similar to the one that proved (4.2) leads to

$$\begin{aligned} & J(x_0, u(\cdot)) \\ &= \mathsf{E}\int_0^\infty [x(t)^T Q x(t) + u(t)^T R u(t)]dt \\ &= x_0^T P x_0 \\ &+\mathsf{E}\int_0^\infty \left\{[u(t) - Kx(t)]^T (R + D^T P D)[u(t) - Kx(t)] + x(t)^T F(P) x(t)\right\} dt, \end{aligned} \tag{4.10}$$

where $K := -(R + D^T P D)^+ (B^T P + D^T P C)$. Since P is feasible to (P), we have $F(P) \succeq 0$. This means that

$$J(x_0, u(\cdot)) \geq x_0^T P x_0 \ \forall u(\cdot) \in U_{ad}^{x_0}$$

for any P feasible to (P).

Note that as yet we cannot conclude that the control in (4.5) is optimal, since we do not know whether or not the control is stabilizing, whereas (4.10) holds only for stabilizing controls. To get around, let us define a feedback control $u^*(t) = Z_U^*(Z_N^*)^{-1}x^*(t)$, which, following Theorem 3.4, is stabilizing. Hence (4.10) with $u(\cdot) = u^*(\cdot)$ and $P = P^*$ yields

$$J(x_0, u^*(\cdot)) = x_0^T P^* x_0 + \mathsf{E}\int_0^\infty [u^*(t) - K^* x^*(t)]^T (R + D^T P^* D)[u^*(t) - K^* x^*(t)]dt, \tag{4.11}$$

with $K^* := -(R + D^T P^* D)^+ (B^T P^* + D^T P^* C)$. Next, we show that

$$\begin{aligned} & [u^*(t) - K^* x^*(t)]^T (R + D^T P^* D)[u^*(t) - K^* x^*(t)] \\ \equiv\ & [u^*(t) - Z_U^*(Z_N^*)^{-1} x^*(t)]^T (R + D^T P^* D)[u^*(t) - Z_U^*(Z_N^*)^{-1} x^*(t)] \\ =\ & 0. \end{aligned} \tag{4.12}$$

To this end, apply the complementary duality. From the relation $\mathcal{L}(P^*)Z^* = 0$, it follows that

$$\begin{aligned} & \begin{bmatrix} R + D^T P^* D, & 0 \\ 0, & F(P^*) \end{bmatrix} \begin{bmatrix} I, & -K^* \\ 0, & I \end{bmatrix} \begin{bmatrix} Z_B^*, & Z_U^* \\ (Z_U^*)^T, & Z_N^* \end{bmatrix} \\ &= \begin{bmatrix} (R + D^T P^* D)(Z_B^* - K^*(Z_U^*)^T), & (R + D^T P^* D)(Z_U^* - K^* Z_N^*) \\ F(P^*)(Z_U^*)^T, & F(P^*) Z_N^* \end{bmatrix} \\ &= \begin{bmatrix} (R + D^T P^* D) Z_B^* + (B^T P^* + D^T P^* C)(Z_U^*)^T, & (R + D^T P^* D) Z_U^* + (B^T P^* + D^T P^* C) Z_N^* \\ F(P^*)(Z_U^*)^T, & F(P^*) Z_N^* \end{bmatrix} \\ &= \begin{bmatrix} 0, & 0 \\ 0, & 0 \end{bmatrix}. \end{aligned}$$

Therefore,

$$(R + D^T P^* D) Z_U^* = -(B^T P^* + D^T P^* C) Z_N^*. \tag{4.13}$$

Repeatedly using the above identity, we get

$$\begin{aligned}
0 &= [u^*(t) - Z_U^*(Z_N^*)^{-1}x^*(t)]^T (R + D^T P^* D)[u^*(t) - Z_U^*(Z_N^*)^{-1}x^*(t)] \\
&= u^*(t)^T (R + D^T P^* D) u^*(t) - 2u^*(t)^T (R + D^T P^* D) Z_U^*(Z_N^*)^{-1} x^*(t) \\
&\quad + x^*(t)^T (Z_N^*)^{-1} (Z_U^*)^T (R + D^T P^* D) Z_U^* (Z_N^*)^{-1} x^*(t) \\
&= u^*(t)^T (R + D^T P^* D) u^*(t) + 2u^*(t)^T (B^T P^* + D^T P^* C) x^*(t) \\
&\quad + x^*(t)^T (Z_N^*)^{-1} (Z_U^*)^T (R + D^T P^* D)(R + D^T P^* D)^+ (R + D^T P^* D) Z_U^* (Z_N^*)^{-1} x^*(t) \\
&= u^*(t)^T (R + D^T P^* D) u^*(t) + 2u^*(t)^T (B^T P^* + D^T P^* C) x^*(t) \\
&\quad + x^*(t)^T (P^* B + C^T P^* D)(R + D^T P^* D)^+ (B^T P^* + D^T P^* C) x^*(t) \\
&= [u^*(t) - K^* x^*(t)]^T (R + D^T P^* D)[u^*(t) - K^* x^*(t)].
\end{aligned}$$

This proves (4.12). It then follows from (4.10) and (4.11) that

$$J(x_0, u^*(\cdot)) = x_0^T P^* x_0 \le J(x_0, u(\cdot)) \qquad \forall u(\cdot) \in U_{ad}^{x_0}.$$

Hence, $u^*(\cdot)$ is optimal. □

A sufficient condition for Theorem 4.4 to hold is the *strict complementarity* of the SDPs (P) and (D), as the following proposition asserts.

PROPOSITION 4.5. *Suppose that* (P) *and* (D) *have strictly complementary optimal solutions* P^* *and* Z^**, respectively, i.e.,* $\mathcal{L}(P^*)Z^* = 0$ *and* $\mathcal{L}(P^*) + Z^* \succ 0$*. Then,* $Z_N^* \succ 0$*.*

Proof. Following Lemma 4.3 and (4.7) we have

$$\mathcal{L}(P^*) = (H^{-1})^T \mathrm{diag}(R + D^T P^* D, 0) H^{-1},$$

where

$$H = \begin{bmatrix} I, & K \\ 0, & I \end{bmatrix} \text{ and } K = -(R + D^T P^* D)^+ (B^T P^* + D^T P^* C).$$

Let $\bar{Z}^* := H^{-1} Z^* (H^{-1})^T$.

It is readily seen that $\bar{Z}_N^* = Z_N^*$. Moreover, diag $(R + D^T P^* D, 0) = H^T \mathcal{L}(P^*) H$ and $\bar{Z}^* = H^{-1} Z^* (H^{-1})^T$ are positive semidefinite and they are complementary to each other. We shall further show that they stay strictly complementary. To see this we first note that the range space of $\mathcal{L}(P^*)$, range$(\mathcal{L}(P^*))$, and the range space of Z^*, range(Z^*), form an orthogonal decomposition of the whole space. Clearly, range$(H^T \mathcal{L}(P^*) H)$ has the same dimension as that of range$(\mathcal{L}(P^*))$; and range$(\bar{Z}^*)$ has the same dimension as that of range(Z^*). Finally, range$(H^T \mathcal{L}(P^*) H)$ and range$(\bar{Z}^*)$ remain orthogonal to each other. Hence they span the whole space too. Therefore,

$$H^T \mathcal{L}(P^*) H + \bar{Z}^* \succ 0.$$

As we noted before, the last diagonal block of the above matrix is Z_N^*. Hence, $Z_N^* \succ 0$, as the proposition stipulates. □

If P^* and Z^* are strictly complementary optimal solutions with

$$R + D^T P^* D \succ 0, \tag{4.14}$$

and $Z_N^* \succ 0$, then by (4.13), the feedback control in (4.9) coincides with the one in (3.5). But if $R + D^T P^* D$ is singular, then these two controls can indeed be *different*; see Example 6.1.

5. Synthesis. To summarize the results we have so far obtained, consider the following statements:

(a) (LQ) is attainable at any $x_0 \in \Re^n$.

(b) (P) has an optimal solution P^* that satisfies:

(i) the generalized stochastic Riccati equation $F(P) = 0$;

(ii) the corresponding feedback control $u^*(t)$ of (4.5) is stabilizing.

(c) (P) and (D) have complementary optimal solutions P^* and Z^*, with $Z_N^* \succ 0$.

The following theorem is a summary of our main results.

THEOREM 5.1. *The following implications hold:*

- (a) $\Rightarrow$ (b(i))*: Theorem* 4.1.
- (b) $\Rightarrow$ (a)*, with the control in* (4.5) *being optimal: Theorem* 4.2.
- (b) $\Rightarrow$ (c)*: Theorem* 3.3.
- (c) $\Rightarrow$ (a)*, with the control in* (4.9) *being optimal: Theorem* 4.4.

Some remarks are in order. Among the above statements, (a) is a direct statement about the solution of the original problem (LQ); (b) and (c), on the other hand, provide two computational approaches to (LQ) via SDP — note that they both imply (a) with the respective optimal feedback controls explicitly given. There are differences, however, between the two; in particular, they have different requirements, and lead to different controls. Computationally, (c) appears to have an edge over (b), as most SDP solvers are based on primal-dual interior point methods. This implies that the iterative solutions produced by such a solver will likely converge to the analytic centers of the primal and dual optimal sets, respectively, which are known to be "maximally complementary" to each other. Therefore, if there is indeed any dual optimal solution with $Z_N^* \succ 0$, then the solver will return such a solution. In this respect, checking $Z_N^* \succ 0$ is much easier than verifying the stabilizing condition in (b(ii)).

Furthermore, Theorem 5.1 also reveals the relationship between (b) and (c): (b) implies (c), whereas (c) implies (b(i)) via (a). That (c) cannot imply the stabilizing condition in (b(ii)) is in itself an interesting fact, which suggests that the two controls in (4.5) and (4.9) are in general intrinsically different. Even when the former is stabilizing, and hence both controls are optimal—since we then have (b) $\Rightarrow$ (c) $\Rightarrow$ (a)—they can still be different (except for the special case of (4.14), where (b) and (c) become equivalent); see Example 6.1.

If (b(ii)) is satisfied, then (b) is reduced to (b(i)). Consequently, (a), (b), and (c) are equivalent. Hence we have the following result.

COROLLARY 5.2. *Suppose that the control in* (4.5) *is stabilizing, Then,* (a), (b), *and* (c) *are equivalent.*

An important special case is when

$$Q \succ 0, \qquad R \succ 0. \tag{5.1}$$

In this case, (P) satisfies the Slater condition because $P = 0$ is a strictly feasible solution, and so does (D) because of the mean-square stabilizability assumption of the original LQ problem. Therefore, (c) holds. On the other hand, by [1, Corollary 5.1] the control in (4.5) must be stabilizing. Hence Corollary 5.2 stipulates that (a) and (b) must hold true as well.

COROLLARY 5.3. *Suppose* (5.1) *holds. Then, the three statements* (a), (b), *and* (c) *hold true.*

Based on the results obtained earlier, it is possible to develop a computational procedure as follows to provide a complete treatment of the stochastic LQ control

problem, and in particular to answer if the problem has an optimal feedback control representable as in (4.5) or in (4.9). The procedure involves only solving some LMIs, an SDP and its dual, for which numerical algorithms have been extensively developed (among others see [11, 19, 20]). Insofar as the complementary primal SDP solution satisfies the generalized Riccati equation (see Lemma 4.3), the procedure can also be viewed as a numerical approach to solving the Riccati equation.

Step 1. Check if the feasible set of (P) is nonempty (which is an LMI condition). If not, then stop: the LQ problem cannot be solved by either the SDP approach or by the Riccati equation; else continue.

Step 2. Check if (D) satisfies the Slater condition, which amounts to solving a system of strict LMIs. If not, then stop: the LQ problem is not mean-square stablizable according to Proposition 3.1(ii) and hence ill-posed; else continue.

Step 3. At this point we know that (P) is feasible and (D) satisfies the Slater condition, and hence (P) has an optimal solution (see, e.g., [14, Theorem 5]). Check if there is any optimal solution of (P) that satisfies $F(P) = 0$. If not, then stop: the LQ problem has no attainable optimal feedback control according to Theorem 4.1; else continue.

Step 4. Check if the control in (4.5) is stabilizing (which can be checked by LMIs according to Proposition 3.1(iii)). If yes, then stop: the control is optimal; else continue.

Step 5. Check if (P) and (D) have complementary optimal solutions P^* and Z^* with $Z_N^* \succ 0$. If yes, then stop: the control $u^*(t) = Z_U^*(Z_N^*)^{-1}x^*(t)$ is optimal; otherwise (LQ) cannot be solved by our SDP approach, nor can it be solved by any other existing method.

Notice that in practical implementation one might as well starts solving (P) and (D) by means of a primal-dual interior point code (e.g., that of using the homogeneous self-dual embedding technique; see [19]), i.e., running Step 5 first. If the result turns out to be positive, then Steps 1–4 are not necessary. On the other hand, even if the result is negative, the algorithm will still tell the feasibility of (P) and (D). This makes it easier to carry out Step 1, followed by subsequent steps.

6. Examples. The first example below demonstrates how the LQ control problem can be solved by the SDP approach developed here, even in the presence of the singularity of $R + D^T P^* D$. It also shows that optimal stabilizing controls can be obtained by both SDP approaches in (b) and (c) of Theorem 5.1, leading to different optimal controls in (4.5) and (4.9), respectively.

Example 6.1. Let $m = n = 1$; $A = C = -1$, $B = D = 1$; $Q = 1$ and $R = -1$. Namely, the problem is this:

$$\begin{array}{ll} \min & \mathsf{E}\int_0^\infty [x(t)^2 - u(t)^2]dt \\ \text{s.t.} & dx(t) = [-x(t) + u(t)]dt + [-x(t) + u(t)]dW(t), \\ & x(0) = x_0. \end{array}$$

This system is mean-square stabilizable, as $u(t) = \alpha x(t)$ is stabilizing for any α with $|\alpha| < 1$. To see this, applying Itô's formula to the system (1) under the above feedback control, we obtain

$$d\mathsf{E}[x(t)^2] = (\alpha^2 - 1)\mathsf{E}[x(t)^2]dt, \;\; \mathsf{E}[x(0)^2] = x_0^2.$$

Hence

$$\mathsf{E}[x(t)^2] = e^{(\alpha^2-1)t}x_0^2, \tag{6.1}$$

which converges to 0 as $t \to +\infty$.

Now, the primal SDP is

$$\begin{array}{ll} \max & p \\ \text{s.t.} & \left[\begin{array}{cc} -1+p, & 0 \\ 0, & 1-p \end{array} \right] \succeq 0. \end{array}$$

The above has an optimal solution $p^* = 1$ (the only feasible solution), which also satisfies the generalized Riccati equation $F(p) = 1 - p = 0$. (Note that singularity occurs in this solution). The feedback control given by (4.5) reduces to $u^*(t) = 0$, which is stabilizing as shown above ($\alpha = 0 < 1$). Therefore, all the tests in Steps 1–4 of the computational procedure presented in section 5 are passed. Consequently, $u^*(t) = 0$ is *one* optimal control of the LQ problem. Moreover, the corresponding objective value is

$$\mathsf{E} \int_0^\infty [x^*(t)]^2 dt = x_0^2 \int_0^\infty e^{-t} dt = x_0^2, \tag{6.2}$$

where the first equality is due to (6.1) with $\alpha = 0$.

Next, we can obtain additional — in fact, infinitely many more — optimal controls by virtue of (c). Indeed, the dual SDP in this case is

$$\begin{array}{ll} \min & -z_b + z_n \\ \text{s.t.} & 1 + z_b - z_n = 0, \\ & z := \left[\begin{array}{cc} z_b, & z_u \\ z_u, & z_n \end{array} \right] \succeq 0. \end{array}$$

It can be directly verified that the above has multiple optimal solutions:

$$(z_b, z_u, z_n) = (z_b, z_u, 1 + z_b),$$

parameterized by (z_u, z_b) with

$$z_b \geq 0, \;\; z_u^2 \leq z_b(1 + z_b). \tag{6.3}$$

In particular, note that the above ensures $z_n = 1 + z_b > 0$. Furthermore, these (parameterized) solutions are all complementary to the primal optimal solution $p^* = 0$. Hence, the test in Step 5 of the numerical procedure is passed, which gives rise to (multiple) optimal controls

$$u^*(t) = z_u z_n^{-1} x^*(t) \equiv \alpha x^*(t).$$

Notice that the feedback gain α satisfies

$$|\alpha| = \frac{z_u}{1 + z_b} \leq \frac{\sqrt{z_b(1 + z_b)}}{1 + z_b} = \sqrt{\frac{z_b}{1 + z_b}} < 1,$$

where the first inequality follows from (6.3). Therefore, these controls are indeed stabilizing by (6.1). Finally, the optimal cost corresponding to these controls is

$$\begin{array}{rcl} J = & \mathsf{E} \int_0^\infty (1 - \alpha^2)[x(t)]^2 dt \\ = & x_0^2 (1 - \alpha^2) \int_0^\infty e^{(\alpha^2 - 1)t} dt \\ = & x_0^2, \end{array}$$

which coincides with (6.2).

The next example illustrates two points: First, when the primal SDP solution satisfies the Riccati equation (i.e., (b(i)) holds), and moreover (4.14) holds, the resulting feedback control may still not be stabilizing (i.e., (b(ii)) fails). When this does happen, the complementarity condition in (c) fails. Second, there indeed exist well-posed stochastic control problems which, however, do not have any attainable optimal control. This calls for approximation methods, which will be presented in the next section.

Example 6.2. Suppose $m = n = 1$; $A = C = 0$, $B = D = 1$; $Q = 4$ and $R = -1$. The control system is as follows:

$$\begin{array}{ll} \min & \mathsf{E}\int_0^\infty [4x(t)^2 - u(t)^2]dt \\ \text{s.t.} & dx(t) = u(t)dt + u(t)dW(t), \\ & x(0) = x_0. \end{array}$$

Consider a feedback control $u(t) = -kx(t)$. Applying Itô's lemma yields

$$d[x(t)]^2 = (k^2 - 2k)[x(t)]^2 - 2k[x(t)]^2 dW(t).$$

Clearly, such a feedback control is stabilizing if and only if $k^2 - 2k < 0$ or $0 < k < 2$. In particular, this implies that $u(t) = -x(t)$ is stabilizing while $u(t) = -2x(t)$ is not.

For any $0 < k < 2$, it follows that $\mathsf{E}\int_0^\infty [x(t)]^2 dt = x_0^2/(2k - k^2)$. Therefore, the control $u(t) = -kx(t)$ has a cost

$$(4 - k^2)\mathsf{E}\int_0^\infty [x(t)]^2 dt = \left(1 + \frac{2}{k}\right) x_0^2.$$

As $k \uparrow 2$ we see that the cost can be arbitrarily close to $2x_0^2$. Nevertheless, this optimum is not attainable when $x_0 \neq 0$.

In terms of the corresponding SDPs, the primal reads

$$\begin{array}{ll} \max & p \\ \text{s.t.} & \left[\begin{array}{cc} -1 + p, & p \\ p, & 4 \end{array} \right] \succeq 0. \end{array}$$

This problem has only one feasible solution $p^* = 2$, which is necessarily the optimal solution. It clearly satisfies the Riccati equation:

$$4 + \frac{p^2}{1 - p} = 0.$$

Hence the control in (4.5) is

$$u^*(t) = -\frac{p^*}{-1 + p^*} x^*(t) = -2x^*(t),$$

which is *not* stabilizing as we discussed before.

Since (b(ii)) does not hold, we expect (c) to fail as well, as (b) and (c) are equivalent under (4.14). So, let us now examine the dual:

$$\begin{array}{ll} \min & 4z_n - z_b \\ \text{s.t.} & 1 + 2z_u + z_b = 0, \\ & z := \left[\begin{array}{cc} z_b, & z_u \\ z_u, & z_n \end{array} \right] \succeq 0. \end{array}$$

This problem is strictly feasible, since the original LQ problem is stabilizable. For instance, $(z_b, z_u, z_n) = (1, -1, 2)$ is a strictly feasible solution. Hence ([23, Theorem 3.1]), the infimum of the dual objective value must coincide with the supremum of the primal objective value, which is 2. This means, should the dual optimal solution z^* exist, it must satisfy $4z_n^* - z_b^* = 2$, or $z_n^* = (2 + z_b^*)/4$. This, along with $z_u^* = -(1 + z_b^*)/2$, leads to

$$z_b^* z_n^* = z_b^* \cdot \frac{2 + z_b^*}{4} < \frac{(1 + z_b^*)^2}{4} = (z_u^*)^2,$$

which violates $z \succeq 0$. Consequently, the dual does not have an attainable optimal solution, and the complementary duality fails.

The third example below illustrates a situation opposite to Example 6.2: (b(i)) fails while (b(ii)) holds. Namely, when $R + D^T P^* D = 0$, the optimal primal solution P^* may *not* satisfy the generalized Riccati equation $F(P) = 0$ (and *vice versa*), even when the corresponding control is stabilizing. In this case, Theorem 4.1 shows that the LQ problem has no attainable optimal control.

Example 6.3. Let $m = n = 1$; $A = -1$, $B = 1$, $C = D = 0$; $Q = 1$ and $R = 0$. This is actually a deterministic system that is mean-square stabilizable, as $u(t) = 0$ is stabilizing. The corresponding SDP reads:

$$\begin{array}{ll} \max & p \\ \text{s.t.} & \left[\begin{array}{cc} 0, & p \\ p, & 1 - 2p \end{array} \right] \succeq 0. \end{array}$$

The above has a unique feasible solution $p^* = 0$, which is hence optimal too. However, the generalized Riccati equation in this case is $F(p) = 1 - 2p = 0$, which has a unique solution $p = \frac{1}{2}$. Therefore, the two solutions are completely different.

Moreover, notice that while (b(i)) fails in this case, (b(ii)) does hold: $u^*(t) = 0$ is indeed stabilizing. On the other hand, in view of Theorem 5.1, (c) $\Rightarrow$ (a) $\Rightarrow$ (b(i)), we expect (c) to fail. Indeed, the dual SDP is

$$\begin{array}{ll} \min & z_n \\ \text{s.t.} & 1 + 2z_u - 2z_n = 0, \\ & z := \left[\begin{array}{cc} z_b, & z_u \\ z_u, & z_n \end{array} \right] \succeq 0. \end{array}$$

This dual problem is strictly feasible, and it has an infimum equal to 0, which is the supremum of the primal. However, the dual optimal solution is not attainable, because whenever $z_n = 0$ we must have $z_u = 1/2$, and hence it is impossible to have $z \succeq 0$.

7. ϵ-Approximation. Examples 6.2 and 6.3 have illustrated that the LQ control problem could be well-posed, but still there exists no attainable optimal control. When this happens we propose to consider (LQ$_\epsilon$), obtained by keeping all the data A, B, C, and D in (LQ) unchanged, and letting $R_\epsilon = R + \epsilon I$ and $Q_\epsilon = Q + \epsilon I$ with $\epsilon > 0$; such a perturbation was already considered in the proof of Theorem 4.1. Recall that the associated SDPs for (LQ$_\epsilon$) are

$$\begin{array}{lll} (\mathrm{P}_\epsilon) & \max & \langle I, P \rangle \\ & \text{s.t.} & \left[\begin{array}{cc} R + \epsilon I + D^T P D, & B^T P + D^T P C \\ P B + C^T P D, & Q + \epsilon I + C^T P C + A^T P + P A \end{array} \right] \succeq 0, \\ & & P \in \mathcal{S}^{n \times n}, \end{array}$$

and

$$\begin{array}{lll}(\mathrm{D}_\epsilon) & \min & \langle R+\epsilon I, Z_B\rangle + \langle Q+\epsilon I, Z_N\rangle \\ & \text{s.t.} & I + Z_U^T B^T + BZ_U + Z_N A^T + AZ_N \\ & & \quad + CZ_N C^T + DZ_U C^T + CZ_U^T D^T + DZ_B D^T = 0, \\ & & \left[\begin{array}{ll} Z_B, & Z_U \\ Z_U^T, & Z_N \end{array}\right] \succeq 0. \end{array}$$

Assuming that the LQ is stablizable and (P) is feasible, both (P_ϵ) and (D_ϵ) satisfy the Slater condition.

THEOREM 7.1. *Suppose* (LQ) *is well-posed. Let* $J_\epsilon^*(x_0)$ *and* $J^*(x_0)$ *be the optimal values of* (LQ_ϵ) *and* (LQ)*, respectively. Then,*

$$\lim_{\epsilon \downarrow 0} J_\epsilon^*(x_0) = J^*(x_0).$$

Proof. Let the optimal solution of (P_ϵ) be P_ϵ^*. In the proof of Theorem 4.1, we proved that

$$u^\epsilon(t) = -(R_\epsilon + D^T P_\epsilon^* D)^+ (B^T P_\epsilon^* + D^T P_\epsilon C) x^\epsilon(t) \tag{7.1}$$

is optimal for (LQ_ϵ), with the corresponding optimal objective value equal to $J_\epsilon^*(x_0) = x_0^T P_\epsilon^* x_0$.

Following the same argument as in the proof of Theorem 4.1, we know that P_ϵ^* is contained in a compact set, with $0 < \epsilon \le \epsilon_0$. Moreover, since by definition $J_\epsilon^*(x_0)$ decreases monotonically as $\epsilon \downarrow 0$, so does P_ϵ^*. Therefore, P_ϵ^* itself also converges as $\epsilon \downarrow 0$.

What remains is to show that $x_0^T P_0^* x_0$ is equal to the true infimum of (LQ), now denoted as $J^*(x_0)$. To this end, first note that

$$x_0^T P_\epsilon^* x_0 = J_\epsilon^*(x_0) \ge J^*(x_0),$$

where the inequality is due to the positive perturbation in (P_ϵ). Letting $\epsilon \to 0$, we obtain

$$x_0^T P_0^* x_0 \ge J^*(x_0).$$

On the other hand, since P_0^* is feasible to (P) (see the proof of Theorem 4.1), it follows from (4.10) that

$$J^*(x_0) \equiv \inf_{u(\cdot) \in U_{ad}^{x_0}} J(x_0, u(\cdot)) \ge x_0^T P_0^* x_0.$$

This completes the proof. □

The above theorem says that the objective value achieved by the perturbed problem is asymptotically optimal. The next result is concerned with the asymptotic optimality of the feedback control.

THEOREM 7.2. *The feedback control* $u^\epsilon(\cdot)$ *constructed by* (7.1) *is asymptotically optimal for* (LQ)*, namely,*

$$\lim_{\epsilon \downarrow 0} J(x_0, u^\epsilon(\cdot)) = J^*(x_0).$$

Proof. Denote by $J_\epsilon(x_0, u(\cdot))$ the cost of the perturbed problem (LQ_ϵ) under an admissible control $u(\cdot) \in U_{ad}^{x_0}$ w.r.t. the initial state x_0. Then for any $\eta > 0$, there is an ϵ_0 such that when $0 < \epsilon < \epsilon_0$,

$$\begin{aligned} J^*(x_0) &\le J(x_0, u^\epsilon(\cdot)) \\ &\le J_\epsilon(x_0, u^\epsilon(\cdot)) \\ &= J_\epsilon^*(x_0) \\ &\le J^*(x_0) + \eta, \end{aligned}$$

where the last inequality is due to Theorem 7.1. This proves our claim. □

Consider Example 6.2. With perturbation, the corresponding primal SDP becomes

$$\begin{array}{ll} \max & p \\ \text{s.t.} & \left[\begin{array}{cc} -1+\epsilon+p, & p \\ p, & 4+\epsilon \end{array} \right] \succeq 0. \end{array}$$

Solving this problem yields

$$p_\epsilon^* = \frac{4+\epsilon+\sqrt{(4+\epsilon)^2+4(4+\epsilon)(-1+\epsilon)}}{2}.$$

Clearly, $p_\epsilon^* = 2 + O(\sqrt{\epsilon})$, and hence the optimal value of (P_ϵ), $p_\epsilon^* x_0^2$, converges to $2x_0^2$ as $\epsilon \downarrow 0$.

8. Concluding remarks. We have developed a systematic approach to the stochastic LQ control problem based on primal-dual SDP, allowing indefinite cost matrices. We have shown that, in addition to its obvious computational advantage, the SDP duality theory provides critical qualitative information about the LQ control problem, in particular, regarding issues such as stability and optimality.

Among the three statements presented in section 5, the strongest is (b), which consists of two parts: (i) the optimal solution to the primal SDP satisfies the generalized Riccati equation, and (ii) the corresponding feedback control is stabilizing. It implies (c), the SDP complementary duality, which, in turn, implies (a): the existence of an optimal control to the LQ problem. Both (b) and (c) hence provide useful computational approaches to solving the LQ problem.

Conversely, our results also provide new insight as to when the LQ problem does *not* possess an optimal solution: Since (b(i)) is implied by (a), if no primal SDP solution satisfies the generalized Riccati equation, then the LQ problem has no attainable optimal control. (For such problems we have developed an ϵ-approximation scheme that yields asymptotic optimal solutions.) However, as the SDP in general possesses multiple optimal solutions, and most SDP solvers usually return a single optimal solution, this result is of more theoretical, as opposed to computational, interest. This limitation is reflected in Step 3 of the procedure outlined in section 5: it requires checking if there is *any* optimal solution of (P) satisfying $F(P) = 0$.

Finally, the gap alluded to in the last step of the same procedure points to an open problem: whether our SDP approach might fail to find an optimal control (a counter-example), or this is simply impossible (a proof).

Acknowledgment. We thank the Associate Editor and the two reviewers for their careful reading of an earlier version of the paper and for their insightful comments.

REFERENCES

[1] M. Ait Rami and X. Y. Zhou, *Linear matrix inequalities, Riccati equations, and indefinite stochastic linear quadratic controls*, IEEE Trans. Automat. Control, 45 (2000), pp. 1131–1143.

[2] A. Albert, *Conditions for positive and nonnegative definiteness in terms of pseudoinverses*, SIAM J. Appl. Math., 17 (1969), pp. 434–440.

[3] F. Alizadeh, *Interior point methods in semidefinite programming with applications to combinatorial optimization*, SIAM J. Optim., 5 (1995), pp. 13–51.

[4] B. D. O. Anderson and J. B. Moore, *Optimal Control—Linear Quadratic Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[5] M. Athens, *Special issue on linear-quadratic-Gaussian problem*, IEEE Trans. Automat. Control, 16 (1971), pp. 527–869.

[6] A. Bensoussan, *Lectures on stochastic control, part* I, in Nonlinear Filtering and Stochastic Control, Lecture Notes in Math. 972, Springer, New York, 1983, pp. 1–39.

[7] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM, Philadelphia, 1994.

[8] S. Chen, X. Li, and X. Y. Zhou, *Stochastic linear quadratic regulators with indefinite control weight costs*, SIAM J. Control Optim., 36 (1998), pp. 1685–1702.

[9] M. H. A. Davis, *Linear Estimation and Stochastic Control*, Chapman and Hall, London, 1977.

[10] L. El Ghaoui, R. Nikoukhah, and F. Delebecque, `LMITOOL`*: A front-end for LMI optimization, user's guide*, 1995. Available via anonymous ftp from ftp.ensta.fr, from the directory /pub/elghaoui/lmitool.

[11] K. Fujisawa, M. Fukuda, M. Kojima, and K. Nakata, *Numerical evaluation of SDPA*, in High Performance Optimization, J.B.G. Frenk et al., eds., Kluwer, Boston, Dordrecht, London, 1999, pp. 267–301.

[12] R. E. Kalman, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana, 5 (1960), pp. 102–119.

[13] M. Kohlmann and X. Y. Zhou, *Relationship between backward stochastic differential equations and stochastic controls: A linear-quadratic approach*, SIAM J. Control Optim., 38 (2000), pp. 1392–1407.

[14] Z. Q. Luo, J. F. Sturm, and S. Zhang, *Duality Results for Conic Convex Programming*, Tech. report 9719/A, Econometric Institute, Erasmus University Rotterdam, Rotterdam, The Netherlands, 1997.

[15] Z. Q. Luo, J. F. Sturm, and S. Zhang, *Conic convex programming and self-dual embedding*, Optim. Methods Softw., 14 (2000), pp. 169–218.

[16] Y. Nesterov and A. Nemirovskii, *Interior Point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia, 1994.

[17] R. Penrose, *A generalized inverse of matrices*, Proc. Cambridge Philos. Soc., 52 (1955), pp. 17–19.

[18] J. F. Sturm, *Primal-Dual Interior Point Approach to Semidefinite Programming*, Ph.D. thesis, Tinbergen Institute Series 156, Erasmus University, Rotterdam, Rotterdam, The Netherlands, 1997.

[19] J. F. Sturm, *Using SeDuMi* 1.02*, a MATLAB toolbox for optimization over symmetric cones*, Optim. Methods Softw., 11/12 (1999), pp. 625–653.

[20] K. C. Toh, M. J. Todd, and R. H. Tütüncü, *SDPT*3*—A* Matlab *software package for semidefinite programming*, Optim. Methods Softw., 11/12 (1999), pp. 545–581.

[21] J. C. Willems, *Least square stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, 16 (1971), pp. 621–634.

[22] W. M. Wonham, *On the separation theorem of stochastic control*, SIAM J. Control, 6 (1968), pp. 312–326.

[23] L. Vandenberghe and S. Boyd, *Semidefinite programming*, SIAM Rev., 38 (1996), pp. 49–95.

[24] V. A. Yakubovich, *The solution of certain linear matrix inequalities in automatic control theory*, Soviet Math. Dokl., 5 (1964), pp. 620–623.

[25] J. Yong and X. Y. Zhou, *Stochastic Controls: Hamiltonian Systems and HJB Equations*, Springer, New York, 1999.

[26] X. Y. Zhou and D. Li, *Continuous-time mean-variance portfolio selection: A stochastic LQ framework*, Appl. Math. Optim., 42 (2000), pp. 19–33.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 3, pp. 824–852

STATIONARY HAMILTON–JACOBI EQUATIONS IN HILBERT SPACES AND APPLICATIONS TO A STOCHASTIC OPTIMAL CONTROL PROBLEM*

SANDRA CERRAI†

Abstract. We study an infinite horizon stochastic control problem associated with a class of stochastic reaction-diffusion systems with coefficients having polynomial growth. The hamiltonian is assumed to be only locally Lipschitz continuous so that the quadratic case can be covered. We prove that the value function V corresponding to the control problem is given by the solution of the stationary Hamilton–Jacobi equation associated with the state system. To this purpose we write the Hamilton–Jacobi equation in integral form, and, by using the smoothing properties of the transition semigroup relative to the state system and the theory of m-dissipative operators, we show that it admits a unique solution. Moreover, the value function V is obtained as the limit of minima for some approximating control problems which admit unique optimal controls and states.

Key words. stochastic reaction-diffusion systems, stationary Hamilton–Jacobi–Bellman equations in infinite dimension, infinite horizon stochastic control problems

AMS subject classifications. 60H15, 60J35 93C20, 93E20

PII. S0363012999359949

1. Introduction.

In the present paper we are concerned with an infinite horizon stochastic control problem associated with the following reaction-diffusion system perturbed by a random term:

(1.1)
$$\begin{cases} \dfrac{\partial y_k}{\partial t}(t,\xi) = \mathcal{A}_k\, y_k(t,\xi) + f_k(\xi, y_1(t,\xi), \dots, y_r(t,\xi)) + z_k(t,\xi) + Q_k \dfrac{\partial^2 w_k}{\partial t \partial \xi}(t,\xi), \\ y_k(0,\xi) = x_k(\xi), \qquad t \geq 0, \quad \xi \in \overline{\mathcal{O}}, \\ \mathcal{B}_k\, y_k(s,\xi) = 0, \qquad \xi \in \partial\mathcal{O}, \qquad k = 1, \dots, r. \end{cases}$$

Here $\mathcal{O}$ is a bounded open set in $\mathbb{R}^d$, $d \leq 3$, with regular boundary. The second order differential operators $\mathcal{A}_k$ are strictly elliptic, have regular coefficients, and are endowed with some boundary conditions $\mathcal{B}_k$. The function $f = (f_1, \dots, f_r) : \overline{\mathcal{O}} \times \mathbb{R}^r \to \mathbb{R}^r$ is twice differentiable, has polynomial growth together with its derivatives, and verifies suitable dissipativity conditions. The linear operators Q_k are bounded and self-adjoint from $L^2(\mathcal{O})$ into itself and are not assumed to be Hilbert–Schmidt in general. Finally, the random fields $\partial^2 w_k / \partial t \partial \xi$ are mutually independent white noises in space and in time, defined on the same stochastic basis $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathbb{P})$, and z_k are square integrable processes adapted to the filtration $\mathcal{F}_t$.

Such a class of systems are of interest in applications and, especially in chemistry and in the present setting, have been widely studied by several authors (see, for example, Friedlin in [18] and Da Prato and Zabczyk in [14]). We recall that in [14] it

*Received by the editors August 4, 1999; accepted for publication (in revised form) February 8, 2001; published electronically September 28, 2001.
http://www.siam.org/journals/sicon/40-3/35994.html
†Dipartimento di Matematica per le Decisioni, Università di Firenze, Via C. Lombroso 6/17, I-50134 Firenze, Italy (cerrai@cce.unifi.it).

is proved that for any initial datum x in the Hilbert space $H = L^2(\mathcal{O}; \mathbb{R}^r)$ and for any adapted control $z \in L^2(\Omega; L^2(0, +\infty; H))$ the system (1.1) admits a unique solution $y(t; x, z)$ in a *generalized* sense that we will specify later. Moreover, if $x \in C(\overline{\mathcal{O}}; \mathbb{R}^r)$ and $z \in L^2(\Omega; L^p(0, +\infty; H))$ with $p > 4/(4 - d)$, such a solution is a *mild* solution.

In correspondence with the system (1.1) we study the following stochastic control problem: *minimizing* the cost functional

$$J(x, z) = \mathbb{E} \int_0^{+\infty} e^{-\lambda t} \left[g(y(t; x, z)) + k(z(t)) \right] dt, \tag{1.2}$$

among all controls $z \in L^2(\Omega; L^2(0, +\infty; H))$ adapted to the filtration $\mathcal{F}_t$. Here $y(t; x, z)$ is the unique solution of (1.1), and $g : H \to \mathbb{R}$ is Lipschitz continuous and bounded. Moreover, $k : H \to (-\infty, +\infty]$ is a measurable mapping such that its Legendre transform K, which is defined by

$$K(x) = \sup_{y \in H} \left\{ - \langle x, y \rangle_H - k(y) \right\}, \qquad x \in H,$$

is Fréchet differentiable and *locally* Lipschitz continuous together with its derivative.

Our aim here is to study the *value function* corresponding to the functional (1.2)

$$V(x) = \inf \left\{ J(x, z) ;\ z \in L^2(\Omega; L^2(0, +\infty; H)),\ \text{adapted} \right\}.$$

Namely, we show that, if A is the realization in H of the differential operator $\mathcal{A} = (\mathcal{A}_1, \dots, \mathcal{A}_r)$, endowed with the boundary conditions $\mathcal{B} = (\mathcal{B}_1, \dots, \mathcal{B}_r)$, and if F is the Nemytskii operator associated with the function $f = (f_1, \dots, f_r)$, then, under the assumption of Lipschitz continuity for K, for any $\lambda > 0$ and $g \in C_b(H)$[1] the infinite dimensional second order nonlinear elliptic problem

$$\lambda \varphi(x) - \frac{1}{2} \mathrm{Tr} \left[Q^2 D^2 \varphi(x) \right] - \langle Ax + F(x), D\varphi(x) \rangle_H + K(D\varphi(x)) = g(x) \tag{1.3}$$

admits a unique differentiable *mild* solution φ. This means that there exists a unique solution $\varphi \in C_b^1(H)$ to the integral problem

$$\varphi(x) = \mathbb{E} \int_0^{+\infty} e^{-\lambda t} \left[g(y(t; x)) - K(D\varphi(y(t; x))) \right] dt,$$

where $y(t; x)$ is the solution of the system (1.1), corresponding to $z = 0$. Moreover, for any $x \in H$ the solution $\varphi(x)$ coincides with the function $V(x)$. When K is only locally Lipschitz continuous, there exists $\mu_0 > 0$ such that the same result holds for any $\lambda > \mu_0$ and $g \in C_b^1(H)$.

It is important to remark that even if we assume $f(\xi, \cdot)$ to be more than once differentiable, nevertheless we are able to prove only C^1-regularity in H for the transition semigroup P_t associated with the system (1.1) (see [8]). Then the solution φ of the problem (1.3) is only C^1, and we can not prove the existence of an optimal state and an optimal control for our control problem. Actually, by following a dynamic

[1] We shall denote by $B_b(H)$ the Banach space of all bounded Borel functions $\varphi : H \to \mathbb{R}$ and by $C_b(H)$ the subspace of uniformly continuous functions. Moreover, we denote by $C_b^k(H)$, $k \in \mathbb{N}$, the subspace of all k-times Fréchet differentiable functions, having uniformly continuous and bounded derivatives, up to the kth order.

programming approach, the optimal state and the optimal control would be given, respectively, by the solution $y^\star(t)$ of the so-called *closed loop equation*

$$dy(t) = [Ay(t) + F(y(t)) - DK(D\varphi(y(t)))]\,dt + Q\,dw(t), \qquad y(0) = x, \tag{1.4}$$

and by

$$z^\star(t) = -DK(D\varphi(y^\star(t))).$$

On the other hand, as $D\varphi$ is only continuous, the mapping $x \mapsto -DK(D\varphi(x))$ is only continuous. Thus we are able only to prove the existence of martingale solutions for the problem (1.4), which are not defined in general in the original stochastic basis $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathbb{P})$, so that the control $z^\star(t)$ is not admissible for our original problem. However, in the case of space dimension $d = 1$ it is possible to prove the existence and uniqueness of solutions for the closed loop equation and then the existence and uniqueness of an optimal control. In what follows, it could be interesting to see if, by introducing the notion of *relaxed controls* (see [17] for the definition and some interesting results in finite dimension), it is possible to prove the existence of an optimal control.

Nevertheless, even if we are not able to prove in general the existence of an optimal control, we can show that the value function V is obtained as the limit of minima of suitable approximating cost functionals J_α, $\alpha \geq 0$, which admit unique optimal controls and unique optimal states and whose value functions coincide with the solutions of suitable approximating Hamilton–Jacobi problems.

Several authors have studied second order Hamilton–Jacobi equations by the approach of viscosity solutions. For the finite dimensional case we refer to the paper by Crandall, Ishii, and Lions [11] and to the book by Fleming and Soner [17], and for the infinite dimensional case we refer to the papers by Lions [25, 26] and to the thesis of Swiech [28]. Other authors have studied regular solutions of second order Hamilton–Jacobi equations, and as far as the infinite dimension is concerned we refer to the works by Barbu and Da Prato [1], Cannarsa and Da Prato [2, 3], Gozzi [20, 21], Haverneau [23] for the evolution case, and by Gozzi and Rouy [22] and Chow and Menaldi [10] for the stationary case. More recently, infinite dimensional Hamilton–Jacobi equations have been studied in connection with some ergodic control problems (see, for example, [19] and [16]).

The main novelty here lies in the fact that we can prove the existence and uniqueness of regular solutions for (1.3) when the nonlinear coefficient F in the state equation has polynomial growth and is not even well defined in the Hilbert space H. Moreover, we can treat both the case of a Lipschitz continuous hamiltonian and the case of a locally Lipschitz hamiltonian so that the quadratic case can be covered.

Due to the difficulties arising from coefficients which are not Lipschitz continuous, the study of mild solutions for the problem (1.3) is quite delicate, and we have to proceed in several steps. We first consider the case of a Lipschitz hamiltonian K, and we prove the existence and uniqueness result for λ large enough. To this purpose, we apply a fixed point argument in the space $C_b^1(H)$, and we use the regularizing properties of the semigroup P_t which have been studied in detail in [7] and [8]. Namely, it has been proved that

$$\varphi \in B_b(H) \Longrightarrow P_t\varphi \in C_b^1(H), \qquad t > 0,$$

and

$$\sup_{x\in H} |D(P_t\varphi)(x)|_H \leq c\,(t \wedge 1)^{-\frac{1+\epsilon}{2}} \sup_{x\in H} |\varphi(x)|$$

for some constant $\epsilon < 1$ depending on Q. Then, if we denote by L the weak generator of P_t (see [4] for the definition and main properties) by proceeding with suitable approximations, we show that the operator

$$N(\varphi) = L\varphi - K(D\varphi)$$

is m-dissipative. This yields the existence and uniqueness of solutions for any $\lambda > 0$. Then we consider a locally Lipschitz hamiltonian K. We approximate it by a sequence of Lipschitz functions, we consider the problems associated with the approximating hamiltonians, and, by a suitable a priori estimate, we get our result, even if in a less general case.

We remark that throughout the paper we have to proceed by several approximations because of the intrinsic difficulties in the study of the system (1.1) and because of the corresponding transition semigroup P_t. Actually, first we have to approximate the reaction term F by Lipschitz continuous functionals F_α in order to get C^2 regularity for the semigroup P_t^α associated with the system

$$dy(t) = [Ay(t) + F_\alpha(y(t))]\, dt + Q\, dw(t), \qquad y(0) = x, \tag{1.5}$$

and then we have to approximate P_t^α by the semigroups $P_t^{\alpha,n}$ associated with the finite dimensional version of (1.5) in order to apply the usual Itô calculus. Unfortunately, the direct approximation of the semigroup P_t by the semigroups $P_t^{\alpha,n}$ does not work.

2. Assumptions. We denote by H the Hilbert space $L^2(\mathcal{O}; \mathbb{R}^r)$, where $\mathcal{O}$ is a bounded open set of $\mathbb{R}^d$, $d \le 3$, having the boundary sufficiently regular. The norm and the scalar product in H are, respectively, denoted by $|\cdot|_H$ and $\langle \cdot, \cdot \rangle_H$. Moreover, we denote by E the Banach space $C(\overline{\mathcal{O}}; \mathbb{R}^r)$, endowed with the usual *sup-norm* $|\cdot|_E$.

$B_b(H)$ is the Banach space of bounded Borel functions $\varphi : H \to \mathbb{R}$, endowed with the *sup-norm*

$$\|\varphi\|_0 = \sup_{x \in H} |\varphi(x)|.$$

$C_b(H)$ is the subspace of uniformly continuous functions. Moreover, $\mathrm{Lip}_b(H)$ denotes the subspace of functions φ such that

$$[\varphi]_{\mathrm{Lip}} = \sup_{\substack{x,y \in H \\ x \neq y}} \frac{|\varphi(x) - \varphi(y)|}{|x - y|_H} < \infty.$$

$\mathrm{Lip}_b(H)$ is a Banach space endowed with the norm

$$\|\varphi\|_{\mathrm{Lip}} = \|\varphi\|_0 + [\varphi]_{\mathrm{Lip}}.$$

For each $k \in \mathbb{N}$, we denote by $C_b^k(H)$ the Banach space of k-times Fréchet differentiable functions, endowed with the norm

$$\|\varphi\|_k = \|\varphi\|_0 + \sum_{h=1}^{k} \sup_{x \in H} |D^h \varphi(x)|_{\mathcal{L}^h(H)}.$$

(Here and in what follows $\mathcal{L}^h(H) = \mathcal{L}(H; \mathcal{L}^{h-1}(H))$, $h \ge 1$, and $\mathcal{L}^0(H) = \mathbb{R}$.) Finally, for any $k \in \mathbb{N}$ and $\theta \in (0,1)$, we denote by $C_b^{k+\theta}(H)$ the subspace of all functions $\varphi \in C_b^k(H)$ such that

$$[\varphi]_{k+\theta} = \sup_{\substack{x,y \in H \\ x \neq y}} \frac{\left|D^k \varphi(x) - D^k \varphi(y)\right|_{\mathcal{L}^k(H)}}{|x - y|_H^\theta} < \infty.$$

$C_b^{k+\theta}(H)$ is a Banach space endowed with the norm

$$\|\varphi\|_{k+\theta} = \|\varphi\|_k + [\varphi]_{k+\theta}.$$

In what follows we shall assume that for any $\xi \in \overline{\mathcal{O}}$ and $\sigma = (\sigma_1, \dots, \sigma_r) \in \mathbb{R}^r$

$$f_k(\xi, \sigma_1, \dots, \sigma_r) = g_k(\xi, \sigma_k) + h_k(\xi, \sigma_1, \dots, \sigma_r), \qquad k = 1, \dots, r.$$

The functions $g_k : \overline{\mathcal{O}} \times \mathbb{R} \to \mathbb{R}$ and $h_k : \overline{\mathcal{O}} \times \mathbb{R}^r \to \mathbb{R}$ are continuous. Moreover, they are assumed to fulfill the following conditions.

Hypothesis 1.

1. *For any $\xi \in \overline{\mathcal{O}}$, the function $h_k(\xi, \cdot)$ is of class C^2 and has bounded derivatives, uniformly with respect to $\xi \in \overline{\mathcal{O}}$. Moreover, the mappings $D_\sigma^j h_k : \overline{\mathcal{O}} \times \mathbb{R}^r \to \mathbb{R}$ are continuous for $j = 1, 2$.*
2. *For any $\xi \in \overline{\mathcal{O}}$, the function $g_k(\xi, \cdot)$ is of class C^2, and there exists $m \geq 0$ such that*

$$\sup_{\xi \in \overline{\mathcal{O}}} \sup_{t \in \mathbb{R}} \frac{|D_t^j g_k(\xi, t)|}{1 + |t|^{2m+1-j}} < \infty.$$

   Moreover, the mappings $D_t^j g_k : \overline{\mathcal{O}} \times \mathbb{R} \to \mathbb{R}$ are continuous for $j = 1, 2$.
3. *If $m \geq 1$, there exist $a > 0$ and $c \in \mathbb{R}$ such that*

$$\sup_{\xi \in \overline{\mathcal{O}}} D_t g_k(\xi, t) \leq -a\, t^{2m} + c, \qquad t \in \mathbb{R}. \tag{2.1}$$

Notice that if c_k and c_{kj} are continuous functions from $\overline{\mathcal{O}}$ into $\mathbb{R}$ for $k = 1, \dots, r$ and $j = 1, \dots, 2m$, and

$$\inf_{\xi \in \overline{\mathcal{O}}} c_k(\xi) > 0,$$

then, for any $k = 1, \dots, r$, the function

$$g_k(\xi, t) = c_k(\xi)\, t^{2m+1} + \sum_{j=1}^{2m} c_{kj}(\xi)\, t^j$$

fulfills the conditions of the Hypothesis 1.

Now we define the operator F by setting for any function $x : \overline{\mathcal{O}} \to \mathbb{R}^r$

$$F(x)(\xi) = f(\xi, x(\xi)), \qquad \xi \in \overline{\mathcal{O}}.$$

If we set $p_\star = 2m+2$ and $q_\star = (2m+2)/(2m+1)$, then F is continuous from $L^{p_\star}(\mathcal{O}; \mathbb{R}^r)$ into $L^{q_\star}(\mathcal{O}; \mathbb{R}^r)$, and if $m \geq 1$, it is twice Fréchet differentiable. In particular, from (2.1) and the mean-value theorem for $x, y \in L^{p_\star}(\mathcal{O}; \mathbb{R}^r)$, it holds that

$$\langle F(x) - F(y), x - y \rangle_H \leq c\, |x - y|_H^2. \tag{2.2}$$

In the same way, we have that the functional F is twice differentiable and dissipative from E into itself. (For more details on the properties of F we refer to [7] and [9].) Notice that due to the growth conditions on f, the functional F is not even well defined in H.

As in [9], we can construct a sequence of functionals $\{F_\alpha\}_\alpha$ which are Lipschitz continuous both in H and in E and such that for any $x, y \in H$

$$\langle F_\alpha(x) - F_\alpha(y), x - y\rangle_H \le c\,|x-y|_H^2 \tag{2.3}$$

for a suitable constant c independent of $\alpha > 0$. Moreover, they are twice Fréchet differentiable in E and for each $j \le 2$ and $R > 0$

$$\lim_{\alpha\to 0} \sup_{|x|_E \le R} |D^j F_\alpha(x) - D^j F(x)|_{\mathcal{L}^j(E)} = 0.$$

Concerning the differential operator $\mathcal{A} = (\mathcal{A}_1, \dots, \mathcal{A}_r)$, we assume that for any $k = 1, \dots, r$

$$\mathcal{A}_k(\xi, D) = \sum_{i,j=1}^d a_{ij}^k(\xi)\frac{\partial^2}{\partial\xi_i\partial\xi_j} + \sum_{i=1}^d b_i^k(\xi)\frac{\partial}{\partial\xi_i}, \qquad \xi \in \overline{\mathcal{O}}.$$

The coefficients a_{ij}^k and b_i^k are of class $C^1(\overline{\mathcal{O}})$, and for any $\xi \in \overline{\mathcal{O}}$ the matrix $[a_{ij}^k(\xi)]$ is symmetric and strictly positive, uniformly with respect to $\xi \in \overline{\mathcal{O}}$. The boundary operators $\mathcal{B}_k$ are given by

$$\mathcal{B}_k(\xi, D) = I \quad \text{or} \quad \mathcal{B}_k(\xi, D) = \sum_{i,j=1}^d a_{ij}^k(\xi)\nu_j(\xi)\frac{\partial}{\partial\xi_i}, \quad \xi \in \overline{\mathcal{O}},$$

where ν is the exterior normal to the boundary of $\mathcal{O}$.

We denote by A the realization in H of the differential operator $\mathcal{A}$ equipped with the boundary conditions $\mathcal{B}$. The unbounded operator $A : D(A) \subset H \to H$ generates an analytic semigroup $e^{t\,A}$, which is not restrictive to assume of negative type. Thus we have

$$\langle Ax, x\rangle_H \le 0, \qquad x \in D(A). \tag{2.4}$$

Notice that each $L^p(\mathcal{O}; \mathbb{R}^r)$, $p \in [1, +\infty]$, is invariant for the semigroup $e^{t\,A}$, and if $p > 1$, then $e^{t\,A}$ is analytic in $L^p(\mathcal{O}; \mathbb{R}^r)$. Moreover, E is invariant for $e^{t\,A}$ as well, and $e^{t\,A}$ generates an analytic semigroup in E which is not strongly continuous. (For the proofs of these facts we refer to [15] and [27].)

Now, for any $k = 1, \dots, r$ we define

$$\mathcal{G}_k(\xi, D) = \sum_{i=1}^d \left(b_i^k(\xi) - \sum_{j=1}^d \frac{\partial a_{ij}^k}{\partial\xi_j}(\xi) \right) \frac{\partial}{\partial\xi_i}, \qquad \xi \in \overline{\mathcal{O}},$$

and by difference we set $\mathcal{C}_k = \mathcal{A}_k - \mathcal{G}_k$. The realization C of the operator $\mathcal{C} = (\mathcal{C}_1, \dots, \mathcal{C}_r)$ with the boundary conditions $\mathcal{B}$ generates in H a self-adjoint analytic semigroup $e^{t\,C}$. In what follows we denote by Q the bounded linear operator of components $Q_1, \dots, Q_r$.

Hypothesis 2.

1. *There exists a complete orthonormal basis* $\{e_k\}$ *in* H *which diagonalizes* C *such that* $\sup_{k\in\mathbb{N}} |e_k|_E < \infty$. *The corresponding set of eigenvalues is denoted by* $\{-\alpha_k\}$.

2. *The bounded linear operator $Q : H \to H$ is nonnegative and diagonal with respect to the complete orthonormal basis $\{e_k\}$ which diagonalizes C. Moreover, if $\{\lambda_k\}$ is the corresponding set of eigenvalues, we have*

$$\sum_{k=1}^{\infty} \frac{\lambda_k^2}{\alpha_k^{1-\gamma}} < +\infty$$

for some $\gamma > 0$.

3. *There exists $\epsilon < 1$ such that*

$$D((-C)^{\frac{\epsilon}{2}}) \subset D(Q^{-1}).$$

If the operator $\mathcal{A}$ with the boundary conditions $\mathcal{B}$ is smooth enough, then $\alpha_k \asymp k^{2/d}$. Thus, if we assume that $\lambda_k \asymp \alpha_k^{-\rho}$, when $d \leq 3$ it is possible to find some ρ such that the conditions of Hypothesis 2 are verified. (For more details see [7] and [8].)

In what follows we shall denote by P_n the projection operator of H onto H_n, the subspace generated by the eigenfunctions $\{e_1, \dots, e_n\}$. Then for any $x \in H$ we define $A_n x = P_n A P_n x$ and $F_{\alpha,n}(x) = P_n(F_\alpha(P_n x))$. It is immediate to check that there exists a constant c independent of $\alpha > 0$ and $n \in \mathbb{N}$ such that

$$\langle F_{\alpha,n}(x) - F_{\alpha,n}(y), x - y \rangle_H \leq c\,|x - y|_H^2. \tag{2.5}$$

Next, let $\{w_k(t)\}$ be a sequence of mutually independent real-valued Brownian motions defined on a stochastic basis $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathbb{P})$. The cylindrical Wiener process $w(t)$ is formally defined as

$$\sum_{k=1}^{\infty} e_k w_k(t),$$

where $\{e_k\}$ is the orthonormal basis of H introduced in Hypothesis 2(1). Under the Hypotheses 2(1) and 2(2) it is possible to show that the linear problem associated with the system (1.1),

$$dz(t) = Az(t)\,dt + Q\,dw(t), \qquad z(0) = 0, \tag{2.6}$$

admits a unique solution $w^A(t)$ which is the mean-square Gaussian process with values in H given by

$$w^A(t) = \int_0^t e^{(t-s)\,A} Q\,dw(s).$$

As shown, for example, in [14], the process w^A belongs to $C([0, +\infty) \times \overline{\mathcal{O}})$, $\mathbb{P}$ almost surely (a.s), and for any $p \geq 1$ and $T > 0$ it holds that

$$\mathbb{E} \sup_{t \in [0,T]} |w^A|_E^p < \infty. \tag{2.7}$$

3. The transition semigroup. By using the notations introduced in the previous section, the system (1.1) can be rewritten as

$$dy(t) = [Ay(t) + F(y(t)) + z(t)]\,dt + Q\,dw(t), \qquad y(0) = x. \tag{3.1}$$

The following theorem is proved in [14] in the uncontrolled case. The proof in the controlled case is analogous; thus we omit it. (For more details we refer also to [7] and [8].)

THEOREM 3.1. *Assume Hypotheses* 1 *and* 2.

1. *For any* $x \in E$ *and for any adapted process* $z \in L^2(\Omega; L^p(0,+\infty;H))$ *with* $p > 4/(4-d)$, *there exists a unique* mild *solution* $y(\cdot;x,z)$ *for the problem* (3.1) *which belongs to* $L^2(\Omega; C((0,T];E) \cap L^\infty(0,T;E))$, *for any* $T > 0$. *This means that*

$$y(t;x,z) = e^{t\,A}x + \int_0^t e^{(t-s)\,A} F(y(s;x,z))\,ds + w^A(t), \tag{3.2}$$

where $w^A(t)$ *is the solution of the linear system* (2.6). *Moreover, it holds that*

$$|y(t;x,z)|_E \le c(t)\left(|x|_E + |z|^{2m+1}_{L^p(0,+\infty;H)} + \sup_{s\in[0,t]} |w^A(s)|^{2m+1}_E\right), \tag{3.3}$$

$\mathbb{P}$-*a.s. for a suitable continuous increasing function* $c(t)$.

2. *For any* $x \in H$ *and for any adapted process* $z \in L^2(\Omega; L^2(0,+\infty;H))$, *there exists a unique* generalized *solution* $y(\cdot;x,z) \in L^2(\Omega; C([0,+\infty);H))$ *for the problem* (3.1). *This means that for any sequence* $\{z_n\} \subset L^2(\Omega; L^p(0,+\infty;H))$ *converging to* z *in* $L^2(\Omega; L^2(0,+\infty;H))$ *and for any sequence* $\{x_n\} \subset E$ *converging to* x *in* H, *the corresponding sequence of mild solutions* $\{y(\cdot;x_n,z_n)\}$ *converges to the process* $y(\cdot;x,z)$ *in* $C([0,T];H)$, $\mathbb{P}$-*a.s., for any fixed* $T > 0$. *Moreover, it holds that*

$$|y(t;x,z)|_H \le c(t)\left(|x|_H + |z|^{2m+1}_{L^2(0,+\infty;H)} + \sup_{s\in[0,t]} |w^A(s)|^{2m+1}_E\right),$$

$\mathbb{P}$-*a.s., for a suitable continuous increasing function* $c(t)$.

3. *The generalized solution* $y(\cdot;x,z)$ *belongs to* $L^{2m+2}(0,+\infty; L^{2m+2}(\mathcal{O};\mathbb{R}^r))$, $\mathbb{P}$-*a.s., and fulfills the integral equation* (3.2).

4. *For any* $x_1, x_2 \in H$ *and* $z_1, z_2 \in L^2(\Omega; L^2(0,+\infty;H))$ *it holds that*

$$|y(t;x_1,z_1) - y(t;x_2,z_2)|_H \le c(t)\left(|x_1 - x_2|_H + |z_1 - z_2|_{L^2(0,t;H)}\right), \tag{3.4}$$

$\mathbb{P}$-*a.s., for a suitable continuous increasing function* $c(t)$.

Next, for any $\alpha > 0$, we introduce the approximating problem

$$dy(t) = (Ay(t) + F_\alpha(y(t)) + z(t))\,dt + Q\,dw(t), \qquad y(0) = x. \tag{3.5}$$

If $x \in H$ and $z \in L^2(\Omega; L^2(0,+\infty;H))$, the system (3.5) admits a unique mild solution $y_\alpha(\cdot;x,z) \in L^2(\Omega; C([0,+\infty);H))$. If $x \in E$ and $z \in L^2(\Omega; L^p(0,+\infty;H))$ with $p > 4/(4-d)$, then $y_\alpha(\cdot;x,z) \in L^2(\Omega; C((0,T];E) \cap L^\infty(0,T;E))$ for any $T > 0$. Moreover, an estimate analogous to (3.3) holds, uniformly with respect to $\alpha > 0$. Namely, there exists an increasing continuous function $c(t)$ independent of α such that

$$|y_\alpha(t;x,z)|_E \le c(t)\left(|x|_E + |z|^{2m+1}_{L^p(0,+\infty;H)} + \sup_{s\in[0,t]} |w^A(s)|^{2m+1}_E\right), \tag{3.6}$$

$\mathbb{P}$-a.s. The following approximation result has been proved already in [9].

PROPOSITION 3.2. *Under the Hypotheses* 1 *and* 2, *for any* $q \ge 1$ *there exists* $p \ge 1$ *such that if* $x \in E$ *and* $z \in L^p(\Omega; L^\infty(0,+\infty;H))$, *then it holds that*

$$\lim_{\alpha\to 0} \mathbb{E}\,|y(t;x,z) - y_\alpha(t;x,z)|^q_E = 0, \qquad \mathbb{P}\text{-}a.s., \tag{3.7}$$

uniformly with respect to (t,x) *in bounded sets of* $[0,+\infty)\times E$ *and* z *in the set*

$$\mathcal{M}_R^2=\left\{ z\in L^2(\Omega;L^2(0,+\infty;H)) \ : \ \sup_{t\geq 0}|z(t)|_H\leq R, \quad \mathbb{P}\text{-}a.s. \right\} \tag{3.8}$$

for any $R\geq 0$.

For any $\alpha>0$ and $n\in\mathbb{N}$, we denote by $y_{\alpha,n}(\cdot;x,z)$ the unique strong solution in $L^2(\Omega;C([0,+\infty);H))$ of the approximating problem

$$dy(t)=(A_ny(t)+F_{\alpha,n}(y(t))+P_nz(t))\,dt+Q_ndw(t), \qquad y(0)=P_nx, \tag{3.9}$$

with $x\in H$ and $z\in L^2(\Omega;L^2(0,+\infty;H))$ adapted. In [9, Lemma 3.4] we have shown that for any fixed $R,T>0$

$$\lim_{n\to+\infty}\sup_{|x|_H\leq R}|y_{\alpha,n}(\cdot;x,z)-y_\alpha(\cdot;x,z)|_{L^2(\Omega;C([0,T];H))}=0. \tag{3.10}$$

Moreover, we have

$$|y_{\alpha,n}(t;x,z)|_H\leq c(t)\left(|x|_H+|z|^{2m+1}_{L^2(0,+\infty;H)}+\sup_{s\in[0,t]}|w^A(s)|^{2m+1}_E\right), \qquad \mathbb{P}\text{-a.s.} \tag{3.11}$$

In what follows we shall denote by $y(t;x)$ the solution of (3.1) with $z=0$. In [7, Theorem 7.4] we have shown that if $f(\xi,\cdot)$ is k-times differentiable, then for any $t\geq 0$ the mapping

$$E\to L^2(\Omega;E), \qquad x\mapsto y(t;x)$$

is k-times Fréchet differentiable. In particular, the first derivative $Dy(t;x)h$ is the unique solution of the linearized problem

$$\frac{dv}{dt}(t)=Av(t)+DF(y(t;x))v(t), \qquad v(0)=h,$$

and it holds that

$$\sup_{x\in E}|Dy(t;x)h|_H\leq e^{c\,t}|h|_H, \qquad \mathbb{P}\text{-a.s.}$$

If $x,h\in H$, then, as shown in [8], the problem above admits a unique generalized solution $v(t;x,h)$ which is not intended to be the mean-square derivative of $y(t;x)$ in general.

In [6] we have proved that, since F_α and $F_{\alpha,n}$ are Lipschitz continuous, $y_\alpha(t;x)$ and $y_{\alpha,n}(t;x)$ are twice mean-square differentiable with respect to $x\in H$ along any direction $h\in H$. In addition, their derivatives belong to $D(A^{1/2})\subset D(Q^{-1})$ and for any $T>0$

$$\begin{aligned}
&\sup_{x\in H}|Dy_\alpha(\cdot;x)h|_{L^\infty(0,T;H)\cap L^2(0,T;D(A^{1/2}))}\leq c_T|h|_H, \qquad \mathbb{P}\text{-a.s.},\\
&\sup_{x\in H}|Dy_{\alpha,n}(t;x)h|_{L^\infty(0,T;H)\cap L^2(0,T;D(A^{1/2}))}\leq c_T|h|_H, \qquad \mathbb{P}\text{-a.s.},
\end{aligned} \tag{3.12}$$

for a constant c_T which is independent of $\alpha > 0$ and $n \in \mathbb{N}$. In [9, Lemma 4.1] we have also proved that

$$\lim_{\alpha\to 0} \mathbb{E} \sup_{|h|_H\le 1} |Dy(\cdot;x)h - Dy_\alpha(\cdot;x)h|^2_{L^\infty(0,T;H)\cap L^2(0,T;D((-A)^{1/2}))} = 0,$$

uniformly with respect to x in bounded sets of E, and in [9, Lemma 4.2] we have proved that

$$\lim_{n\to+\infty} \mathbb{E} \sup_{|h|_H\le 1} |Dy_\alpha(\cdot;x)h - Dy_{\alpha,n}(\cdot;x)h|^2_{L^\infty(0,T;H)\cap L^2(0,T;D((-A)^{1/2}))} = 0,$$

uniformly with respect to x in bounded sets of H.

Next we define the transition semigroup P_t corresponding to the system (3.1) by setting for any $\varphi \in B_b(H)$ and $x \in H$

$$P_t\varphi(x) = \mathbb{E}\,\varphi(y(t;x)), \qquad t \ge 0.$$

In an analogous way, we define the semigroups P_t^α and $P_t^{\alpha,n}$ associated, respectively, to the systems (3.5) and (3.9). Due to (3.7), for any $\varphi \in C_b(H)$ and $R > 0$

$$\lim_{\alpha\to 0} \sup_{|x|_E\le R} |P_t^\alpha\varphi(x) - P_t\varphi(x)| = 0, \tag{3.13}$$

uniformly for t in bounded sets of $[0,+\infty)$. Moreover, due to (3.10) we have that

$$\lim_{n\to+\infty} \sup_{|x|_H\le R} |P_t^{\alpha,n}\varphi(x) - P_t^\alpha\varphi(x)| = 0, \tag{3.14}$$

uniformly for t in bounded sets of $[0,+\infty)$. It is important to notice that all the properties of the semigroup P_t which we are going to describe are fulfilled by the semigroups P_t^α and $P_t^{\alpha,n}$ as well.

From (3.4) it easily follows that P_t maps $C_b(H)$ into itself as a contraction. In general P_t is not strongly continuous in $C_b(H)$. (See [4] for a counter example even in finite dimension.) Nevertheless, as $y(\cdot;x) \in L^2(\Omega; C([0,+\infty);H))$ for any fixed $x \in H$, by the dominated convergence theorem, we have that if $\varphi \in C_b(H)$, then the mapping

$$[0,+\infty) \to \mathbb{R}, \qquad t \mapsto P_t\varphi(x)$$

is continuous. Thus, by proceeding as in [4], we define the generator L of P_t as the unique closed operator $L : D(L) \subset C_b(H) \to C_b(H)$ such that

$$R(\lambda,L)\varphi(x) = \int_0^{+\infty} e^{-\lambda t} P_t\varphi(x)\,dt, \qquad \lambda > 0,$$

for any fixed $\varphi \in C_b(H)$ and $x \in H$. In a similar way we define the generators L_α and $L_{\alpha,n}$ corresponding, respectively, to the semigroups P_t^α and $P_t^{\alpha,n}$.

In [4] it is shown that for any $\varphi \in D(L)$ and $x \in H$ the mapping

$$[0,+\infty) \to \mathbb{R}, \qquad t \mapsto P_t\varphi(x)$$

is differentiable and

$$\frac{d}{dt}P_t\varphi(x) = L(P_t\varphi)(x) = P_t(L\varphi)(x).$$

The same holds for L_α and $L_{\alpha,n}$. In particular, if $\varphi \in C_b^2(H)$, we have that $P_s^{\alpha,n}\varphi \in D(L_{\alpha,n})$, for any $\alpha > 0$, $n \in \mathbb{N}$, and $s \geq 0$, and

$$L_{\alpha,n}(P_s^{\alpha,n}\varphi) = \mathcal{L}_{\alpha,n}(P_s^{\alpha,n}\varphi), \tag{3.15}$$

where the differential operator $\mathcal{L}_{\alpha,n}$ is defined by

$$\mathcal{L}_{\alpha,n}\varphi(x) = \frac{1}{2}\mathrm{Tr}\left[Q_n^2 D^2\varphi(x)\right] + \langle A_n x + F_{\alpha,n}(x), D\varphi(x)\rangle_H, \qquad x \in H. \tag{3.16}$$

Actually, if we define $\psi = \lambda P_s^{\alpha,n}\varphi - \mathcal{L}_{\alpha,n}(P_s^{\alpha,n}\varphi)$ for some $\lambda > 0$, we have that

$$R(\lambda, L_{\alpha,n})\psi(x) = \int_0^{+\infty} e^{-\lambda t}\left[\lambda P_{t+s}^{\alpha,n}\varphi(x) - P_t^{\alpha,n}\mathcal{L}_{\alpha,n}(P_s^{\alpha,n}\varphi)(x)\right] dt.$$

It is not difficult to prove that, in general, if φ is twice differentiable, then

$$P_t^{\alpha,n}(\mathcal{L}_{\alpha,n}\varphi)(x) = \mathcal{L}_{\alpha,n}(P_t^{\alpha,n}\varphi)(x).$$

Thus, as $P_s^{\alpha,n}\varphi \in C_b^2(H)$, from the Itô formula we have

$$P_t^{\alpha,n}\mathcal{L}_{\alpha,n}(P_s^{\alpha,n}\varphi)(x) = \mathcal{L}_{\alpha,n}\left(P_{t+s}^{\alpha,n}\varphi\right)(x) = \frac{d}{dt}(P_{t+s}^{\alpha,n}\varphi(x)).$$

This allows us to conclude that

$$R(\lambda, L_{\alpha,n})\psi(x) = -\int_0^{+\infty} \frac{d}{dt}(e^{-\lambda t}P_{t+s}^{\alpha,n}\varphi(x))\, dt = P_s^{\alpha,n}\varphi(x),$$

so that $P_s^{\alpha,n}\varphi \in D(L_{\alpha,n})$ and (3.15) holds.

In [8] we have proved that the semigroup P_t has a smoothing effect. Namely, it maps $B_b(H)$ into $C_b^1(H)$ for any $t > 0$, and for $i \leq j = 0, 1$ it holds that

$$\|P_t\varphi\|_j \leq c\,(t \wedge 1)^{-\frac{(j-i)(1+\epsilon)}{2}}\|\varphi\|_i, \tag{3.17}$$

where ϵ is the constant introduced in Hypothesis 2(3). As far as the semigroups P_t^α and $P_t^{\alpha,n}$ are concerned, in [6] it is proved that they map $B_b(H)$ into $C_b^2(H)$ for any $t > 0$, and

$$\|P_t^\alpha\varphi\|_j + \|P_t^{\alpha,n}\varphi\|_j \leq c_\alpha\,(t \wedge 1)^{-\frac{(j-i)(1+\epsilon)}{2}}\|\varphi\|_i \tag{3.18}$$

for any $i \leq j \leq 2$, for some constant c_α independent of n. Moreover, if $i \leq j \leq 1$, the constant c_α is independent of α as well.

We conclude, recalling that in [9] it has been proved that if $\varphi \in C_b(H)$, then for any $R > 0$

$$\lim_{\alpha\to 0}\sup_{|x|_E \leq R}|D(P_t^\alpha\varphi)(x) - D(P_t\varphi)(x)|_H = 0, \tag{3.19}$$

uniformly for t in bounded sets of $[\delta, +\infty)$ with $\delta > 0$. Moreover, it has been proved that

$$\lim_{n\to+\infty}\sup_{|x|_H \leq R}|D(P_t^{\alpha,n}\varphi)(x) - D(P_t^\alpha\varphi)(x)|_H = 0, \tag{3.20}$$

uniformly for t in bounded sets of $[\delta, +\infty)$ with $\delta > 0$.

4. The Hamilton–Jacobi equation. We are here concerned with the stationary Hamilton–Jacobi equation

$$\lambda\varphi(x) - L\varphi(x) + K(D\varphi(x)) = g(x), \qquad x \in H. \tag{4.1}$$

Our aim is to show that such an equation admits a unique solution $\varphi(\lambda, g)$ for any $\lambda > 0$ and $g \in C_b(H)$. To this purpose we first prove a regularity result for the elements of $D(L)$.

LEMMA 4.1. *Assume Hypotheses* 1 *and* 2. *Then* $D(L) \subset C_b^1(H)$, *and for any* $\lambda > 0$ *and* $g \in C_b(H)$ *it holds that*

$$\|R(\lambda, L)g\|_1 \le \rho(\lambda)\|g\|_0, \tag{4.2}$$

where $\rho(\lambda) = c(\lambda^{\frac{\epsilon-1}{2}} + \lambda^{-1})$.

Proof. We recall that if $\varphi \in C_b(H)$, then $P_t\varphi \in C_b^1(H)$ for any $t > 0$. Thus for any $x, h \in H$ and $\lambda > 0$ we have

$$R(\lambda, L)g(x+h) - R(\lambda, L)g(x) = \int_0^{+\infty} e^{-\lambda t}(P_t g(x+h) - P_t g(x))\, dt$$

$$= \int_0^{+\infty} e^{-\lambda t} \langle D(P_t g)(x), h\rangle_H \, dt + E(x, h),$$

where

$$E(x, h) = \int_0^{+\infty} e^{-\lambda t} \int_0^1 \langle D(P_t g)(x + \theta h) - D(P_t g)(x), h\rangle_H \, d\theta \, dt.$$

Due to (3.17) we have

$$\left| \int_0^{+\infty} e^{-\lambda t} \langle D(P_t g)(x), h\rangle_H \, dt \right|$$

$$\le c \int_0^{+\infty} e^{-\lambda t}(t \wedge 1)^{-\frac{1+\epsilon}{2}} \, dt \, |h|_H \, \|g\|_0 = c \left(\lambda^{\frac{\epsilon-1}{2}} + \lambda^{-1}\right) |h|_H \, \|g\|_0.$$

Moreover, as $D(P_t g)$ is continuous in H, by the dominated convergence theorem we easily have that

$$\lim_{|h|_H \to 0} \frac{|E(x, h)|}{|h|_H} = 0.$$

This implies that $R(\lambda, L)g \in C_b^1(H)$, and for any $x, h \in H$

$$\langle D(R(\lambda, L)g)(x), h\rangle_H = \int_0^{+\infty} e^{-\lambda t} \langle D(P_t g)(x), h\rangle_H \, dt \tag{4.3}$$

so that the estimate (4.2) holds true. □

Remark 4.2. Notice that due to (3.18) we can repeat the arguments used above, and we can show that both $D(L_\alpha)$ and $D(L_{\alpha,n})$ are contained in $C_b^1(H)$, and a formula analogous to (4.3) holds for the derivatives of $R(\lambda, L_\alpha)g$ and $R(\lambda, L_{\alpha,n})g$ when $g \in C_b(H)$. In particular, it holds that

$$\|R(\lambda, L_\alpha)g\|_1 + \|R(\lambda, L_{\alpha,n})g\|_1 \le \rho(\lambda)\|g\|_0. \tag{4.4}$$

Moreover, as

$$\|P_t^\alpha\varphi\|_i + \|P_t^{\alpha,n}\varphi\|_i \le c_\alpha\,(t\wedge 1)^{-\frac{(i-j)(1+\epsilon)}{2}}\|\varphi\|_j, \qquad j\le i\le 2,$$

for a constant c_α independent of $n\in\mathbb{N}$, by interpolation we have that for any $\theta_1,\theta_2\in[0,1]$

$$\|P_t^\alpha\varphi\|_{1+\theta_1} + \|P_t^{\alpha,n}\varphi\|_{1+\theta_1} \le c_\alpha\,(t\wedge 1)^{-\frac{(\theta_1-\theta_2+1)(1+\epsilon)}{2}}\|\varphi\|_{\theta_2}.$$

By proceeding as in the proof of the previous lemma, this implies that if $\varphi\in C_b^{\theta_2}(H)$, then $R(\lambda,L_\alpha)\varphi$ and $R(\lambda,L_{\alpha,n})\varphi$ are in $C_b^{1+\theta_1}(H)$ for any $\theta_1<\theta_2+(1-\epsilon)/(1+\epsilon)$ and

$$\|R(\lambda,L_\alpha)\varphi\|_{1+\theta_1} + \|R(\lambda,L_{\alpha,n})\varphi\|_{1+\theta_1} \le c_\alpha\left(\lambda^{\frac{(\theta_1-\theta_2+1)(\epsilon+1)}{2}-1}+\lambda^{-1}\right)\|g\|_{\theta_2}. \tag{4.5}$$

In particular, we have that $D(L_\alpha)$ and $D(L_{\alpha,n})$ are contained in $C_b^{1+\theta}(H)$ for any $\theta<(1-\epsilon)/(\epsilon+1)$.

4.1. Lipschitz hamiltonian K. In the proof of the existence and uniqueness of solutions for the problem (4.1) we proceed in several steps. First we assume the Lipschitz continuity of the hamiltonian K.

Hypothesis 3. *The mapping $K:H\to\mathbb{R}$ is Fréchet differentiable and Lipschitz continuous together with its derivative. Moreover, $K(0)=0$.*

Notice that the condition $K(0)=0$ is not restrictive, as we can substitute g by $g-K(0)$.

By using the Lemma 4.1 we get the following result.

PROPOSITION 4.3. *Under Hypotheses* 1, 2, *and* 3, *there exists $\lambda_0>0$ such that* (4.1) *admits a unique solution $\varphi(\lambda,g)\in C_b^1(H)$ for any $\lambda>\lambda_0$ and for any $g\in C_b(H)$.*

Proof. The equation (4.1) is equivalent to the equation

$$\varphi = R(\lambda,L)\,(g-K(D\varphi)) = \Gamma(\lambda,g)(\varphi).$$

Due to Lemma 4.1, if $\varphi\in C_b^1(H)$ and $g\in C_b(H)$, then $\Gamma(\lambda,g)(\varphi)\in C_b^1(H)$. Thus if we show that for some $\lambda_0>0$ the mapping $\Gamma(\lambda,g)$ is a contraction in $C_b^1(H)$ for any $\lambda>\lambda_0$, our thesis follows.

As K is Lipschitz continuous for any $\varphi_1,\varphi_2\in C_b^1(H)$, we have

$$\|R(\lambda,L)(K(D\varphi_1)-K(D\varphi_2))\|_1 \le c\,\rho(\lambda)\|\varphi_1-\varphi_2\|_1.$$

Thus, if we choose λ_0 such that $c\rho(\lambda_0)=1$, we have that $\Gamma(\lambda,g)$ is a contraction in $C_b^1(H)$ for any $\lambda>\lambda_0$. This implies that it admits a unique fixed point $\varphi\in C_b^1(H)$, which is the unique solution of (4.1) in $C_b^1(H)$. □

Remark 4.4. By using (4.4) it is possible to prove that there exists $\lambda_0>0$ sufficiently large such that the mappings

$$\Gamma_\alpha(\lambda,g)(\varphi) = R(\lambda,L_\alpha)(g-K(D\varphi)), \qquad \alpha>0,$$

are contractions in $C_b^1(H)$ for any $\lambda>\lambda_0$ and for any $g\in C_b(H)$, and the approximating Hamilton–Jacobi equations

$$\lambda\varphi - L_\alpha\varphi + K(D\varphi) = g \tag{4.6}$$

admit a unique solution $\varphi_\alpha(\lambda, g) \in C_b^1(H)$. Moreover, as the function $\rho(\lambda)$ in (4.4) does not depend on $\alpha > 0$, the constant λ_0 does not depend on α either.

LEMMA 4.5. *Under Hypotheses* 1, 2, *and* 3, *for any* $\lambda > 0$ *and* $g \in C_b(H)$ *we have*

$$\lim_{\alpha\to 0} \sup_{|x|_E \le R} \left| D^j \left(\Gamma_\alpha^k(\lambda, g)(0) - \Gamma^k(\lambda, g)(0) \right)(x) \right|_{\mathcal{L}^j(H)} = 0, \qquad j = 0, 1, \tag{4.7}$$

for any $k \in \mathbb{N}$ *and* $R > 0$.

Proof. We proceed by induction. For $k = 1$ the limit (4.7) is trivially verified. Assume that (4.7) holds for some $k \ge 1$. We show that this implies that (4.7) holds for $k + 1$. We have

$$\begin{aligned} &D^j \left(\Gamma_\alpha^{k+1}(\lambda, g)(0) - \Gamma^{k+1}(\lambda, g)(0) \right) \\ &= D^j \left(R(\lambda, L_\alpha) \left[g - K \left(D(\Gamma_\alpha^k(\lambda, g)(0)) \right) \right] - R(\lambda, L) \left[g - K \left(D(\Gamma^k(\lambda, g)(0)) \right) \right] \right). \end{aligned}$$

In general, if $f \in C_b(H)$ and $\{f_\alpha\}$ is any bounded generalized sequence of $C_b(H)$ such that for any $R > 0$

$$\lim_{\alpha\to 0} \sup_{|x|_E \le R} |f_\alpha(x) - f(x)| = 0, \tag{4.8}$$

then for any $R > 0$ and $j = 0, 1$ we have

$$\lim_{\alpha\to 0} \sup_{|x|_E \le R} |D^j (R(\lambda, L_\alpha)(f_\alpha - f))(x)|_H = 0. \tag{4.9}$$

Indeed, as the formula (4.3) holds for the derivative of $R(\lambda, L_\alpha)$, as well, for any $x \in H$ we have

$$D^j (R(\lambda, L_\alpha)(f_\alpha - f))(x) = \int_0^{+\infty} e^{-\lambda t} D^j \left(P_t^\alpha (f_\alpha - f) \right)(x)\, dt.$$

If x lies in a bounded set of E, due to (2.7) and (3.6) the solution $y_\alpha(t; x)(\omega)$ lies in a bounded set of E for $\mathbb{P}$-almost all $\omega \in \Omega$. Therefore, by (4.8) for any $R > 0$ this yields

$$\lim_{\alpha\to 0} \sup_{|x|_E \le R} |(f_\alpha - f)(y_\alpha(t; x))| = 0, \qquad \mathbb{P}\text{-a.s.}, \tag{4.10}$$

and by applying the dominated convergence theorem we get (4.9) for $j = 0$. As proved in [5], for any $t > 0$ we have

$$\langle D(P_t^\alpha (f_\alpha - f))(x), h \rangle_H = \frac{1}{t} \mathbb{E} \left(f_\alpha - f \right)(y_\alpha(t; x)) \int_0^t \left\langle Q^{-1} D y_\alpha(s; x) h, dw(s) \right\rangle_H,$$

where $D y_\alpha(t; x) h$ is the mean-square derivative of $y_\alpha(t; x)$ along the direction $h \in H$. Hence, thanks to (3.12), by interpolation we easily get

$$|D(P_t^\alpha (f_\alpha - f))(x)|_H \le c\, (t \wedge 1)^{-\frac{1+\epsilon}{2}} \left(\mathbb{E} \left| (f_\alpha - f)(y_\alpha(t; x)) \right|^2 \right)^{1/2},$$

and, thanks to (4.10), this implies (4.9) for $j = 1$.

Thus, since from the inductive hypothesis and the Lipschitz continuity of K the sequence $\{K(D[\Gamma^k_\alpha(\lambda,g)(0)])\}$ and $K(D[\Gamma^k(\lambda,g)(0)])$ fulfill (4.8), we can conclude that for any $R>0$

(4.11)
$$\lim_{\alpha\to0}\sup_{|x|_E\le R}\left|D^j\left(R(\lambda,L_\alpha)\left[K(D[\Gamma^{k+1}_\alpha(\lambda,g)(0)]))-K(D[\Gamma^{k+1}_\alpha(\lambda,g)(0)])\right]\right)(x)\right|=0.$$

Now, if $f\in C^1_b(H)$, for any $x\in H$ we have

$$D^j\left[(R(\lambda,L_\alpha)-R(\lambda,L))\,(g-K(Df))\right](x)$$

$$=\int_0^{+\infty}e^{-\lambda\,t}D^j\left[(P^\alpha_t-P_t)(g-K(Df))\right](x)\,dt.$$

Then, by using the estimates (3.17) and (3.18) and the limits (3.13) and (3.19), we get that

$$\lim_{\alpha\to0}\sup_{|x|_E\le R}\left|D^j\left[(R(\lambda,L_\alpha)-R(\lambda,L))\,(g-K(Df))\right](x)\right|_H=0$$

for any $R>0$. As $\Gamma^k(\lambda,L)(0)\in C^1_b(H)$, this implies that

$$\lim_{\alpha\to0}\sup_{|x|_E\le R}\left|D^j\left[(R(\lambda,L_\alpha)-R(\lambda,L))\,(g-\Gamma^k(\lambda,g)(0))\right](x)\right|_H=0,$$

and recalling (4.11) we can conclude that

$$\lim_{\alpha\to0}\sup_{|x|_E\le R}\left|D^j\left(\Gamma^{k+1}_\alpha(\lambda,g)(0)-\Gamma^{k+1}(\lambda,g)(0)\right)(x)\right|=0.$$

By induction this yields (4.7). □

In the next proposition we show that the solution $\varphi(\lambda,g)$ of the problem (4.1) can be approximated by the solutions $\varphi_\alpha(\lambda,g)$ of the problems (4.6).

PROPOSITION 4.6. *Assume Hypotheses* 1, 2, *and* 3. *Then, if* λ_0 *is the constant introduced in the Proposition* 4.3, *for any* $\lambda>\lambda_0$ *and* $g\in C_b(H)$ *it holds that*

$$\lim_{\alpha\to0}\sup_{|x|_E\le R}\left|D^j\left(\varphi(\lambda,g)-\varphi_\alpha(\lambda,g)\right)(x)\right|_{\mathcal{L}^j(H)}=0,\qquad j=0,1,\tag{4.12}$$

for any $R>0$. *In particular, for any* $\lambda>0$ *we have*

(4.13)
$$\lim_{\alpha\to0}\sup_{|x|_E\le R}\left|D^j\left[\varphi(\lambda,g)-\varphi_\alpha(\lambda+\lambda_0,g+\lambda_0\,\varphi(\lambda,g))\right](x)\right|_{\mathcal{L}^j(H)}=0,\qquad j=0,1.$$

Proof. Let us fix λ_0 as in Proposition 4.3. We have seen that $\varphi=\varphi(\lambda,g)$ and $\varphi_\alpha=\varphi_\alpha(\lambda,g)$ are, respectively, the unique fixed points of the mappings $\Gamma(\lambda,g)$ and $\Gamma_\alpha(\lambda,g)$. Since for any $\lambda>\lambda_0$ and $g\in C_b(H)$ the contraction constants of $\Gamma_\alpha(\lambda,g)$ are the same for all $\alpha>0$, for any $\epsilon>0$ there exists $k_\epsilon\in\mathbb{N}$ such that

$$\|\Gamma^{k_\epsilon}(\lambda,g)(0)-\varphi\|_1+\sup_{\alpha>0}\|\Gamma^{k_\epsilon}_\alpha(\lambda,g)(0)-\varphi_\alpha\|_1\le\epsilon.$$

Thus for $j=0,1$ and $x\in H$ we have

$$\left|D^j\left(\varphi-\varphi_\alpha\right)(x)\right|\le\epsilon+\left|D^j\left(\Gamma^{k_\epsilon}(\lambda,g)(0)-\Gamma^{k_\epsilon}_\alpha(\lambda,g)(0)\right)(x)\right|,$$

and due to (4.7) this implies (4.12). Now, since $\varphi(\lambda, g) = \varphi(\lambda + \lambda_0, g + \lambda_0\, \varphi(\lambda, g))$, by using (4.12) we can conclude that (4.13) holds true. □

Remark 4.7. For any $\alpha > 0$ and $n \in \mathbb{N}$, consider the problem

$$\lambda\varphi - L_{\alpha,n}\varphi + K_n(D\varphi) = g_n, \tag{4.14}$$

where $K_n(x) = K(P_n x)$ and $g_n(x) = g(P_n x)$ for each $n \in \mathbb{N}$ and $x \in H$. By proceeding as for the problems (4.1) and (4.6), it is possible to show that there exists λ_0 large enough such that for any $g \in C_b(H)$ and $\lambda > \lambda_0$ there exists a unique solution $\varphi_{\alpha,n}(\lambda, g) \in C_b^1(H)$. Such a solution is given by the unique fixed point of the mapping

$$\Gamma_{\alpha,n}(\lambda, g)(\varphi) = R(\lambda, L_{\alpha,n})(g_n - K_n(D\varphi)).$$

By using arguments analogous to those used in the Lemma 4.5, due to the estimates (3.11) and (3.18), and due to the limits (3.14) and (3.20), there exists $\lambda_0 > 0$ such that for $\lambda > \lambda_0$ and $g \in C_b(H)$ it holds that

$$\lim_{n\to+\infty} \sup_{|x|_H \le R} \left| D^j \left(\Gamma_{\alpha,n}^k(\lambda, g)(0) - \Gamma_\alpha^k(\lambda, g)(0) \right)(x) \right|_{\mathcal{L}^j(H)} = 0, \qquad j = 0, 1,$$

for any $\alpha > 0$, $k \in \mathbb{N}$, and $R > 0$. Thus, by proceeding as in the proof of Proposition 4.6, due to (3.14) and (3.20) it is possible to verify that there exists $\lambda_0 > 0$ such that if $\lambda > \lambda_0$, then for any $\alpha > 0$, and $R > 0$ it holds that

$$\lim_{n\to+\infty} \sup_{|x|_H \le R} \left| D^j \left[\varphi_\alpha(\lambda, g) - \varphi_{\alpha,n}(\lambda, g) \right](x) \right|_{\mathcal{L}^j(H)} = 0. \tag{4.15}$$

In the next proposition we show that if the datum g belongs to $C_b^1(H)$, then the approximating problems (4.6) and (4.14) have a solution of class C^2.

LEMMA 4.8. *Under Hypotheses* 1, 2, *and* 3, *if* $g \in C_b^1(H)$ *and* $\lambda > 0$, *then the solutions* $\varphi_\alpha(\lambda, g)$ *and* $\varphi_{\alpha,n}(\lambda, g)$ *of the problems* (4.6) *and* (4.14) *belong to* $C_b^2(H)$. *Moreover, for any* $R > 0$ *and* $\lambda > 0$

$$\sup_{\|g\|_1 \le R} \|\varphi_\alpha(\lambda, g)\|_2 < \infty. \tag{4.16}$$

Proof. We prove the lemma only for the problem (4.6), as the proof for the problem (4.14) is identical.

As shown in Remark 4.2, $D(L_\alpha) \subset C_b^{1+\theta}(H)$ for any $\theta < (1-\epsilon)/(1+\epsilon)$. Thus, if $\varphi_\alpha(\lambda, g)$ is the solution of the problem (4.6), we have that $\varphi_\alpha(\lambda, g) \in C_b^{1+\theta_0}(H)$ for some $0 < \theta_0 < (1-\epsilon)/(1+\epsilon)$. As we have

$$\varphi_\alpha(\lambda, g) = R(\lambda, L_\alpha)\left(g - K(D\varphi_\alpha(\lambda, g))\right),$$

by using again Remark 4.2 it follows that $\varphi_\alpha(\lambda, g) \in C_b^{1+2\,\theta_0}(H)$. Therefore, by repeating this argument a finite number of steps we get that $\varphi_\alpha(\lambda, g) \in C_b^2(H)$.

The estimate (4.16) follows as above by applying (4.5) a finite number of times. □

Due to (3.15), the previous lemma implies that if $g \in C_b^1(H)$, then $\varphi_{\alpha,n} = \varphi_{\alpha,n}(\lambda, g)$ is a strict solution of the problem (4.14); that is,

$$\lambda\varphi_{\alpha,n} - \mathcal{L}_{\alpha,n}\varphi_{\alpha,n} + K_n(D\varphi) = g_n,$$

where $\mathcal{L}_{\alpha,n}$ is the differential operator introduced in (3.16).

Now, for any $\varphi \in D(L)$ we define

$$N(\varphi) = L\varphi - K(D\varphi). \tag{4.17}$$

In the same way, for any $\alpha > 0$ and $n \in \mathbb{N}$ we define $N_\alpha(\varphi) = L_\alpha\varphi - K(D\varphi)$ and $N_{\alpha,n}(\varphi) = L_{\alpha,n}\varphi - K_n(D\varphi)$.

THEOREM 4.9. *Under Hypotheses* 1, 2, *and* 3, *the operator* N *defined by* (4.17) *is m-dissipative. Thus for any* $\lambda > 0$ *and for any* $g \in C_b(H)$ *there exists a unique solution* $\varphi(\lambda, g) \in D(L)$ *for the problem* (4.1).

Thanks to Proposition 4.3, in order to show that N is m-dissipative, it suffices to show that N is dissipative. To this purpose, we first give the following preliminary result.

LEMMA 4.10. *Assume that Hypotheses* 1, 2, *and* 3 *hold. Then there exists* $\lambda_0 > 0$ *such that for any* $\lambda > \lambda_0$ *and* $\varphi_1, \varphi_2 \in D(L_\alpha)$

$$\|\varphi_1 - \varphi_2\|_0 \le \frac{1}{\lambda}\,\|\lambda(\varphi_1 - \varphi_2) - (N_\alpha(\varphi_1) - N_\alpha(\varphi_2))\|_0\,.$$

Proof. We set $g_1 = \lambda\varphi_1 - N_\alpha(\varphi_1)$ and $g_2 = \lambda\varphi_2 - N_\alpha(\varphi_2)$, and for any $n \in \mathbb{N}$ we set $g_{1,n}(x) = g_1(P_n x)$ and $g_{2,n}(x) = g_2(P_n x)$, $x \in H$. Then for λ large enough there exist $\varphi_{1,n}$ and $\varphi_{2,n}$ in $D(L_{\alpha,n})$ such that

$$\lambda\varphi_{1,n} - N_{\alpha,n}(\varphi_{1,n}) = g_{1,n}, \qquad \lambda\varphi_{2,n} - N_{\alpha,n}(\varphi_{2,n}) = g_{2,n}.$$

If we show that

$$\|\varphi_{1,n} - \varphi_{2,n}\|_0 \le \frac{1}{\lambda}\|g_{1,n} - g_{2,n}\|_0, \tag{4.18}$$

we are done. Actually, for any $x \in H$ this implies that

$$|\varphi_{1,n}(x) - \varphi_{2,n}(x)| \le \frac{1}{\lambda}\|g_{1,n} - g_{2,n}\|_0 \le \frac{1}{\lambda}\|g_1 - g_2\|_0,$$

and due to (4.15) we can take the limit as $n \to +\infty$, and we get

$$|\varphi_1(x) - \varphi_2(x)| \le \frac{1}{\lambda}\|g_1 - g_2\|_0.$$

By taking the supremum for $x \in H$, we can conclude.

Thus in order to conclude the proof we have to show that the operator $N_{\alpha,n}$ fulfills (4.18). The operator $L_{\alpha,n}$ satisfies the same conditions of the operator $\mathcal{L}$ studied in [5]; thus, thanks to [5, Proposition 7.5],

$$D(L_{\alpha,n}) = \left\{ \varphi \in \bigcap_{p\ge 1} W^{2,p}_{\rm loc}(\mathbb{R}^n) \cap C_b(\mathbb{R}^n)\,;\, \mathcal{L}_{\alpha,n}\varphi \in C_b(\mathbb{R}^n) \right\},$$

$$L_{\alpha,n}\varphi = \mathcal{L}_{\alpha,n}\varphi.$$

Now we remark that

$$K_n(D\varphi_{1,n}(x)) - K_n(D\varphi_{2,n}(x))$$

$$= \left\langle \int_0^1 DK_n(\lambda D\varphi_{1,n}(x) + (1-\lambda)D\varphi_{2,n}(x))\,d\lambda, D\varphi_{1,n}(x) - D\varphi_{2,n}(x) \right\rangle;$$

thus, if we set

$$U_{\alpha,n}(x) = \int_0^1 DK_n(\lambda D\varphi_{1,n}(x) + (1-\lambda)D\varphi_{2,n}(x))\,d\lambda,$$

we have

$$\lambda(\varphi_{1,n} - \varphi_{2,n})(x) - \mathcal{L}_{\alpha,n}(\varphi_{1,n} - \varphi_{2,n})(x)$$

$$+ \langle U_{\alpha,n}(x), D(\varphi_{1,n} - \varphi_{2,n})(x)\rangle = g_{1,n}(x) - g_{2,n}(x).$$

Since the function $U_{\alpha,n}$ is uniformly continuous, as $\varphi_{1,n}$ and $\varphi_{2,n}$ belong to $C_b^1(H)$, the operator $\mathcal{N}_{\alpha,n}$ defined by

$$\mathcal{N}_{\alpha,n}\psi(x) = \mathcal{L}_{\alpha,n}\psi(x) - \langle U_{\alpha,n}(x), D\psi(x)\rangle$$

is of the same type as the operator $\mathcal{L}$ studied in [5]. Therefore, we can adapt the proof of [5, Lemma 7.4] to the present situation, and we obtain

$$\|\varphi_{1,n} - \varphi_{2,n}\|_0 \le \frac{1}{\lambda}\|\lambda(\varphi_{1,n} - \varphi_{2,n}) - \mathcal{N}_{\alpha,n}(\varphi_{1,n} - \varphi_{2,n})\|_0 = \frac{1}{\lambda}\|g_{1,n} - g_{2,n}\|_0. \qquad \square$$

Proof of Theorem 4.9. Let us fix $\lambda > 0$ and $\varphi_1, \varphi_2 \in D(L)$, and let us define $g_1 = \lambda\varphi_1 - N(\varphi_1)$ and $g_2 = \lambda\varphi_2 - N(\varphi_2)$. If λ_0 is the maximum between the constant introduced in Remark 4.4 and the constant introduced in Lemma 4.10, for any $\alpha > 0$ there exist $\varphi_{1,\alpha}, \varphi_{2,\alpha} \in D(L_\alpha)$ such that

$$(\lambda + \lambda_0)\,\varphi_{1,\alpha} - N_\alpha\varphi_{1,\alpha} = g_1 + \lambda_0\,\varphi_1, \qquad (\lambda + \lambda_0)\,\varphi_{2,\alpha} - N_\alpha\varphi_{2,\alpha} = g_2 + \lambda_0\,\varphi_2,$$

and

$$\|\varphi_{1,\alpha} - \varphi_{2,\alpha}\|_0 \le \frac{1}{\lambda + \lambda_0}\|(g_1 - g_2) + \lambda_0\,(\varphi_1 - \varphi_2)\|_0.$$

Thus for any $x \in H$ we have

$$|\varphi_{1,\alpha}(x) - \varphi_{2,\alpha}(x)| \le \frac{1}{\lambda + \lambda_0}\|g_1 - g_2\|_0 + \frac{\lambda_0}{\lambda + \lambda_0}\|\varphi_1 - \varphi_2\|_0.$$

Now, if $x \in E$, due to (4.13) we can take the limit in the left-hand side as α goes to zero, and we get

$$|\varphi_1(x) - \varphi_2(x)| \le \frac{1}{\lambda + \lambda_0}\|g_1 - g_2\|_0 + \frac{\lambda_0}{\lambda + \lambda_0}\|\varphi_1 - \varphi_2\|_0.$$

As φ_1 and φ_2 are continuous in H, the estimate above holds also for $x \in H$, and by taking the supremum for $x \in H$ it follows that

$$\|\varphi_1 - \varphi_2\|_0 - \frac{\lambda_0}{\lambda + \lambda_0}\|\varphi_1 - \varphi_2\|_0 \le \frac{1}{\lambda + \lambda_0}\|g_1 - g_2\|_0,$$

so that

$$\|\varphi_1 - \varphi_2\|_0 \le \frac{1}{\lambda}\|g_1 - g_2\|_0. \qquad \square$$

4.2. Locally Lipschitz hamiltonian K. We first prove an a priori estimate which is crucial in order to prove the m-dissipativity of the operator N in the case of a locally Lipschitz hamiltonian K.

PROPOSITION 4.11. *Assume that Hypotheses* 1, 2, *and* 3 *hold. Then there exists some $\mu_0 > 0$, which does not depend on K, such that if $g \in C_b^1(H)$ and $\lambda > \mu_0$, then*

$$\|D\varphi(\lambda, g)\|_0 \le \|Dg\|_0. \tag{4.19}$$

Proof. Let us fix $\lambda, \mu > 0$ and $g \in C_b^1(H)$, and let us consider $\varphi_\alpha = \varphi_\alpha(\lambda + \mu, g + \mu\,\varphi(\lambda, g))$ and $\varphi_{\alpha,n} = \varphi_{\alpha,n}(\lambda + \mu, g + \mu\,\varphi(\lambda, g))$. Since $g \in C_b^1(H)$, then $\varphi_{\alpha,n}$ belongs to $C_b^2(H)$, and it is a strict solution of the problem

$$(\lambda + \mu)\,\varphi - N_{\alpha,n}(\varphi) = g_n + \mu\,\varphi_n(\lambda, g),$$

where $\varphi_n(\lambda, g)(x) = \varphi(\lambda, g)(P_n x)$. The problem above can be rewritten as

$$(\lambda + \mu)\,\varphi(x) - \frac{1}{2}\sum_{h=1}^{n} \lambda_h^2 D_h^2\varphi(x) - \sum_{h,k=1}^{n} a_{hk}x_k D_h\varphi(x)$$

$$- \langle F_\alpha(P_n x), D\varphi(x)\rangle_H + K(P_n D\varphi(x)) = g(P_n x) + \mu\,\varphi(\lambda, g)(P_n x),$$

where $D_h\varphi(x) = \langle D\varphi(x), e_h\rangle_H$ and $a_{hk} = \langle Ae_k, e_h\rangle_H$. By differentiating with respect to x_j, by setting $\psi_h = D_h\varphi$, for $h = 1, \dots, n$, and by multiplying each side by ψ_j, we get

$$(\lambda + \mu)\,\psi_j^2 - \frac{1}{2}\sum_{h=1}^{n} \lambda_h^2 \psi_j D_h^2\psi_j - \sum_{h,k=1}^{n} a_{hk}x_k\psi_j D_h\psi_j - \sum_{h=1}^{n} a_{hj}\psi_h\psi_j$$

$$- \sum_{h=1}^{n} \langle F_{\alpha,n}, e_h\rangle\,\langle \psi_j D\psi_j, e_h\rangle - \sum_{h=1}^{n} \langle DF_{\alpha,n}e_j\psi_j, e_h\rangle\,\psi_h$$

$$+ \sum_{h=1}^{n} D_h K(P_n D\varphi_{\alpha,n})\psi_j D_h\psi_j = \langle Dg_n, e_j\psi_j\rangle + \mu\,\langle D\varphi_n(\lambda, g), e_j\psi_j\rangle\,.$$

Then we sum up over j and by setting $z(x) = |D\varphi_{\alpha,n}(x)|_H^2$ and by taking into account that

$$\left(D_h^2\psi_j\right)\psi_j = \frac{1}{2}D_h^2(\psi_j^2) - (D_h\psi_j)^2,$$

we have

$$2(\lambda + \mu)\,z(x) - \frac{1}{2}\mathrm{Tr}\,\left[Q_n^2 D^2 z(x)\right] + \sum_{h,j=1}^{n} \lambda_h^2 (D_h\psi_j)^2(x) - \langle A_n x, Dz(x)\rangle$$

$$-2\,\langle A_n D\varphi_{\alpha,n}(x), D\varphi_{\alpha,n}(x)\rangle + \langle DK(D\varphi_{\alpha,n}(x)), Dz(x)\rangle - \langle F_\alpha(P_n x), Dz(x)\rangle$$

$$-2\,\langle DF_\alpha(P_n x)D\varphi_{\alpha,n}(x), D\varphi_{\alpha,n}(x)\rangle = 2\,\langle Dg(P_n x) + \mu\,D\varphi(\lambda, g)(P_n x), D\varphi_{\alpha,n}(x)\rangle\,.$$

Therefore, by using (2.3) and (2.4) it follows that

$$2(\lambda+\mu)\,z(x)-\frac{1}{2}\mathrm{Tr}\,\left[Q_n^2D^2z(x)\right]-\langle A_nx,Dz(x)\rangle-\langle F_\alpha(P_nx),Dz(x)\rangle$$

$$+\langle DK(D\varphi_{\alpha,n}(x)),Dz(x)\rangle\le 2\,\langle Dg(x)+\mu\,D\varphi(\lambda,g)(P_nx),D\varphi_{\alpha,n}(P_nx)\rangle+\gamma\,z(x)$$

$$\le 2\,(\|Dg\|_0+\mu\,\|D\varphi(\lambda,g)\|_0)\,|D\varphi_{\alpha,n}(x)|_H+\gamma\,z(x)$$

for a suitable constant $\gamma\in\mathbb{R}$ depending only on F and A.

Now let us consider the equation

$$dy(t)=[A_ny(t)+F_{\alpha,n}(y(t))+U_{\alpha,n}(y(t))]\,dt+Q_n\,dw(t),\qquad y(0)=P_nx, \tag{4.20}$$

where $U_{\alpha,n}(x)=-DK(D\varphi_{\alpha,n}(x))$ for any $x\in H$. If $g\in C_b^1(H)$, then $\varphi_{\alpha,n}\in C_b^2(H)$, and then the mapping $U_{\alpha,n}:H\to H$ is Lipschitz continuous. This implies that there exists a unique strong solution $y_{\alpha,n}(\cdot;x)\in L^2(\Omega;C([0,+\infty);H))$ for (4.20). If we denote by $R_t^{\alpha,n}$ the corresponding transition semigroup, it is possible to show that the solution of the problem

$$(2(\lambda+\mu)-\gamma)\psi(x)-\frac{1}{2}\mathrm{Tr}\,\left[Q_n^2D^2\psi(x)\right]-\langle A_nx,D\psi(x)\rangle-\langle F_\alpha(P_nx),D\psi(x)\rangle$$

$$+\langle DK(D\varphi_{\alpha,n}(x)),D\psi(x)\rangle=2\,(\|Dg\|_0+\mu\,\|D\varphi(\lambda,g)\|_0)\;|D\varphi_{\alpha,n}(x)|_H$$

for any $\lambda>\gamma$ is given by

$$\psi(x)=2\,(\|Dg\|_0+\mu\,\|D\varphi(\lambda,g)\|_0)\int_0^{+\infty}e^{-(2(\lambda+\mu)-\gamma)\,t}R_t^{\alpha,n}\,(|D\varphi_{\alpha,n}|_H)\,(x)\,dt.$$

(See [5] for a proof.) Thus by a comparison argument we have that

$$|D\varphi_{\alpha,n}(x)|_H^2\le\frac{2}{2(\lambda+\mu)-\gamma}\;(\|Dg\|_0+\mu\,\|D\varphi(\lambda,g)\|_0)\,|D\varphi_{\alpha,n}(x)|_H,$$

and if we take $\lambda>1+\gamma/2=\mu_0$, it follows that

$$|D\varphi_{\alpha,n}(\lambda+\mu,g+\mu\,\varphi(\lambda,g))(x)|_H\le\frac{1}{1+\mu}\,(\|Dg\|_0+\mu\,\|D\varphi(\lambda,g)\|_0)\,.$$

Due to (4.13) and (4.15), if μ is large enough, we can take first the limit as n goes to infinity and then the limit as α goes to zero, and for any $x\in E$ we get

$$|D\varphi(x)|_H\le\frac{1}{1+\mu}\,(\|Dg\|_0+\mu\,\|D\varphi(\lambda,g)\|_0)\,.$$

As $\varphi(\lambda,g)\in C_b^1(H)$, the same estimate holds for $x\in H$, and then, by taking the supremum for $x\in H$, we get

$$\|D\varphi(\lambda,g)\|_0\le\frac{1}{1+\mu}\,(\|Dg\|_0+\mu\,\|D\varphi(\lambda,g)\|_0)\,,$$

which immediately yields (4.19). □

Remark 4.12. It is immediate to check that the proof of the previous proposition adapts to the problem (4.6). Thus there exists $\lambda_0 > 0$, which is clearly independent of $\alpha > 0$, such that for any $\lambda > \lambda_0$ and $g \in C_b^1(H)$

$$\|D\varphi_\alpha(\lambda, g)\|_0 \le \|Dg\|_0.$$

From now on we shall assume that K fulfills the following assumption.

Hypothesis 4. *The hamiltonian $K : H \to \mathbb{R}$ is Fréchet differentiable and is locally Lipschitz continuous, together with its derivative. Moreover, $K(0) = 0$.*

We want to show that under the hypotheses above the problem (4.1) admits a unique solution for any $\lambda > \mu_0$ and $g \in C_b^1(H)$. To this purpose, for any $r > 0$ let K_r be a Fréchet differentiable function such that

$$K_r(x) = \begin{cases} K(x) & \text{if } |x|_H \le r, \\ K\left(\dfrac{(r+1)x}{|x|_H}\right) & \text{if } |x|_H > r+1. \end{cases} \tag{4.21}$$

It is immediate to check that K_r is Lipschitz continuous, together with its derivative, for each $r > 0$, and $K_r(x) = K(x)$ if $|x|_H \le r$.

THEOREM 4.13. *Under Hypotheses* 1, 2, *and* 4 *there exists $\mu_0 > 0$ such that for any $\lambda > \mu_0$ and $g \in C_b^1(H)$ there exists a unique solution $\varphi(\lambda, g) \in D(L)$ for the problem* (4.1).

Proof. For any $r > 0$ and $g \in C_b^1(H)$ we define $\varphi_r(\lambda, g)$ as the solution of the problem $\lambda\varphi - L\varphi + K_r(D\varphi) = g$. Due to Proposition 4.11 there exists $\mu_0 > 0$ such that for any $\lambda > \mu_0$

$$\sup_{r>0} \|D\varphi_r(\lambda, g)\|_0 \le \|Dg\|_0.$$

Thus, if we fix $r > \|g\|_1$, we have that $K_r(D\varphi_r(\lambda, g)) = K(D\varphi_r(\lambda, g))$, and then

$$\lambda\varphi_r(\lambda, g) - L\varphi_r(\lambda, g) + K(D\varphi_r(\lambda, g)) = g. \qquad \square$$

Remark 4.14. The operator N is dissipative. Actually, fix $\lambda > 0$ and $\varphi_1, \varphi_2 \in D(L)$, and define $g_i = \lambda\varphi_i - N(\varphi_i)$ for $i = 1, 2$. If we take $r \ge \max(\|\varphi_1\|_1, \|\varphi_2\|_1)$, we have

$$g_i = \lambda\varphi_i - L\varphi_i + K_r(D\varphi_i), \qquad i = 1, 2.$$

Thus we can apply Theorem 4.9 to the hamiltonian K_r, and we get

$$\|\varphi_1 - \varphi_2\|_0 \le \frac{1}{\lambda}\|g_1 - g_2\|_0,$$

so that N is dissipative. In particular, N is closable, and its closure $\bar{N}$ is m-dissipative, so that for any $\lambda > 0$ and $g \in C_b(H)$ there exists a unique solution to the problem

$$\lambda\varphi - \bar{N}(\varphi) = g.$$

5. Application to the control problem. Let $k : H \to (-\infty, +\infty]$ be a measurable mapping such that its Legendre transform

$$K(x) = \sup \{ -\langle x, y\rangle_H - k(y) ;\ y \in H \}, \qquad x \in H,$$

fulfills Hypothesis 4. It is possible to show that if k is strictly convex and continuously Fréchet differentiable, if

$$\lim_{|y|_H \to +\infty} \frac{k(y)}{|y|_H} = 0,$$

and if $Dk : H \to \mathcal{L}(H)$ has a continuous inverse which is Lipschitz continuous on bounded subsets of H, then Hypothesis 4 is verified. An easy example is given by $k(y) = |y|_H^2$.

For any $\lambda > 0$ and $g \in C_b(H)$ we consider the *cost functional*

$$J(x; z) = \mathbb{E} \int_0^{+\infty} e^{-\lambda t} \left[g(y(t)) + k(z(t)) \right] dt, \tag{5.1}$$

where $y(t) = y(t; x, z)$ is the unique solution of the system (3.1). The corresponding value function is defined as

$$V(x) = \inf \left\{ J(x; z) ;\ z \in L^2(\Omega; L^2(0, +\infty; H)) \text{ adapted} \right\}.$$

Our aim is to prove that if φ is the unique solution of the Hamilton–Jacobi equation (4.1), then $V(x) = \varphi(x)$ for any $x \in H$. To this purpose we first prove the following preliminary result.

LEMMA 5.1. *Assume Hypotheses* 1, 2, *and* 4. *If $\varphi = \varphi(\lambda, g)$ is the solution of the problem* (4.1) *in $C_b^1(H)$ and if $y(t) = y(t; x, z)$ is the solution of the controlled system* (3.1), *we have*

$$J(x; z) = \varphi(x) + \mathbb{E} \int_0^{+\infty} e^{-\lambda t} \left[K(D\varphi(y(t))) + \langle z(t), D\varphi(y(t))\rangle_H + k(z(t)) \right] dt. \tag{5.2}$$

Proof. If $r \geq \|D\varphi(\lambda, g)\|_0$ and if K_r is defined as in (4.21), then we have $K(D\varphi(x)) = K_r(D\varphi(x))$ for any $x \in H$, and the problem (4.1) can be rewritten as

$$\lambda\varphi - L\varphi + K_r(D\varphi) = g.$$

Now we fix a sequence $\{g_k\} \subset C_b^1(H)$ converging to g in $C_b(H)$ and for any $k, n \in \mathbb{N}$ and $\alpha > 0$ we denote by $\varphi_{\alpha,n}^k = \varphi_{\alpha,n}^k(\lambda + \mu, g_k + \mu\,\varphi(\lambda, g))$ the solution of the problem

$$(\lambda + \mu)\,\varphi - L_{\alpha,n}\varphi + K_{r,n}(D\varphi) = g_{k,n} + \mu\,\varphi_n(\lambda, g), \tag{5.3}$$

where $K_{r,n}(x) = K_r(P_n x)$, $g_{k,n}(x) = g_k(P_n x)$, and $\varphi_n(\lambda, g)(x) = \varphi(\lambda, g)(P_n x)$, and μ is some positive constant to be determined later. Since g_k and $\varphi(\lambda, g)$ are continuously differentiable, due to Lemma 4.8 we have that $\varphi_{\alpha,n}^k$ belongs to $C_b^2(H)$. Then, since $y_{\alpha,n}(t; x, z)$ is a strong solution of the problem (3.9), we can apply the Itô formula to the mapping $t \mapsto e^{-\lambda t}\varphi_{\alpha,n}^k(y_{\alpha,n}(t))$, and we get

$$d\left(e^{-\lambda t}\varphi_{\alpha,n}^k(y_{\alpha,n}(t))\right) = e^{-\lambda t}\left\langle D\varphi_{\alpha,n}^k(y_{\alpha,n}(t)), Q_n\, dw(t)\right\rangle_H$$

$$+e^{-\lambda t}\left((\mathcal{L}_{\alpha,n} - \lambda)\varphi_{\alpha,n}^k(y_{\alpha,n}(t)) + \left\langle P_n z(t), D\varphi_{\alpha,n}^k(y_{\alpha,n}(t))\right\rangle_H\right).$$

Recalling that $\varphi^k_{\alpha,n}$ is the solution of (5.3) and that (3.15) holds, we have

$$(\mathcal{L}_{\alpha,n}-\lambda)\varphi^k_{\alpha,n}=\mu\,\varphi^k_{\alpha,n}+K_{r,n}(D\varphi^k_{\alpha,n})-g_{k,n}-\mu\,\varphi_n(\lambda,g).$$

Then, by integrating with respect to $t\in[0,T]$ and by taking the expectation, we get

$$e^{-\lambda\,T}P^{\alpha,n}_T\varphi^k_{\alpha,n}-\varphi^k_{\alpha,n}=\mu\,\mathbb{E}\int_0^T e^{-\lambda\,t}\left(\varphi^k_{\alpha,n}-\varphi(\lambda,g)\right)(y_{\alpha,n}(t))\,dt$$

$$+\mathbb{E}\int_0^T e^{-\lambda\,t}\left(K_r(D\varphi^k_{\alpha,n}(y_{\alpha,n}(t)))-g_k(y_{\alpha,n}(t))+\left\langle z(t),D\varphi^k_{\alpha,n}(y_{\alpha,n}(t))\right\rangle_H\right)dt.$$

Due to (3.10) and (4.15), if μ is large enough, we can take the limit as n goes to infinity, and we get

$$e^{-\lambda\,T}P^{\alpha}_T\varphi^k_{\alpha}(x)-\varphi^k_{\alpha}(x)=\mu\,\mathbb{E}\int_0^T e^{-\lambda\,t}\left(\varphi^k_{\alpha}-\varphi(\lambda,g)\right)(y_{\alpha}(t))\,dt$$

$$=\mathbb{E}\int_0^T e^{-\lambda\,t}\left[K_r(D\varphi^k_{\alpha}(y_{\alpha}(t)))-g_k(y_{\alpha}(t))+\left\langle z(t),D\varphi^k_{\alpha}(y_{\alpha}(t))\right\rangle_H\right]dt,$$

where $\varphi^k_\alpha=\varphi^k_\alpha(\lambda+\mu,g_k+\mu\,\varphi(\lambda,g))$ is the solution of the problem

$$(\lambda+\mu)\varphi-L_\alpha\varphi+K_r(D\varphi)=g_k+\mu\,\varphi(\lambda,g).$$

By taking the limit as T goes to infinity, this yields

$$\begin{aligned}-\varphi^k_\alpha&=\mu\,\mathbb{E}\int_0^{+\infty} e^{-\lambda\,t}\left(\varphi^k_{\alpha}-\varphi(\lambda,g)\right)(y_{\alpha}(t))\,dt\\&+\mathbb{E}\int_0^{+\infty} e^{-\lambda\,t}\left[K_r(D\varphi^k_{\alpha}(y_{\alpha}(t)))-g_k(y_{\alpha}(t))+\left\langle z(t),D\varphi^k_{\alpha}(y_{\alpha}(t))\right\rangle_H\right]dt.\end{aligned}\tag{5.4}$$

We remark that for any $h,k\in\mathbb{N}$ and $\alpha>0$ we have

$$\varphi^k_\alpha-\varphi^h_\alpha=R(\lambda+\mu,L_\alpha)\left[g_k-g_h-\left(K_r(D\varphi^k_\alpha)-K_r(D\varphi^h_\alpha)\right)\right],$$

and then, due to (4.2),

$$\|\varphi^k_\alpha-\varphi^h_\alpha\|_1\le\rho(\lambda+\mu)\left(\|g_k-g_h\|_0+c_r\|D\varphi^k_\alpha-D\varphi^h_\alpha\|_0\right),$$

where c_r is the Lipschitz constant of K_r. Therefore, if μ is sufficiently large, we have $\rho(\lambda+\mu)c_r<1$, so that

$$\|\varphi^k_\alpha-\varphi^h_\alpha\|_1\le\frac{\rho(\lambda+\mu)}{1-\rho(\lambda+\mu)c_r}\|g_k-g_h\|_0.\tag{5.5}$$

This means that the sequence $\{\varphi^k_\alpha\}$ converges to some φ_α in $C^1_b(H)$. It is immediate to check that φ_α coincides with $\varphi_\alpha(\lambda+\mu,g+\mu\,\varphi(\lambda,g))$, and then, by taking the limit as k goes to infinity in (5.4), due to the dominated convergence theorem we can conclude that

$$\begin{aligned}-\varphi_\alpha&=\mu\,\mathbb{E}\int_0^{+\infty} e^{-\lambda\,t}\left(\varphi_{\alpha}-\varphi(\lambda,g)\right)(y_{\alpha}(t))\,dt\\&+\mathbb{E}\int_0^{+\infty} e^{-\lambda\,t}\left[K_r(D\varphi_{\alpha}(y_{\alpha}(t)))-g(y_{\alpha}(t))+\left\langle z(t),D\varphi_{\alpha}(y_{\alpha}(t))\right\rangle_H\right]dt.\end{aligned}\tag{5.6}$$

If $x \in E$ and $z \in L^p(\Omega; L^\infty(0,+\infty;H))$ with p as in the Proposition 3.2, we can use (3.7) and (4.13), and by taking the limit as α goes to zero we have

$$\varphi(x) + \mathbb{E}\int_0^{+\infty} e^{-\lambda t}\left[K(D\varphi(y(t))) - g(y(t)) + \langle z(t), D\varphi(y(t))\rangle_H\right] dt = 0.$$

Notice that here we have replaced K_r by K, as we fixed $r \geq \|D\varphi(\lambda,g)\|_0$. Since $\varphi \in C_b^1(H)$ and $y(t;x,z)$ depends continuously on $x \in H$ and $z \in L^2(\Omega; L^2(0,+\infty;H))$, the same identity holds for $x \in H$ and $z \in L^2(\Omega; L^2(0,+\infty;H))$. Then, recalling how $J(x;z)$ is defined, if we rearrange all terms, we get (5.2). □

THEOREM 5.2. *Assume that Hypotheses* 1, 2, *and* 4 *hold. Then there exists* μ_0 *such that for any* $\lambda > \mu_0$ *and* $g \in C_b^1(H)$ *the value function* V *corresponding to the cost functional* (5.1) *coincides with the solution* $\varphi(\lambda,g)$ *of the Hamilton–Jacobi equation* (4.1).

Moreover, for any $x \in E$ *we have*

$$V(x) = \lim_{\alpha\to 0} \min\left\{ J_\alpha(x;z)\,;\ z \in L^2(\Omega; L^2(0,+\infty;H)) \text{ adapted}\right\},$$

where $\{J_\alpha(x,z)\}$ *is a sequence of cost functionals which admit unique optimal controls and states and whose value functions* V_α *coincide with the solution of the problems*

$$(\lambda+\lambda_0)\varphi - L_\alpha\varphi + K_r(D\varphi) = g + \lambda_0\,\varphi(\lambda,g)$$

for some $\lambda_0 > 0$ *large enough and* $r \geq \|D\varphi(\lambda,g)\|_0$.

Proof. In Theorem 4.13 we have seen that, if $\lambda > \mu_0$ and $g \in C_b^1(H)$, there exists a unique solution $\varphi(\lambda,g) \in C_b^1(H)$ for (4.1). Due to (5.2) and to the definition of K, we have that $V(x) \geq \varphi(\lambda,g)(x)$ for any $x \in H$. Now we try to prove the opposite inequality. To this purpose we proceed by approximation.

We fix $r \geq \|D\varphi(\lambda,g)\|_0$, and for any $\alpha > 0$ we define the cost functional

$$J_\alpha(x;z) = \mathbb{E}\int_0^{+\infty} e^{-\lambda t}\left[g(y_\alpha(t;x,z)) + k(z(t))\right] dt$$

$$+\lambda_0\,\mathbb{E}\int_0^{+\infty} e^{-\lambda t}\left[(\varphi(\lambda,g) - \varphi_\alpha)\,(y_\alpha(t;x,z))\right] dt$$

$$+\mathbb{E}\int_0^{+\infty} e^{-\lambda t}\left[K(D\varphi_\alpha(y_\alpha(t;x,z))) - K_r(D\varphi_\alpha(y_\alpha(t;x,z)))\right] dt,$$

where $\varphi_\alpha = \varphi_\alpha(\lambda+\lambda_0, g+\lambda_0\,\varphi(\lambda,g))$ is the solution of the problem

$$(\lambda+\lambda_0)\varphi - L_\alpha\varphi + K_r(D\varphi) = g + \lambda_0\,\varphi(\lambda,g),$$

and λ_0 is the constant introduced in Proposition 4.3 corresponding to the hamiltonian K_r. We denote by $V_\alpha(x)$ the corresponding value function. Thanks to (5.6) we easily have that $V_\alpha(x) \geq \varphi_\alpha(x)$ for any $x \in H$. In fact, it is possible to show that $V_\alpha(x) = \varphi_\alpha(x)$. Indeed, for each $x \in H$ the function

$$H \to \mathbb{R}, \qquad z \mapsto -\langle z, D\varphi_\alpha(x)\rangle_H - k(z)$$

attains its maximum at $z = -DK(D\varphi_\alpha(x))$. Then, if we show that the *closed loop equation*

$$dy(t) = \left[Ay(t) + F_\alpha(y(t)) - DK(D\varphi_\alpha(y(t)))\right] dt + Q\,dw(t), \qquad y(0) = x, \tag{5.7}$$

has a unique adapted solution $y^\star_\alpha(t)$, we have that for the control

$$z^\star_\alpha(t) = -DK(D\varphi_\alpha(y^\star_\alpha(t)))$$

it holds that $J_\alpha(x, z^\star_\alpha) = \varphi_\alpha(x)$. This means that $V_\alpha(x) = \varphi_\alpha(x)$, and there exists a unique optimal control and a unique optimal state for the minimizing problem corresponding to the cost functional $J_\alpha(x; z)$.

If $g \in C^1_b(H)$, then due to Lemma 4.8 $\varphi_\alpha \in C^2_b(H)$, so that the mapping

$$U_\alpha : H \to H, \qquad x \mapsto -DK(D\varphi_\alpha(x))$$

is Lipschitz continuous. This implies that the closed loop equation admits a unique solution.

For any $\alpha > 0$ the optimal control relative to the functional $J_\alpha(x; z)$ is $z^\star_\alpha(t) = -DK(D\varphi_\alpha(y^\star_\alpha(t)))$. According to Proposition 4.11 we have

$$\|D\varphi_\alpha\|_0 \le \|Dg\|_0 + \lambda_0 \|D\varphi(\lambda, g)\|_0,$$

and then, since DK is bounded on bounded sets, there exists $R > 0$ such that

$$\sup_{\alpha>0} \sup_{t\ge 0} |z^\star_\alpha(t)|_H = R, \qquad \mathbb{P}\text{-a.s.}$$

This implies that

$$V_\alpha(x) = \inf \left\{ J_\alpha(x; z) \,:\, z \in \mathcal{M}^2_R \right\},$$

where $\mathcal{M}^2_R$ is the subset of admissible controls introduced in (3.8).

Now, recalling Proposition 4.6, we have that for any $x \in E$

$$\lim_{\alpha\to 0} V_\alpha(x) = \lim_{\alpha\to 0} \varphi_\alpha(\lambda + \lambda_0, g + \lambda_0\varphi(\lambda, g))(x) = \varphi(\lambda, g)(x).$$

Thus, if we show that

$$\lim_{\alpha\to 0} \sup_{z\in \mathcal{M}^2_R} |J_\alpha(x; z) - J(x; z)| = 0, \tag{5.8}$$

it immediately follows that $V(x) = \varphi(\lambda, g)(x)$ for $x \in E$.

Due to Proposition 3.2, we have that

$$\lim_{\alpha\to 0} \mathbb{E}\, |g(y_\alpha(t; x, z)) - g(y(t; x, z))| = 0,$$

uniformly for (t, x) in bounded sets of $[0, +\infty) \times E$ and $z \in \mathcal{M}^2_R$. Hence, if we fix $\epsilon > 0$ and $M > 0$ such that

$$\int_M^{+\infty} e^{-\lambda\, t}\, dt \le \frac{\epsilon}{2\, \|g\|_0},$$

we have

$$\mathbb{E} \int_0^{+\infty} e^{-\lambda\, t}\, [g(y_\alpha(t; x, z)) - g(y(t; x, z))]\, dt$$

$$\le \epsilon + \int_0^M e^{-\lambda\, t} \mathbb{E}\, |g(y_\alpha(t; x, z)) - g(y(t; x, z))|\, dt,$$

so that, due to the arbitrariness of $\epsilon > 0$,

$$\lim_{\alpha\to 0} \sup_{z\in \mathcal{M}_R^2} \mathbb{E}\int_0^{+\infty} e^{-\lambda t}\left[g(y_\alpha(t;x,z)) - g(y(t;x,z))\right]\,dt = 0.$$

Thanks to Lemma 4.8, we have that for $j = 0, 1$

$$\lim_{\alpha\to 0} \sup_{|x|_E\le R} \left|D^j\left(\varphi_\alpha - \varphi\right)(x)\right|_{\mathcal{L}^j(H)} = 0. \tag{5.9}$$

Moreover, thanks to (3.6),

$$\sup_{z\in \mathcal{M}_R^2} \sup_{t\in[0,T]} |y_\alpha(t;x,z)|_E < +\infty, \qquad \mathbb{P}\text{-a.s.}$$

for any $T > 0$. Thus, by using the same arguments as above, we have

$$\lim_{\alpha\to 0} \sup_{z\in \mathcal{M}_R^2} \mathbb{E}\int_0^{+\infty} e^{-\lambda t}\left[(\varphi_\alpha - \varphi)(y_\alpha(t;x,z))\right]\,dt = 0.$$

Finally, since the sequence $\{\varphi_\alpha\}$ is bounded in $C_b^1(H)$, recalling that K and K_r are bounded on bounded sets and $K(D\varphi(x)) = K_r(D\varphi(x))$ for any $x \in H$, by using (5.9) and by arguing as above, we have

$$\lim_{\alpha\to 0} \sup_{z\in \mathcal{M}_R^2} \mathbb{E}\int_0^{+\infty} e^{-\lambda t}\left[K(D\varphi_\alpha(y_\alpha(t;x,z))) - K_r(D\varphi_\alpha(y_\alpha(t;x,z)))\right]\,dt = 0.$$

Therefore, we can conclude that (5.8) holds for any $x \in E$, and then $V(x) = \varphi(x)$ for $x \in E$.

Now assume that $x \in H$. We fix a sequence $\{x_n\} \subset E$ converging to x in H. For each $n \in \mathbb{N}$ we have $V(x_n) = \varphi(x_n)$ and

$$J(x_n;z) - J(x;z) = \mathbb{E}\int_0^{+\infty} e^{-\lambda t}\left[g(y(t;x_n,z)) - g(y(t;x,z))\right]\,dt.$$

Then, due to (3.4), if $\varphi \in C_b^1(H)$, we easily get

$$\lim_{n\to+\infty} \sup_{z\in \mathcal{M}_R^2} J(x_n;z) = J(x;z),$$

so that we can conclude that $V(x) = \varphi(x)$ for any $x \in H$. □

We have seen that if we assume the hamiltonian K to be Lipschitz continuous, then for any $\lambda > 0$ and $g \in C_b(H)$ there exists a unique solution $\varphi(\lambda, g)$ in $D(L) \subset C_b^1(H)$ to the problem (4.1). This allows us to have a stronger version of the previous theorem in the case of Lipschitz K.

THEOREM 5.3. *Assume that Hypotheses* 1, 2, *and* 3 *hold. Then for any* $\lambda > 0$ *and* $g \in \mathrm{Lip}_b(H)$ *the value function* V *corresponding to the cost functional* (5.1) *coincides with the solution* $\varphi(\lambda, g)$ *of the Hamilton–Jacobi equation* (4.1).

Moreover, for any $x \in E$ *we have*

$$V(x) = \lim_{\alpha\to 0} \min\left\{ J_\alpha(x;z)\,;\ z \in L^2(\Omega; L^2(0,+\infty;H)) \text{ adapted}\right\},$$

where $\{J_\alpha(x,z)\}$ is a sequence of cost functionals which admit unique optimal controls and states and whose value functions V_α coincide with the solution of the problems

$$(\lambda+\lambda_0)\varphi - L_\alpha\varphi + K(D\varphi) = g + \lambda_0\,\varphi(\lambda,g)$$

for some $\lambda_0 > 0$.

Proof. By arguing as in the proof of the previous theorem, we have the thesis for any $g \in C_b^1(H)$ and $\lambda > 0$. Thus, in order to conclude, we have to show that for any $g \in \mathrm{Lip}_b(H)$ the approximating closed loop equation

$$du(t) = [Au(t) + F_\alpha(u(t)) - DK(D\varphi_\alpha(u(t)))]\,dt + Q\,dw(t), \qquad u(0) = x,$$

admits a unique adapted solution $u_\alpha^\star(t)$.

If $g \in \mathrm{Lip}_b(H)$, we can find a bounded sequence $\{g_k\} \subset C_b^1(H)$ converging to g in $C_b(H)$. For each $k \in \mathbb{N}$ there exists a unique solution $\varphi_{\alpha,k}$ for the Hamilton–Jacobi problem

$$(\lambda+\lambda_0)\,\varphi - L_\alpha\varphi + K_r(D\varphi) = g_k + \lambda_0\varphi(\lambda,g).$$

Then, since $g_k + \lambda_0\varphi(\lambda,g) \in C_b^1(H)$, as proved above, the corresponding closed loop equation has a unique solution $y^\star_{\alpha,k}(t)$. If we show that for any $T > 0$ the sequence $\{y^\star_{\alpha,k}\}$ converges to some $y^\star_\alpha$ in $C([0,T];H)$, $\mathbb{P}$-a.s. and in mean-square, then we easily have that $y^\star_\alpha$ is the solution of the closed loop equation (5.7).

For $k,h \in \mathbb{N}$ we define $v_\alpha^{k,h}(t) = y^\star_{\alpha,k}(t) - y^\star_{\alpha,h}(t)$. We have that $v_\alpha^{k,h}$ is the solution of the problem

$$\begin{aligned}\frac{dv}{dt}(t) = \quad & Av(t) + F_\alpha(y^\star_{\alpha,k}(t)) - F_\alpha(y^\star_{\alpha,h}(t)) - DK(D\varphi_{\alpha,k}(y^\star_{\alpha,k}(t))) \\ & + DK(D\varphi_{\alpha,h}(y^\star_{\alpha,h}(t))), \qquad v(0) = 0.\end{aligned}$$

By multiplying each side by $v_\alpha^{k,h}(t)$ and recalling (2.3) and (2.4), we have

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}|v_\alpha^{k,h}(t)|_H^2 \le c\,|v_\alpha^{k,h}(t)|_H^2 \\ &+|DK(D\varphi_{\alpha,k}(y^\star_{\alpha,k}(t))) - DK(D\varphi_{\alpha,h}(y^\star_{\alpha,h}(t)))|_H|v_\alpha^{k,h}(t)|_H.\end{aligned}$$

Since DK is locally Lipschitz continuous and according to Lemma 4.11 we have

$$\sup_{k\in\mathbb{N}}\|D\varphi_{\alpha,k}\|_0 \le \sup_{k\in\mathbb{N}}(\|Dg_k\|_0 + \lambda_0\,\|D\varphi\|_0) < \infty,$$

we obtain

$$\begin{aligned}&|DK(D\varphi_{\alpha,k}(y^\star_{\alpha,k}(t))) - DK(D\varphi_{\alpha,h}(y^\star_{\alpha,h}(t)))|_H \\ &\le c\,\left|D\varphi_{\alpha,k}(y^\star_{\alpha,k}(t)) - D\varphi_{\alpha,h}(y^\star_{\alpha,h}(t))\right|_H.\end{aligned}$$

For each $x,y \in H$ we have

$$|D\varphi_{\alpha,k}(x) - D\varphi_{\alpha,k}(y)|_H \le c\,\|\varphi_{\alpha,k}\|_2|x-y|_H,$$

and then, since the sequence $\{g_k\}$ is bounded in $C_b^1(H)$, from (4.16) it follows that

$$|D\varphi_{\alpha,k}(y^\star_{\alpha,k}(t)) - D\varphi_{\alpha,k}(y^\star_{\alpha,h}(t))|_H \le c_\alpha\,|v_\alpha^{k,h}(t)|_H.$$

Moreover, if λ_0 is large enough, due to (5.5) we have

$$\|D\varphi_{\alpha,k} - D\varphi_{\alpha,h}\|_0 \le c_\alpha \, \|g_k - g_h\|_0.$$

Therefore, we can conclude that

$$|DK(D\varphi_{\alpha,k}(y^\star_{\alpha,k}(t))) - DK(D\varphi_{\alpha,h}(y^\star_{\alpha,h}(t)))|_H \le c_\alpha \, |v^{k,h}_\alpha(t)|_H + c_\alpha \, \|g_k - g_h\|_0,$$

so that from the Young inequality we have

$$\frac{1}{2}\frac{d}{dt}|v^{k,h}_\alpha(t)|^2_H \le c_\alpha \, |v^{k,h}_\alpha(t)|^2_H + c_\alpha \, \|g_k - g_h\|^2_0.$$

By the Gronwall lemma this yields

$$\sup_{t\in[0,T]} |y^\star_{\alpha,k}(t) - y^\star_{\alpha,h}(t)|_H \le c_T \|g_k - g_h\|_0, \qquad \mathbb{P}\text{-a.s.},$$

for some constant C_T. Thus $y^\star_{\alpha,k}$ converges to some $y^\star_\alpha$ in $C([0,T];H)$, $\mathbb{P}$-a.s and in mean-square, and it is not difficult to check that $y^\star_\alpha$ is the solution of the closed loop equation corresponding to the datum g. □

By proceeding as in [9, Theorem 7.3] it is possible to show that when the space dimension d equals 1, under suitable assumptions there exist an optimal control and the corresponding optimal state.

THEOREM 5.4. *Assume that the space dimension d equals* 1.

1. *If the constant m in Hypothesis* 1 *is less than or equal to* 1, *then there exists a unique optimal control for the minimizing problem associated with the functional J. Furthermore, the optimal control $z^\star$ is related to the corresponding optimal state $y^\star$ by the feedback formula*

$$z^\star(t) = -DK[DV(y^\star(t))], \qquad t \in [0,T].$$

2. *If DK can be extended as a Lipschitz continuous mapping from $E^\star$ into itself, then the same conclusion holds for any $x \in E$.*

REFERENCES

[1] V. BARBU AND G. DA PRATO, *Hamilton-Jacobi Equations in Hilbert Spaces*, Research Notes in Mathematics 86, Pitman, Boston, 1983.

[2] P. CANNARSA AND G. DA PRATO, *Second-order Hamilton–Jacobi equations in infinite dimensions*, SIAM J. Control Optim., 29 (1991), pp. 474–492.

[3] P. CANNARSA AND G. DA PRATO, *Direct solution of a second-order Hamilton-Jacobi equation in Hilbert spaces*, in Stochastic Partial Differential Equations and Applications, Pitman Res. Notes Math. Ser. 268, G. Da Prato and L. Tubaro, eds., Longman, Harlow, UK, 1992, pp. 72–85.

[4] S. CERRAI, *A Hille Yosida theorem for weakly continuous semigroups*, Semigroup Forum, 49 (1994), pp. 349–367.

[5] S. CERRAI, *Elliptic and parabolic equations in $\mathbb{R}^n$ with coefficients having polynomial growth*, Comm. Partial Differential Equations, 21 (1996), pp. 281–317.

[6] S. CERRAI, *Differentiability with respect to initial datum for solutions of SPDE'S with no Fréchet differentiable drift term*, Comm. Appl. Anal., 2 (1998), pp. 249–270.

[7] S. CERRAI, *Smoothing properties of transition semigroups relative to SDE's with values in Banach spaces*, Probab. Theory Related Fields, 113 (1999), pp. 85–114.

[8] S. CERRAI, *Differentiability of Markov semigroups for stochastic reaction-diffusion equations and applications to control*, Stochastic Process. Appl., 83 (1999), pp. 15–37.

[9] S. CERRAI, *Optimal control problems for stochastic reaction-diffusion systems with non-Lipschitz coefficients*, SIAM J. Control Optim., 39 (2001), pp. 1779–1816.

[10] P. L. Chow and J. L. Menaldi, *Infinite dimensional Hamilton-Jacobi equations in Gauss-Sobolev spaces*, J. Nonlinear Anal., 29 (1997), pp. 415–426.
[11] M. G. Crandall, H. Ishii, and P. L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.
[12] G. Da Prato and A. Debussche, *Control of the stochastic Burgers model of turbulence*, SIAM J. Control Optim., 37 (1999), pp. 1123–1149.
[13] G. Da Prato and A. Debussche, *Dynamic programming for the stochastic Burgers equation*, Ann. Mat. Pura Appl. (4), 178 (2000), pp. 143–174.
[14] G. Da Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Cambridge University Press, Cambridge, UK, 1992.
[15] E. B. Davies, *Heat Kernels and Spectral Theory*, Cambridge University Press, Cambridge, UK, 1989.
[16] T. E. Duncan, B. Maslowski, and B. Pasik-Duncan, *Ergodic boundary point control of stochastic semilinear systems*, SIAM J. Control Optim., 36 (1998), pp. 1020–1047.
[17] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.
[18] M. Freidlin, *Markov Processes and Differential Equations: Asymptotic Problems*, Lectures Math. ETH Zürich, Birkhäuser-Verlag, Basel, 1996.
[19] B. Goldys and B. Maslowski, *Ergodic control of semilinear stochastic equations and the Hamilton-Jacobi equation*, J. Math. Anal. Appl., 234 (1999), pp. 592–631.
[20] F. Gozzi, *Regularity of solutions of a second order Hamilton-Jacobi equation and application to a control problem*, Comm. Partial Differential Equations, 20 (1995), pp. 775–826.
[21] F. Gozzi, *Global regular solutions of second order Hamilton-Jacobi equations in Hilbert spaces with locally Lipschitz nonlinearities*, J. Math. Anal. Appl., 198 (1996), pp. 399–443.
[22] F. Gozzi and E. Rouy, *Regular solutions of second-order stationary Hamilton-Jacobi equations*, J. Differential Equations, 130 (1996), pp. 201–234.
[23] T. Haverneanu, *Existence for the dynamic programming equation of control diffusion processes in Hilbert spaces*, Nonlinear Anal., 9 (1985), pp. 619–629.
[24] P. L. Lions, *Viscosity solutions of fully nonlinear second order equations and optimal control in infinite dimensions* I: *The case of bounded stochastic evolutions*, Acta Math., 161 (1998), pp. 243–278.
[25] P. L. Lions, *Viscosity solutions of fully nonlinear second order equations and optimal control in infinite dimensions* II: *Optimal control of Zakai's equation*, in Stochastic Partial Differential Equations and Applications, Lecture Notes in Math. 1390, G. DaPrato and L. Tubaro, eds., Springer-Verlag, Berlin, 1989, pp. 147–170.
[26] P. L. Lions, *Viscosity solutions of fully nonlinear second order equations and optimal control in infinite dimensions* III: *Uniqueness of viscosity solutions for general second order equations*, J. Funct. Anal., 86 (1989), pp. 1–18.
[27] A. Lunardi, *Analytic Semigroups and Optimal Regularity in Parabolic Problems*, Birkäuser-Verlag, Basel, 1995.
[28] A. Swiech, *Viscosity Solutions of Fully Nonlinear Partial Differential Equations with Unbounded Terms in Infinite Dimensions*, Ph.D. thesis, University of California Santa Barbara, Santa Barbara, CA, 1993.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 3, pp. 853–881

VIABILITY KERNELS AND CAPTURE BASINS OF SETS UNDER DIFFERENTIAL INCLUSIONS*

JEAN-PIERRE AUBIN†

Abstract. This paper provides a characterization of viability kernels and capture basins of a target viable in a constrained subset as a *unique* closed subset between the target and the constrained subset satisfying tangential conditions or, by duality, normal conditions. It is based on a method devised by Hélène Frankowska for characterizing the value function of an optimal control problem as generalized (contingent or viscosity) solutions to Hamilton–Jacobi equations. These abstract results, interesting by themselves, can be applied to epigraphs of functions or graphs of maps and happen to be very efficient for solving other problems, such as stopping time problems, dynamical games, boundary-value problems for systems of partial differential equations, and impulse and hybrid control systems, which are the topics of other companion papers.

Key words. differential inclusion, control system, viability kernel, capture basin, Hamilton–Jacobi equations, local viability, backward invariance

AMS subject classifications. 49A52, 49J24, 49K24, 49L25

PII. S036301290036968X

1. Introduction.

We consider in this paper a differential inclusion $x' \in F(x)$ (summarizing the dynamics of a control system) and two subsets C and K of a finite dimensional vector space X such that $C \subset K$. Here, K is regarded as a *constrained subset* in which the solution must evolve until possibly reaching the subset C regarded as a *target*.

DEFINITION 1.1.

1. *The subset* $\mathrm{Viab}_F(K)$ *of initial states* $x_0 \in K$ *such that at least one solution* $x(\cdot)$ *to differential inclusion* $x' \in F(x)$ *starting at* x_0 *is viable in* K *for all* $t \geq 0$ *is called the* viability kernel *of* K *under* F. *A subset* K *is a* repeller *under* F *if its viability kernel is empty.*
2. *The subset* $\mathrm{Capt}_F^K(C)$ *of initial states* $x_0 \in K$ *such that* C *is reached in finite time before possibly leaving* K *by at least one solution* $x(\cdot)$ *to differential inclusion* $x' \in F(x)$ *starting at* x_0 *is called the* viable-capture basin *of* C *in* K, *and* $\mathrm{Capt}_F(C) := \mathrm{Capt}_F^X(C)$ *is said to be the* capture basin *of* C.
3. *The subset*

$$\mathrm{Viab}_F(K, C) \ := \ \mathrm{Viab}_F(K \backslash C) \cup \mathrm{Capt}_F^K(C)$$

of initial states $x_0 \in K$ *such that at least one solution* $x(\cdot)$ *to differential inclusion* $x' \in F(x)$ *starting at* x_0 *is viable in* K *for all* $t \geq 0$ *or viable in* K *until it reaches* C *in finite time is called the* viability kernel of K with target C *under* F.

A subset $C \subset K$ *is said to be* isolated in K by F *if it coincides with its viability kernel* K *with target* C*:*

$$\mathrm{Viab}_F(K, C) \ = \ C.$$

*Received by the editors March 20, 2000; accepted for publication (in revised form) November 20, 2000; published electronically September 28, 2001.
http://www.siam.org/journals/sicon/40-3/36968.html
†Université Paris-Dauphine, Centre de Recherche Viabilité, Jeux, Contrôle, F-75775 Paris Cedex 16, France (J.P.Aubin@wanadoo.fr, aubin@viab.dauphine.fr).

The subset $\mathrm{Env}_F(C) := \mathrm{Capt}_{-F}(C)$ is known under various names such as *invariance envelope* or *accessibility map* or *controlled map* of C. (See [45] for properties of invariance envelopes under Lipschitz maps and [6, 8, 9] for Marchaud maps.) Henri Poincaré introduced the concept of *shadow* (in French, *ombre*) of K, which is the set of initial points of K from which (all) solutions leave K in finite time. It is thus equal to the complement $K\backslash\mathrm{Viab}_F(K)$ of the viability kernel of K, which has been introduced in the context of differential inclusions in [1]. The concept of *viability kernel with a target* by a Lipschitz set-valued map has been introduced and studied in [48], where the viability kernel algorithm designed in [50] (see also the survey [31]) has been extended for approximating the viability kernel with a target.

One could regard the viability kernel $\mathrm{Viab}(K)$ of K as the viability kernel $\mathrm{Viab}(K, \emptyset)$ of K with the empty set as a target:

$$\mathrm{Viab}(K) \;=\; \mathrm{Viab}(K, \emptyset) \;\text{ and }\; \mathrm{Capt}^K(\emptyset) \;=\; \emptyset.$$

Therefore, the viability kernel $\mathrm{Viab}(K, C)$ of K with target C coincides with the capture basin $\mathrm{Capt}^K(C)$ of C viable in K whenever the viability kernel $\mathrm{Viab}(K\backslash C)$ is empty, i.e., whenever $K\backslash C$ is a repeller:

$$\mathrm{Viab}(K\backslash C) \;=\; \emptyset \;\Rightarrow\; \mathrm{Viab}(K, C) \;=\; \mathrm{Capt}^K(C).$$

This happens, in particular, when K is a repeller, or when the viability kernel $\mathrm{Viab}(K)$ of K is contained in the target C.

Consequently, the concept of viability kernel with a target allows us to study *both the viability kernel of a closed subset and the viable-capture basin of a target.*

These subsets can be characterized in diverse ways through tangential conditions. We recall that the contingent cone $T_L(x)$ to $L \subset X$ at $x \in L$ is the set of directions $v \in X$ such that there exist sequences $h_n > 0$ converging to 0 and v_n converging to v satisfying $x + h_n v_n \in K$ for every n.

One of our objectives is to prove the following characterizations of the viability kernels and capture basins.

THEOREM 1.2. *Let us assume that F is Marchaud and that the target $C \subset K$ and K are closed. The viability kernel* $\mathrm{Viab}_F(K, C)$ *of the subset K with target C is*

1. *the* largest *closed subset D satisfying $C \subset D \subset K$ and*

$$D\backslash C \text{ is locally viable under } F \;(\forall\, x \in D\backslash C,\; F(x) \cap T_D(x) \neq \emptyset).$$

2. *If, furthermore, K is assumed to be backward invariant under F and F to be Lipschitz, the viability kernel* $\mathrm{Viab}_F(K, C)$ *is the* unique *closed subset $D \subset K$ satisfying the following.*

$$(1.1)\quad \begin{cases} \text{(i)} & D\backslash C \text{ is locally viable under } F \;(\forall\, x \in D\backslash C,\; F(x) \cap T_D(x) \neq \emptyset), \\ \text{(ii)} & D \text{ is backward invariant under } F \;(\forall\, x \in D,\; F(x) \subset -T_D(x)), \\ \text{(iii)} & K\backslash D \text{ is a repeller.} \end{cases}$$

The uniqueness properties of the viability kernel and the viable-capture basins are obtained thanks to the *Frankowska method*, consisting in introducing (local) backward invariance together with (local) forward viability of subsets. Indeed, Hélène Frankowska did point out in [39, 40] the backward invariance and local forward viability properties of the epigraph of the value function of an optimal control problem. She proved that the epigraph of the value function of an optimal control problem—assumed to be only lower semicontinuous—is backward invariant and viable under a

(natural) auxiliary system. It allowed her to characterize the value functions as unique solutions of contingent inequalities and, by duality, to obtain lower semicontinuous (or bilateral) solutions to Hamilton–Jacobi partial differential equations, obtained by other methods in [19]. (See also [18] for more details on this point of view.) Furthermore, when the value function is continuous, she proved that its epigraph is viable and its hypograph invariant [35, 36, 37, 38]. By duality, she proved that the latter property is equivalent to the fact that the value function is a viscosity solution of the associated Hamilton–Jacobi equation in the sense of Crandall and Lions in [32]. This *epigraphical approach* in the field of Hamilton–Jacobi equations has since been taken up by other authors.

Actually, we can spare the assumption that K is backward invariant in the above theorem if we are ready to trade the property that D is backward invariant with the weaker property that D satisfies $\mathrm{Capt}_F^K(D) = D$. Indeed, we shall derive this theorem from Theorem 4.4 below, which does not assume that K is backward viable.

Not only is the concept of the viability kernel naturally important in the framework of economic models and biological problems having motivated viability theory in the first place, but it happens that the notions of equilibria, of absorbing sets, of basins of absorption, of attractors, of "permanence," of "fluctuation," of "Lyapunov stability," of optimal Lyapunov functions, and of value function of an intertemporal optimization problem as well as other dynamical features can be studied by using the concept of the *viability kernel* as a mathematical tool (see [2, 3, 4, 5] for applications and further references).

The concept of the viable-capture basin also plays a fundamental role for solving first-order partial differential equations (see [12, 13, 14, 15, 16], chapter 8 of [2], [11] without boundary conditions, and [6, 7, 8, 9] for the Dirichlet boundary-value problems for such systems). Finally, the viability kernel algorithm allows us to compute the viability kernel (see [31, 47, 50]). Nonemptiness of the viability kernel is studied in [22, 23, 24]. Extension of this concept to impulse and hybrid control systems can be found in [17].

We shall conclude this paper by describing (without proofs that will be given in a forthcoming companion paper) an application of these results to optimal discounted intertemporal control. Consider the evolution of a control system with (multivalued) feedbacks:

$$\begin{cases} \text{(i)} & x'(t) = f(x(t), u(t)), \\ \text{(ii)} & u(t) \in P(x(t)), \end{cases}$$

where the state $x(\cdot)$ ranges over a finite dimensional vector-space X and the control $u(\cdot)$ ranges over another finite dimensional vector-space $\mathcal{M}$. The problem is to minimize a functional of the form

$$\begin{cases} J(t, x; (x(\cdot), u(\cdot))) \\ \\ := \; e^{\int_0^t \mathbf{m}(x(s),u(s))ds}\mathbf{c}(T-t, x(t)) + \displaystyle\int_0^t e^{\int_0^\tau \mathbf{m}(x(s),u(s))ds}\mathbf{l}(x(\tau), u(\tau))d\tau \end{cases}$$

over the set $\mathcal{S}(x)$ of solutions $(x(\cdot), u(\cdot))$ to a control system

$$V(T, x) \; := \inf_{(x(\cdot),u(\cdot))\in\mathcal{S}(x)} \inf_{t\in[0,T]} J(t, x; (x(\cdot), u(\cdot)))$$

or,

$$W(T, x) \; := \inf_{(x(\cdot),u(\cdot))\in\mathcal{S}(x)} \sup_{t\in[0,T]} J(t, x; (x(\cdot), u(\cdot))).$$

The connection between these problems and the basic viability theorems is simple. For instance, the epigraph of the value function is the capture basin of the epigraph of the cost function $\mathbf{c}$ under the auxiliary control system governed by the dynamics

$$g(x,y,u) \;=\; (f(x,u), -\mathbf{m}(x,y)y - \mathbf{l}(x,u)),$$

viable in the epigraph of an adequate function. This being checked, it will be sufficient to translate the properties of capture basins stated in Theorem 1.2 in terms of value functions, the tangential conditions characterizing capture basins becoming the Hamilton–Jacobi–Bellman variational inequalities of which the value function is an (adequately generalized) solution. It is enough to observe that the contingent cone to the epigraph of a function is, by definition, the epigraph of the contingent epiderivative of this function.

When we are studying the viability kernels with targets under differential inclusions, we observe that they are not specific to differential inclusions. They involve only few properties[1] of the solution map $\mathcal{S}$ associating with any initial state x the set $\mathcal{S}(x)$ of pairs $t \mapsto (x(t), u(t))$ that are solutions to the above control system starting at x at initial time 0. These properties of the solution map are common to other control problems, such as

1. control problems with memory (see the contributions of [42, 43], some of them being presented in [2])—previously known under the name of functional control problems, the new fashion calling them "path dependent control systems,"
2. parabolic type partial differential inclusions (see the contributions of [51, 52, 53, 54, 55], some of them being presented in [2])—also known as distributed control systems;
3. "mutational equations" governing the evolution in metric spaces, including "morphological equations" governing the evolution of sets (see [4], for instance).

Although these problems are not covered in this paper by lack of place, we shall make another step in abstraction by gathering these common properties of the solution map under the name of *evolutionary systems* and study the properties of viability kernels with targets in this general framework. In the case of differential inclusions, we shall use the viability and invariance theorems for characterizing them in terms of tangential conditions.

The paper is organized as follows. Section 2 introduces evolutionary systems. The third section defines hitting and exit functions. Viability kernels and capture basins are defined and characterized in section 4 for general evolutionary systems. Their characterizations in terms of tangential conditions or, by duality, in terms of normal conditions, are provided in the fifth section. The sixth provides useful stability results. The last section summarizes the applications of the above theorems to optimal control and stopping time problems.

2. Evolutionary systems.

2.1. Definition of evolutionary systems. The following results dealing with viability kernel and capture basins are valid for any *evolutionary system* described by a set-valued map $\mathcal{S}$ mapping some topological space X (most often, a topological

[1] These are the translation and concatenation properties of the set-valued map $x \rightsquigarrow \mathcal{S}(x)$, as well as continuity properties of this set-valued map.

vector-space) to the space $\mathcal{C}(0,\infty;X)$ of continuous functions $x(\cdot)$ from $\mathbf{R}_+$ to X, supplied with the topology of uniform convergence on compact intervals.

It can be the solution map associated with a differential inclusion $x' \in F(x)$ on a finite dimensional vector space X, with a differential inclusion with memory $x'(t) \in F(T(t)x)$ or with a mutational equation $\overset{\circ}{x} \ni f(x)$ on metric spaces.

DEFINITION 2.1. *An evolutionary system is a set-valued map* $\mathcal{S} : X \rightsquigarrow \mathcal{C}(0,\infty;X)$ *satisfying the following.*

1. *The* translation property. *Let* $x(\cdot) \in \mathcal{S}(x)$. *Then for all* $T \geq 0$, *the function* $y(\cdot)$ *defined by* $y(t) := x(t+T)$ *is a solution* $y(\cdot) \in \mathcal{S}(x(T))$ *starting at* $x(T)$.
2. *The* concatenation property. *Let* $x(\cdot) \in \mathcal{S}(x)$ *and* $T \geq 0$. *Then for every* $y(\cdot) \in \mathcal{S}(x(T))$, *the function* $z(\cdot)$, *defined by*

$$z(t) := \begin{cases} x(t) & \text{if} \quad t \in [0,T], \\ y(t-T) & \text{if} \quad t \geq T, \end{cases}$$

belongs to $\mathcal{S}(x)$.

We shall associate with $\mathcal{S}$ *its* backward evolutionary system $\mathcal{S}_- : X \rightsquigarrow \mathcal{C}(0,\infty;X)$ *defined by* $y(\cdot) \in \mathcal{S}_-(x)$ *if and only if there exists a solution* $z(\cdot) \in \mathcal{S}(x)$ *such that for every* $T \geq 0$, *the function* $x(\cdot)$, *defined by*

$$x(t) := \begin{cases} y(T-t) & \text{if} \quad t \in [0,T], \\ z(t-T) & \text{if} \quad t \geq T, \end{cases}$$

belongs to $\mathcal{S}(x)$.

We observe that $\mathcal{S}_{-\,-} = \mathcal{S}$.

The viability and capturability issues use the notion of evolutions viable in a subset.

DEFINITION 2.2. *Let* $K \subset X$ *be a subset of* X. *A function* $t \in [0,T] \mapsto x(t) \in X$ *is said to be* viable *in* K *on* $[0,T]$ *if*

$$\forall\, t \in [0,T], \;\; x(t) \;\in\; K,$$

and viable in K *if* $T = +\infty$.

The following results dealing with these issues shall use only the translation and concatenation properties and topological properties such that the upper semicompactness[2] and/or lower semicontinuity of the evolutionary system $\mathcal{S} : x \in X \rightsquigarrow \mathcal{S}(x) \subset \mathcal{C}(0,\infty;X)$.

Before proceeding further, let us recall that differential inclusions provide examples of evolutionary systems.

2.2. Evolutionary systems associated with differential inclusions. Let $X := \mathbf{R}^n$ be a finite dimensional vector space, and let $F : X \rightsquigarrow X$ be a strict[3]

[2]A set-valued map $F : X \rightsquigarrow Y$ is said to be *upper semicompact* at x if for every sequence x_n converging to x and for every sequence $y_n \in F(x_n)$, there exists a subsequence y_{n_p} converging to some $y \in F(x)$. It is said to be lower semicontinuous at x if and only if for any $y \in F(x)$ and for any sequence of elements $x_n \in \mathrm{Dom}(F)$ converging to x, there exists a sequence of elements $y_n \in F(x_n)$ converging to y.

[3]This means that for every $x \in X$, $F(x) \neq \emptyset$. We denote by

$$\mathrm{Graph}(F) := \{(x,y) \in X \in Y \mid y \in F(x)\}$$

the *graph* of a set-valued map $F : X \rightsquigarrow Y$ and by $\mathrm{Dom}(F) := \{x \in X | F(x) \neq \emptyset\}$ its *domain*.

set-valued map. We denote by $\mathcal{S}_F(x) \subset \mathcal{C}(0,\infty;X)$ the set of *absolutely continuous functions* $t \mapsto x(t) \in X$ satisfying

$$\text{for almost all } t \geq 0, \quad x'(t) \in F(x(t)),$$

starting at time 0 at x: $x(0) = x$. The set-valued map $\mathcal{S}_F : X \rightsquigarrow \mathcal{C}(0,\infty;X)$ is called the *solution map* (or the set-valued flow) associated with F.

Without assumptions, the solution map $\mathcal{S}_F$ may have empty values. However, whenever the solution map $\mathcal{S}_F : X \rightsquigarrow \mathcal{C}(0,\infty;X)$ associated with the differential inclusion $x' \in F(x)$ is strict, it obviously satisfies the *translation property* and the *concatenation property*.

One can also observe that the backward evolutionary system $\mathcal{S}_{F_-}$ is the solution map $\mathcal{S}_{-F}$ associated with $-F$.

2.2.1. Marchaud differential inclusions.

DEFINITION 2.3 (Marchaud map). *We shall say that F is a* Marchaud map *if*

$$\begin{cases} \text{(i)} & \text{the graph and the domain of } F \text{ are nonempty and closed,} \\ \text{(ii)} & \text{the values } F(x) \text{ of } F \text{ are convex,} \\ \text{(iii)} & \text{the growth of } F \text{ is linear:} \\ & \exists\, c > 0 \mid \forall\, x \in X, \quad \|F(x)\| := \sup_{v \in F(x)} \|v\| \leq c(\|x\| + 1). \end{cases}$$

We recall the following version of the important Theorem 3.5.2 of [2] stating that the solution map is strict and upper semicompact.

STATEMENT 1. *Assume that $F : X \rightsquigarrow X$ is Marchaud. Then the solution map $\mathcal{S}_F$ is an upper semicompact evolutionary system from X into the space of continuous functions supplied with the topology of uniform convergence on compact intervals.*

2.2.2. Lipschitz differential inclusions.

DEFINITION 2.4. *The set-valued map F is said to be Lipschitz if there exists a constant $\lambda > 0$ such that*

$$\forall\, x, y \in X, \quad F(x) \subset F(y) + B(0, \lambda\|x - y\|).$$

The Filippov theorem (see Theorem 5.3.1 of [2]) implies that whenever F is Lipschitz, the associated evolutionary system is lower semicontinuous.

STATEMENT 2. *Assume that $F : X \rightsquigarrow X$ is Lipschitz. Then the solution map $\mathcal{S}_F$ is a lower semicontinuous evolutionary system from X into the space of continuous functions supplied with the topology of uniform convergence on compact intervals.*

3. Exit and hitting time functions. We shall associate with an evolutionary system $\mathcal{S} : X \rightsquigarrow \mathcal{C}(0,\infty;X)$ the concepts of upper exit time function of a subset K and the lower hitting function (or minimal time function) of a target and study their continuity (actually, semicontinuity) properties.

DEFINITION 3.1. *Let $K \subset X$ be a subset. The functional $\tau_K : \mathcal{C}(0,\infty;X) \mapsto \mathbf{R}_+ \cup \{+\infty\}$ associating with $x(\cdot)$ its* exit time *$\tau_K(x(\cdot))$ defined by*

$$\tau_K(x(\cdot)) := \inf\{t \in [0,\infty[\mid x(t) \notin K\}$$

is called the exit functional.

Let $C \subset K$ be a target. We introduce the (constrained) hitting functional *$\varpi_{(K,C)}$ defined by*

$$\varpi_{(K,C)}(x(\cdot)) := \inf\{t \geq 0 \mid x(t) \in C \text{ and } \forall s \in [0,t],\ x(s) \in K\}$$

associating with $x(\cdot)$ *its* hitting time*, introduced in* [29]*). When* $K := X$*, we set*

$$\varpi_C(x(\cdot)) \;=\; \varpi_{(X,C)}(x(\cdot)) : \mathcal{C}(0,\infty;X) \mapsto \mathbf{R}_+ \cup \{+\infty\}$$

and call it the hitting functional *(or* minimal time functional*).*

We use the convention $\inf\{\emptyset\} := +\infty$, and we observe that

$$\text{if } \varpi_{(X,C)}(x(\cdot)) < +\infty, \text{ then } \varpi_C(x(\cdot)) \;=\; \varpi_{(X,C)}(x(\cdot)) \;\leq\; \tau_C(x(\cdot)).$$

We also note that

$$\forall\, s \in [0, \varpi_C(x(\cdot))],\;\; \varpi_C(x(\cdot + s)) \;=\; \varpi_C(x(\cdot)) - s \tag{3.1}$$

and that if $K_1 \subset K_2$ and $D_1 \supset D_2$, then $\varpi_{(K_1,D_1)}(x(\cdot)) \leq \varpi_{(K_2,D_2)}(x(\cdot))$. Let us point out that

$$\varpi_{\bigcup_{i=1}^{n} D_i}(x(\cdot)) \;=\; \min_{i=1,\ldots,n} \varpi_{D_i}(x(\cdot)).$$

Therefore,

$$\forall\, x \in K\backslash C,\;\; \tau_{K\backslash C}(x(\cdot)) = \min(\varpi_C(x(\cdot), \tau_K(x(\cdot)))$$

since

$$\tau_{K\backslash C}(x) \;=\; \varpi_{X\backslash(K\backslash C)}(x(\cdot)) \;=\; \varpi_{C\cup(X\backslash K)} \;=\; \min(\varpi_C(x(\cdot)), \varpi_{X\backslash K}(x(\cdot))).$$

Definition 3.2. *Consider an evolutionary system* $\mathcal{S} : X \rightsquigarrow \mathcal{C}(0,+\infty;X)$*. Let* $C \subset K$ *and* K *be two subsets.*

The function $\tau_K^\sharp : K \mapsto \mathbf{R}_+ \cup \{+\infty\}$ *defined by*

$$\tau_K^\sharp(x) \;:=\; \sup_{x(\cdot)\in\mathcal{S}(x)} \tau_K(x(\cdot))$$

is called the upper exit function*.*

The function $\varpi^\flat_{(K,C)} : K \mapsto \mathbf{R}_+ \cup \{+\infty\}$ *defined by*

$$\varpi^\flat_{(K,C)}(x) \;:=\; \inf_{x(\cdot)\in\mathcal{S}(x)} \varpi_{(K,C)}(x(\cdot))$$

is called the lower constrained hitting function*, and the function*

$$\varpi^\flat_C(x) \;:=\; \inf_{x(\cdot)\in\mathcal{S}(x)} \varpi_C(x(\cdot))$$

is called the lower hitting function*.*

Statement 3. *Let* $\mathcal{S} : X \rightsquigarrow \mathcal{C}(0,+\infty;X)$ *be a strict upper semicompact map, and let* C *and* K *be two closed subsets such that* $C \subset K$*. Then the hitting function* $\varpi^\flat_{(K,C)}$ *is lower semicontinuous and the exit function* $\tau_K^\sharp$ *is upper semicontinuous. Furthermore, for any* $x \in \mathrm{Dom}(\varpi^\flat_{(K,C)})$*, there exists at least one solution* $x^\flat(\cdot) \in \mathcal{S}(x)$ *which hits* C *as soon as possible before possibly leaving* K*,*

$$\varpi^\flat_{(K,C)}(x) \;=\; \varpi_{(K,C)}(x^\flat(\cdot)),$$

and for any $x \in \mathrm{Dom}(\tau_K^\sharp)$*, there exists at least one solution* $x^\sharp(\cdot) \in \mathcal{S}(x)$ *which remains viable in* K *as long as possible:*

$$\tau_K^\sharp(x) \;=\; \tau_K(x^\sharp(\cdot)).$$

This statement is a consequence of the more general Theorem 6.2 dealing with upper hypolimits of upper exit functions and epilimits of lower constrained epifunctions of subsets that is proved below. See also [29, 30].

4. Viability kernels and capture basins. We shall answer in this section questions such as the following.

- Starting from K, is it possible to remain viable in K (as long as possible)?
- Starting outside of a target $C \subset K$, is it possible to reach it (as fast as possible) while being viable in the subset K?

These two very natural questions lead to the introduction of the following concepts.

4.1. Viability kernels with targets. We now define the viability kernels, the capture basins, and the viable-capture basins of a subset under a set-valued map.

DEFINITION 4.1. *Let $\mathcal{S} : X \rightsquigarrow \mathcal{C}(0, +\infty; X)$ be a set-valued evolutionary system, and let $C \subset K \subset X$ be two subsets, C being regarded as a target and K as a constrained set.*

1. *The subset K is said to be* locally viable *under $\mathcal{S}$ if from any initial state $x \in K$ starts at least one solution viable in K on a nonempty interval and* viable *if this solution is viable on $[0, +\infty[$. We shall say that K* captures *the target C if from any initial state $x \in K$ starts at least one solution viable in K until it may reach the target C, and we say that K* finitely captures *the target C if it reaches it in finite time.*
2. *The subset $\mathrm{Viab}(K, C)$ of initial states $x_0 \in K$ such that at least one solution $x(\cdot) \in \mathcal{S}(x_0)$ starting at x_0 is viable in K for all $t \geq 0$ or viable in K until it reaches C in finite time is called the* viability kernel of K with target C *under $\mathcal{S}$. A subset $C \subset K$ is said to be* isolated in K by $\mathcal{S}$ *if it coincides with its viability kernel:*

$$\mathrm{Viab}(K, C) \;=\; C.$$

3. *The subset $\mathrm{Capt}^K(C)$ of initial states $x_0 \in K$ such that C is reached in finite time before possibly leaving K by at least one solution $x(\cdot) \in \mathcal{S}(x_0)$ starting at x_0 is called the* viable-capture basin *of C in K, and*

$$\mathrm{Capt}(C) \;:=\; \mathrm{Capt}^X(C)$$

is said to be the capture basin *of C.*
4. *When the target $C = \emptyset$ is the empty set, we set*

$$\mathrm{Viab}(K) \;:=\; \mathrm{Viab}(K, \emptyset) \;\; \textit{and} \;\; \mathrm{Capt}^K(\emptyset) \;=\; \emptyset,$$

and we say that $\mathrm{Viab}(K)$ is the viability kernel of K. A subset K is a repeller *under $\mathcal{S}$ if its viability kernel is empty, or, equivalently, if the empty set is isolated in K.*

In other words, the viability kernel $\mathrm{Viab}(K)$ is the subset of initial states $x_0 \in K$ such that at least one solution $x(\cdot) \in \mathcal{S}(x_0)$ starting at x_0 is viable in K for all $t \geq 0$. Furthermore, we observe that

$$\mathrm{Viab}(K, C) \;=\; \mathrm{Viab}(K \backslash C) \cup \mathrm{Capt}^K(C). \tag{4.1}$$

Therefore, the viability kernel $\mathrm{Viab}(K, C)$ of K with target C coincides with the capture basin $\mathrm{Capt}^K(C)$ of C viable in K whenever the viability kernel $\mathrm{Viab}(K \backslash C)$ is empty, i.e., whenever $K \backslash C$ is a repeller:

$$\mathrm{Viab}(K \backslash C) \;=\; \emptyset \;\Rightarrow\; \mathrm{Viab}(K, C) \;=\; \mathrm{Capt}^K(C). \tag{4.2}$$

Consequently, the concept of the viability kernel with a target allows us to study both the viability kernel of a closed subset and the viable-capture basin of a target.

Remark. If subsets K_i capture a given target $C \subset K_i$ for all $i \in I$, so does their union $\bigcup_{i\in I} K_i$. However, the intersection of two subsets K_1 and K_2 capturing a same target C does not necessarily capture C, since starting from a state of $K_1 \cap K_2$, there may exist two different solutions that are viable in K_1 or in K_2 but no solution viable in $K_1 \cap K_2$. □

We observe that the *viability kernel* is characterized by

$$\mathrm{Viab}(K) \;:=\; \{x \in K \mid \tau^\sharp_K(x) \;=\; +\infty\}$$

and that the *viable-capture basin*

$$\mathrm{Capt}^K(C) \;:=\; \{x \in K \mid \varpi^\flat_{(K,C)}(x) \;<\; +\infty\}$$

is the domain of the constrained hitting function $\varpi^\flat_{(K,C)}$.

To say that K is a repeller under $\mathcal{S}$ amounts to saying that the exit function $\tau^\sharp_K$ is finite on K, and to say that $K\backslash C$ is a repeller amounts to saying that all solutions $x(\cdot) \in \mathcal{S}(x)$ starting from $x \in K\backslash C$ reach C or leave K in finite time, i.e., satisfy $\tau_{K\backslash C}(x(\cdot)) = \min(\varpi_C(x(\cdot), \tau_K(x(\cdot)))) < +\infty$.

The viability kernel $\mathrm{Viab}(K, C)$ of K with target C captures C.

PROPOSITION 4.2. *The viability kernel* $\mathrm{Viab}(K, C)$ *of* K *with target* C *is the largest subset of* K *capturing* C*, and the viability kernel* $\mathrm{Viab}(K)$ *of* K *is the largest viable subset of* K*.*

Proof. First, any subset D such that $C \subset D \subset K$ capturing C is obviously contained in the viability kernel $\mathrm{Viab}(K, C)$ with target C.

For proving that the viability kernel $\mathrm{Viab}(K, C)$ with target C captures the target C, take $x_0 \in \mathrm{Viab}(K, C)$, and prove that there exists a solution $x(\cdot) \in \mathcal{S}(x_0)$ starting at x_0 viable in $\mathrm{Viab}(K, C)$ until it possibly reaches C. Indeed, there exists a solution $x(\cdot) \in \mathcal{S}(x_0)$ viable in K until some time $T \geq 0$, either finite when it reaches C or infinite. Then for all $t \in [0, T[$, the function $y(\cdot)$ defined by $y(\tau) := x(t + \tau)$ is a solution $y(\cdot) \in \mathcal{S}(x(t))$ starting at $x(t)$ and viable in K until it reaches C at time $T - t$. Hence $x(t)$ does belong to $\mathrm{Viab}(K, C)$ for every $t \in [0, T[$. □

Furthermore, we derive from Proposition 4.2 and Theorem 6.4 below the following proposition.

PROPOSITION 4.3. *Let us assume that the map* $\mathcal{S}$ *is upper semicompact and that* $C \subset K$ *and* K *are closed. Then the viability kernel* $\mathrm{Viab}(K, C)$ *with a target is the largest closed subset of* K *capturing* C*, and the viability kernel* $\mathrm{Viab}(K)$ *is the largest viable closed subset of* K*.*

4.2. Characterization of the viability kernel with a target. The first characterization is stated in the following theorem.

THEOREM 4.4. *Let us assume that* $\mathcal{S}$ *is upper semicompact and that the subsets* $C \subset K$ *and* K *are closed. The viability kernel* $\mathrm{Viab}(K, C)$ *of a subset* K *with target* C *under* $\mathcal{S}$ *is the* unique *closed subset satisfying* $C \subset D \subset K$*, and*

$$\text{(4.3)} \qquad \begin{cases} \text{(i)} & D\backslash C \textit{ is} \text{ locally viable } \textit{under } \mathcal{S}, \\ \text{(ii)} & D \textit{ is} \text{ isolated } \textit{in } K \textit{ by } \mathcal{S} \ (\mathrm{Viab}(K, D) \;=\; D). \end{cases}$$

It follows from Theorem 4.6 characterizing the viability kernel as the largest closed subset $D \subset K$ such that $D\backslash C$ is locally viable and from Theorem 4.7 characterizing the viability kernel as the smallest subset D isolated in K.

We begin with necessary conditions.

PROPOSITION 4.5. *Let us consider a closed subset C of K. Then the following hold.*

1. *If $D \supset C$ captures C, then $D\backslash C$ is locally viable under $\mathcal{S}$.*
2. *If $D_1 \supset C$ captures C and $D_2 \supset D_1$ captures D_1, then D_2 captures C (transitivity of the capturability property).*

Consequently, the viability kernel $\mathrm{Viab}(K, C)$ *of a subset K with target C under $\mathcal{S}$ satisfies the following properties:*

$$\begin{cases} \text{(i)} & \mathrm{Viab}(K,C)\backslash C \text{ is locally viable and } \mathrm{Viab}(K) \text{ is viable under } \mathcal{S}, \\ \text{(ii)} & \mathrm{Viab}(K,C) \text{ is isolated in } K \text{ by } \mathcal{S} \ (\mathrm{Viab}(K, \mathrm{Viab}(K,C)) = \mathrm{Viab}(K,C)). \end{cases}$$

Proof. For proving the first statement, take $x_0 \in D\backslash C$, and prove that there exists a solution $x(\cdot) \in \mathcal{S}(x_0)$ starting at x_0 viable in $D\backslash C$ on a nonempty interval. Indeed, since C is closed, there exists $\eta > 0$ such that $B(x_0, \eta) \cap C = \emptyset$, so that $x(t) \in B(x_0, \eta) \cap D \subset D\backslash C$ on some nonempty interval.

For proving that D_2 captures C, take any $x_0 \in D_2$. There exists a solution $x(\cdot) \in \mathcal{S}(x_0)$ viable in D_2 forever or else, until it possibly reaches the subset D_1 of D_2 at some *finite time* $T > 0$ at $x(T) \in D_1$. In this case, for any $t \geq T$, $x(t)$ remains in D_1, and thus, in D_2, until it possibly reaches C. Hence D_2 captures C.

In particular, if we take $D_1 := \mathrm{Viab}(K, C)$ and $D_2 := \mathrm{Viab}(K, \mathrm{Viab}(K, C))$, we infer that $D_1 = D_2$ since D_1 is the largest subset of K capturing the target C. □

We now proceed with the proof of the sufficiency.

THEOREM 4.6. *Assume that $\mathcal{S}$ is upper semicompact. Let $C \subset K$ be closed subsets.*

Then the viability kernel $\mathrm{Viab}(K, C)$ *of K with target C under $\mathcal{S}$ is the* largest closed subset $D \subset K$ *containing C such that $D\backslash C$ is locally viable.*

In particular, K captures C if and only if $K\backslash C$ is locally viable.

Proof. When $C = \emptyset$, this is Proposition 4.3. Otherwise, Theorem 6.4 and Proposition 4.5 imply that the viability kernel $\mathrm{Viab}(K, C)$ of K with target C under $\mathcal{S}$ is a closed subset such that $\mathrm{Viab}(K, C)\backslash C$ is locally viable.

Let $D \subset K$ containing C such that $D\backslash C$ is locally viable. Since $C \subset \mathrm{Viab}(K, C)$, let us take x in $D\backslash C$ and show that it belongs to $\mathrm{Viab}(K, C)$. Either there exists a solution $x(\cdot) \in \mathcal{S}(x)$ viable in $D \subset K$ forever or, if not, by Statement 3, there exists a solution $x^\sharp(\cdot) \in \mathcal{S}(x)$ that maximizes $\tau_D(x(\cdot))$,

$$\tau_D^\sharp(x) := \sup_{x(\cdot) \in \mathcal{S}(x)} \tau_D(x(\cdot)) = \tau_D(x^\sharp(\cdot)),$$

and thus that leaves D at $x^\sharp := x^\sharp(\tau_D^\sharp(x)) \in D$. Actually, this point belongs to C.

Otherwise, since $D\backslash C$ is locally viable, one could associate with $x^\sharp \in D\backslash C$ a solution $y(\cdot) \in \mathcal{S}(x^\sharp)$ and $T > 0$ such that $y(\tau) \in D\backslash C$ for all $\tau \in [0, T]$. Concatenating this solution to $x^\sharp(\cdot)$, we obtain a solution viable in D on an interval $[0, \tau_D^\sharp(x) + T]$, which contradicts the definition of $x^\sharp(\cdot)$. □

THEOREM 4.7. *Let $C \subset K$. Then the viability kernel* $\mathrm{Viab}(K, C)$ *is the* smallest *subset D between C and K isolated in K by $\mathcal{S}$.*

Proof. Proposition 4.5 implies that the viability kernel $\mathrm{Viab}(K, C)$ is isolated in K by $\mathcal{S}$. Conversely, since D is isolated in K by $\mathcal{S}$, we infer that $\mathrm{Viab}(K, C) \subset \mathrm{Viab}(K, D) = D$. □

4.3. Isolated subsets. We need to characterize further isolated subsets for enriching the above characterization theorem.

First, we point out the following.

PROPOSITION 4.8. *Let C and K be two subsets such that $C \subset K$. Then the following properties are equivalent.*

1. *C is isolated in K by $\mathcal{S}$:* $\mathrm{Viab}(K, C) = C$.
2. *For all $x \in K\backslash C$, all solutions reach $X\backslash K$ in finite time before (possibly) hitting C.*
3. $\mathrm{Viab}(K) = \mathrm{Viab}(C)$, *and* $\mathrm{Capt}^K(C) = C$.
4. *$K\backslash C$ is a repeller and* $\mathrm{Capt}^K(C) = C$.

Isolated subsets enjoy local backward invariance properties discovered by Hélène Frankowska in her studies of Hamilton–Jacobi equations associated with value functions of optimal control problems under state constraints that play a crucial role in the characterization of viability kernels with a target. Indeed, there is a close connection between isolation in K and local backward invariance relatively to K.

DEFINITION 4.9. *We shall say that a subset $C \subset K$ is* locally backward invariant relatively to K *under $\mathcal{S}$ if for every $x \in C$, all backward solutions starting from x and viable in K on an interval $[0, T]$ are viable in C on $[0, T]$, i.e., if for every $x \in C$, for every $t_0 \in]0, +\infty[$, and for all solutions $x(\cdot)$ arriving at x at time t_0 such that there exists $s \in [0, t_0[$ such that $x(\cdot)$ is viable in K on the interval $[s, t_0]$, then $x(\cdot)$ is viable in C on the same interval.*

Naturally, if $C \subset K$ is locally backward invariant, it remains locally backward invariant relatively to K. If K is itself locally backward invariant, any subset locally backward invariant relatively to K is locally backward invariant.

If $C \subset K$ is locally backward invariant relatively to K, then $C \cap \mathrm{Int}(K)$ is locally backward invariant, and from any $x \in C \cap \partial K$, all backward solutions $y(\cdot) \in \mathcal{S}_-(x)$ satisfy

$$\begin{cases} \text{either } \exists\, T > 0 \quad \text{such that } \forall\, t \in [0, T], \quad x(t) \in C, \\ \text{or } \exists\ t_n \to 0+ \mid y(t_n) \in X\backslash K. \end{cases}$$

THEOREM 4.10. *A closed subset $C \subset K$ is locally backward invariant relatively to K if and only if* $\mathrm{Capt}^K(C) = C$.

Proof. Assume that C is locally backward invariant relatively to K, and consider $x \in \mathrm{Capt}^K(C)\backslash C$. There exists a solution $x(\cdot) \in \mathcal{S}(x)$ viable in K until it reaches C at time $T := \varpi_C(x(\cdot)) \geq 0$ at $c = x(\varpi_C(x(\cdot)))$. Since C is closed, then $T > 0$ is positive. Let $z(\cdot) \in \mathcal{S}_-(x)$, and let $y(\cdot)$ be the function defined by

$$y(t) := \begin{cases} x(T - t) & \text{if} \quad t \in [0, T], \\ z(t - T) & \text{if} \quad t \geq T. \end{cases}$$

Then $y(\cdot) \in \mathcal{S}_-(c)$ and is viable in K on the interval $[0, \varpi_C(x(\cdot))]$. Since C is assumed to be locally backward invariant relatively to K, then $y(t) \in C$ for all $t \in [0, \varpi_C(x(\cdot))]$, and, in particular, $y(T) = x$ belongs to C. We have obtained a contradiction.

The converse statement follows from the next theorem.

PROPOSITION 4.11. *The viability kernel* $\mathrm{Viab}(K, C)$ *of K with a target $C \subset K$ and the viable-capture basin* $\mathrm{Capt}^K(C)$ *are locally backward invariant relatively to K. Consequently, every subset $C \subset K$ isolated in K is locally backward invariant relatively to K.*

Proof. Let us consider $x \in \mathrm{Viab}(K, C)$ and $z(\cdot) \in \mathcal{S}(x)$ viable in K until it possibly reaches C. Let us consider a backward solution $y(\cdot) \in \mathcal{S}_-(x)$ viable in K

such that $\tau_K(y(\cdot)) > 0$. (This is always the case whenever $x \in \mathrm{Int}(K)$.) For every $T \in [0, \tau_K(y(\cdot))[$, we associate with it the solution $x(\cdot) \in \mathcal{S}(x(T))$ defined by

$$x(t) := \begin{cases} y(T-t) & \text{if} \quad t \in [0,T], \\ z(t-T) & \text{if} \quad t \geq T, \end{cases}$$

starting at $y(T) \in K$ viable in K until it possibly reaches C. This means that $y(T) \in \mathrm{Viab}(K,C)$ for every $T \in [0, \tau_K(y(\cdot))[$, i.e., that the backward solution $y(\cdot) \in \mathcal{S}_-(x)$ is viable in $\mathrm{Viab}(K,C)$ on the interval $[0, \tau_K(y(\cdot))]$.

In other words, for every $x \in \mathrm{Viab}(K,C)$, every backward solution viable in K on some time interval is actually viable in $x \in \mathrm{Viab}(K,C)$ on the same interval. □

We derive the following characterization.

PROPOSITION 4.12. *Let us consider a closed subset $C \subset K$. Then C is isolated in K by $\mathcal{S}$ if and only if*

1. *C is locally backward invariant relatively to K, and*
2. *$K \backslash C$ is a repeller.*

Putting together these results, we obtain Theorem 4.13, characterizing viability kernels with targets, and Theorem 4.14, characterizing capture basins.

THEOREM 4.13. *Let us assume that $\mathcal{S}$ is upper semicompact and that the subsets $C \subset K$ and K are closed. The viability kernel* $\mathrm{Viab}(K,C)$ *of a subset K with target C under $\mathcal{S}$ is the* unique *closed subset satisfying $C \subset D \subset K$ and*

$$(4.4) \qquad \begin{cases} \text{(i)} & D \backslash C \text{ is locally viable under } \mathcal{S}, \\ \text{(ii)} & D \text{ is locally backward invariant relatively to } K \text{ under } \mathcal{S}, \\ \text{(iii)} & K \backslash D \text{ is a repeller under } \mathcal{S}. \end{cases}$$

Therefore, Theorem 4.13 and (4.2) imply that when $K \backslash C$ is a repeller, the above theorem implies a characterization of the viable-capture basins.

THEOREM 4.14. *Let us assume that $\mathcal{S}$ is upper semicompact and that a closed subset $C \subset K$ satisfies the property*

$$(4.5) \qquad \mathrm{Viab}(K \backslash C) = \emptyset.$$

Then the viable-capture basin $\mathrm{Capt}^K(C)$ *is the* unique *closed subset D satisfying $C \subset D \subset K$ and*

$$(4.6) \qquad \begin{cases} \text{(i)} & D \backslash C \text{ is locally viable under } \mathcal{S}, \\ \text{(ii)} & D \text{ is locally backward invariant relatively to } K \text{ under } \mathcal{S}. \end{cases}$$

4.4. Viability kernels of backward invariant sets. We obtain further properties when K is backward invariant under $\mathcal{S}$. To begin with, *the capture basin* $\mathrm{Capt}(C) := \mathrm{Capt}^X(C)$ *is contained in K and equal to* $\mathrm{Capt}^K(C)$, *so that*

$$\mathrm{Viab}(K,C) = \mathrm{Viab}(K \backslash C) \cup \mathrm{Capt}(C).$$

THEOREM 4.15. *A subset K is invariant under a set-valued map $\mathcal{S}$ if and only if its complement $X \backslash K$ is backward invariant under $\mathcal{S}$.*

Proof. To say that K is not invariant under $\mathcal{S}$ amounts to saying that there exists a solution $x(\cdot) \in \mathcal{S}(x_0)$ and $T > 0$ such that

$$x(0) \in K \ \& \ x(T) \in X \backslash K.$$

Let $z(\cdot) \in \mathcal{S}_-(x_0)$ be a backward solution, and define the function $y(\cdot)$ by

$$y(t) = \begin{cases} x(T-t) & \text{if} \quad t \in [0,T], \\ z(t-T) & \text{if} \quad t \geq T. \end{cases}$$

It is a backward solution starting at $y(0) = x(T) \in X\backslash K$ and satisfying $y(T) = x_0 \in K$. This amounts to saying that the complement $X\backslash K$ of K is not backward invariant. □

We then derive the following theorem.

THEOREM 4.16. *Assume that $\mathcal{S}$ is upper semicompact, that $C \subset K$ and K are closed, and that K is backward invariant under $\mathcal{S}$. Then the viability kernel* $\mathrm{Viab}(K,C)$ *of K with target C under $\mathcal{S}$ is the* unique *closed subset D satisfying $C \subset D \subset K$ and*

$$\begin{cases} \text{(i)} & D\backslash C \text{ is locally viable under } \mathcal{S}, \\ \text{(ii)} & D \text{ is backward invariant under } \mathcal{S} \text{ (or, equivalently, } X\backslash D \text{ is invariant under } \mathcal{S}), \\ \text{(iii)} & K\backslash D \text{ is a repeller under } \mathcal{S}. \end{cases}$$

Proof. To say that K is backward invariant amounts to saying that the complement of K is invariant thanks to Theorem 4.15. Therefore, $\mathrm{Viab}(K,C)$ being isolated, all solutions starting from $K\backslash\mathrm{Viab}(K,C)$ leave K in finite time before possibly hitting C. Actually, they never reach C because the complement $X\backslash K$ is invariant. Hence we have checked that the complement $X\backslash\mathrm{Viab}(K,C)$ of the viability kernel of K with target C is invariant. Theorem 4.15 implies that the viability kernel $\mathrm{Viab}(K,C)$ of K with target C is backward invariant. □

4.5. The barrier property. The boundary of the viability kernel satisfies the barrier property.

DEFINITION 4.17. *If $D \subset K$, the* boundary $\partial_K(D)$ of D relative to K *is the subset*

$$\partial_K(D) := \overline{D} \cap \overline{(K\backslash D)},$$

and the subset $\partial D := \partial_X(D)$ is called the boundary *of D. We shall say that a subset $D \subset K$ enjoys the* barrier property relative to K under $\mathcal{S}$ *if its boundary $\partial_K(D)$ of D relative to K is locally invariant with respect to D: For every $x \in \partial_K(D)$, all solutions starting from x viable in D are actually viable in the boundary $\partial_K(D)$ of D relative to K until they reach the boundary of K.*

We see at once that

$$\partial_K(D) \cap \mathrm{Int}(K) = \partial D \cap \mathrm{Int}(K)$$

and that

$$\text{if } D \subset \mathrm{Int}(K), \text{ then } \partial_K(D) = \partial D.$$

Remark on the barrier property. The "barrier property" of the viability kernel of a closed subset has been discovered by Marc Quincampoix in [46] and generalized by Pierre Cardaliaguet in [25, 26, 27, 28] for differential games. It plays an important role in control theory and the theory of differential games, because every solution starting from the boundary of the viability kernel can either remain in the boundary or leave the viability kernel, or, equivalently, no solution starting from outside the viability

kernel can cross its boundary. Such solutions can remain only on the boundary of the viability kernel, or leave it.

This is a semipermeability property of the viability kernel, which is very important in terms of interpretation. Viability is indeed a very fragile property, which cannot be reestablished from the outside. *In other words, love it or leave it.* □

THEOREM 4.18. *If $\mathcal{S}$ is upper semicompact and lower semicontinuous, then the viability kernel* $\mathrm{Viab}(K,C)$ *of a closed subset K with a closed target $C \subset K$ under $\mathcal{S}$ enjoys the barrier property relative to K.*

Proof. Let x belong to $\partial_K(\mathrm{Viab}(K,C))$, and let $x(\cdot) \in \mathcal{S}(x)$ be a solution viable in K forever ($\varpi^\sharp_{(K,C)}(x(\cdot)) = +\infty$) or until it reaches C at finite time $\varpi^\sharp_{(K,C)}(x(\cdot)) < +\infty$. Let $x_n \in K\backslash\mathrm{Viab}(K,C)$ converge to x. Since $\mathcal{S}$ is lower semicontinuous by Statement 3, there exists a solution $x_n(\cdot) \in \mathcal{S}(x_n)$ converging to $x(\cdot)$ uniformly over compact intervals. Since $\mathrm{Viab}(K,C)$ is isolated, we know that for every n,

$$\forall\, t \leq \tau_K(x_n(\cdot)),\ \ x_n(t) \in K\backslash\mathrm{Viab}(K,C).$$

Since $\varpi_{\partial K}(x_n(\cdot)) \ \leq\ \tau_K(x_n(\cdot))$ and since the functional $x(\cdot) \mapsto \varpi_{\partial K}(x(\cdot))$ is lower semicontinuous, we infer that for every $t < \varpi_{\partial K}(x(\cdot))$ there exists $N > 0$ such that for any $n \geq N$,

$$t < \varpi_{\partial K}(x_n(\cdot)) \ \leq\ \tau_K(x_n(\cdot)),$$

and thus that $x_n(t)$ belongs to $K\backslash\mathrm{Viab}(K,C)$. Taking the limit, we infer that $x(t)$ belongs to $\overline{K\backslash\mathrm{Viab}(K,C)}$. Hence $x(t)$ belongs to the boundary $\partial_K(\mathrm{Viab}(K,C))$ of the viability kernel relative to K whenever $t < \varpi_{\partial K}(x(\cdot))$. □

5. Frankowska's and viscosity property of viability kernels. We restrict now our study to the case of viability kernels with targets under evolutionary systems defined by the solution maps of differential inclusions $x' \in F(x)$. In this case, the viability and invariance theorems characterize the viability and invariance properties by tangential conditions, as it was mentioned in the introduction, or, equivalently,[4] by normal conditions. We recall that the (regular) *normal cone*[5] $N_L(x) := T_L(x)^-$ to a subset L at $x \in L$ is the polar cone to the contingent cone $T_L(x)$ (see, for instance, [10] or [49] for more details). We denote by

$$\forall\, p \in X^\star,\ \ \sigma(K,p) \ :=\ \sup_{x\in K}\langle p,x\rangle$$

the *support function* of K.

$$\forall\, x \in K\backslash R^{-1}(K),\ \ F(x) \cap T_K(x) \neq \emptyset.$$

5.1. The basic viability and invariance theorems.

STATEMENT 4. *Assume that F is Marchaud. The two following statements hold true.*

1. *If K is closed, then K is (globally) viable under F if and only if*

$$\forall\, x \in K,\ \ F(x) \cap T_K(x) \neq \emptyset,$$

[4]The equivalence between tangential and normal conditions was first noticed in a different context in [41]. A simpler proof of this fact was given by Hélène Frankowska and appeared in [14] and in Theorem 3.2.4 of [2]. Other proofs were provided later in [24] and [57].

[5]One can replace if wished this normal cone $N_L(x)$ by the smaller subset $x - \Pi_L(x)$ of *normal proximals* to L at x, where $\Pi_L(x)$ denotes the set of best approximations of x by elements of L.

or, equivalently, in dual form, if and only if

$$\forall\, x \in K,\ \forall\, p \in N_K(x),\ \ \sigma(F(x), -p)\ \geq\ 0.$$

2. *If $C \subset K$ is closed, then K captures C by F if and only if*

$$\forall\, x \in K\backslash C,\ \ F(x) \cap T_K(x) \neq \emptyset,$$

or, equivalently, in dual form, if and only if

$$\forall\, x \in K\backslash C,\ \forall\, p \in N_K(x),\ \ \sigma(F(x), -p)\ \geq\ 0.$$

STATEMENT 5. *Assume that F is Lipschitz. The two following statements hold true.*

1. *If K is closed, then K is (globally) invariant under F if and only if*

$$\forall\, x \in K,\ \ F(x)\ \subset\ T_K(x),$$

or, equivalently, in dual form, if and only if

$$\forall\, x \in K,\ \forall\, p \in N_K(x),\ \sigma(F(x), p)\ \leq\ 0.$$

2. *If $C \subset K$ is closed, then K absorbs C by F if and only if*

$$\forall\, x \in K\backslash C,\ \ F(x)\ \subset\ T_K(x),$$

or, equivalently, in dual form, if and only if

$$\forall\, x \in K\backslash C,\ \forall\, p \in N_K(x),\ \sigma(F(x), p)\ \leq\ 0.$$

3. *If $C \subset K$ is closed, then C is backward invariant under F relatively to K if and only if*

$$(5.1)\qquad \begin{cases} \text{(i)} & \forall\, x \in C \cap \mathrm{Int}(K),\ \ -F(x)\ \subset\ T_C(x), \\ \text{(ii)} & \forall\, x \in C \cap \partial K,\ \ -F(x)\ \subset\ T_C(x) \cup T_{X\backslash K}(x), \end{cases}$$

or, equivalently, in normal form, if and only if

$$(5.2)\begin{cases} \text{(i)} & \forall\, x \in C \cap \mathrm{Int}(K),\ \ \forall\, p \in N_C(x),\ \ \sigma(F(x), -p)\ \leq\ 0, \\ \text{(ii)} & \forall\, x \in C \cap \partial K,\ \ \forall\, p \in N_C(x) \cap N_{\overline{X\backslash K}}(x),\ \ \sigma(F(x), -p)\ \leq\ 0. \end{cases}$$

5.2. Tangential and normal characterizations of viability kernels with targets. Using the viability theorem, Statement 1, and the invariance theorem, Statement 5, we deduce that the viability kernels and the viable-capture basins enjoy tangential and normal characterizations.

For that purpose, we introduce the following Frankowska property.

DEFINITION 5.1. *Let us consider a set-valued map $F : X \rightsquigarrow X$ and two subsets $C \subset K$ and K. We shall say that a subset D between C and K satisfies the* Frankowska property *with respect to F if*

$$(5.3)\qquad \begin{cases} \text{(i)} & \forall\, x \in D\backslash C,\ \ F(x) \cap T_D(x)\ \neq\ \emptyset, \\ \text{(ii)} & \forall\, x \in D \cap \mathrm{Int}(K),\ \ -F(x)\ \subset\ T_D(x), \\ \text{(iii)} & \forall\, x \in D \cap \partial K,\ \ -F(x)\ \subset\ T_D(x) \cup T_{X\backslash K}(x), \end{cases}$$

or, equivalently, by duality, satisfying the "normal conditions"

$$(5.4)\quad \begin{cases} \text{(i)} & \forall\, x \in D\backslash C,\ \forall\, p \in N_D(x),\ \ \sigma(F(x), -p) \geq 0, \\ \text{(i)} & \forall\, x \in D \cap \mathrm{Int}(K),\ \forall\, p \in N_D(x),\ \ \sigma(F(x), -p) \leq 0, \\ \text{(ii)} & \forall\, x \in D \cap \partial K,\ \forall\, p \in N_D(x) \cap N_{\overline{X\backslash K}}(x),\ \ \sigma(F(x), -p) \leq 0. \end{cases}$$

When K is assumed further to be backward locally invariant, the above conditions (5.3) *and* (5.4) *boil down to*

$$(5.5)\quad \begin{cases} \text{(i)} & \forall\, x \in D\backslash C,\ \ F(x) \cap T_D(x) \neq \emptyset, \\ \text{(ii)} & \forall\, x \in D,\ \ -F(x) \subset T_D(x), \end{cases}$$

and

$$(5.6)\quad \begin{cases} \text{(i)} & \forall\, x \in D\backslash C,\ \forall\, p \in N_D(x),\ \ \sigma(F(x), -p) = 0, \\ \text{(ii)} & \forall\, x \in D,\ \forall\, p \in N_D(x),\ \ \sigma(F(x), -p) \leq 0, \end{cases}$$

respectively.

We deduce from the characterization theorem, Theorem 4.13, its tangential and normal formulations.

THEOREM 5.2. *Let us assume that F is Marchaud and that $C \subset K$ and K are closed. The viability kernel* $\mathrm{Viab}_F(K, C)$ *of the subset K with target C under F is*

1. *the* largest *closed subset D of K satisfying*

$$\forall\, x \in D\backslash C,\ \ F(x) \cap T_D(x) \neq \emptyset.$$

2. *When F is assumed to be also Lipschitz, the viability kernel* $\mathrm{Viab}_F(K, C)$ *is the* unique *closed subset $D \subset K$ satisfying*
   (a) *the Frankowska property* (5.3) *(or its dual formulation* (5.4)*);*
   (b) *$K\backslash D$ is a repeller.*

As a consequence, we obtain the following tangential characterization of viable-capture basins.

THEOREM 5.3. *Let us assume that F is Marchaud, that K is closed, and that a closed subset C satisfies* $\mathrm{Viab}_F(K\backslash C) = \emptyset$. *Then the viable-capture basin* $\mathrm{Capt}_F^K(C)$ *is*

1. *the* largest *closed subset D satisfying $C \subset D \subset K$ and*

$$(5.7)\qquad \forall\, x \in D\backslash C,\ \ F(x) \cap T_D(x) \neq \emptyset.$$

2. *If F is Lipschitz, the viable-capture basin* $\mathrm{Capt}_F^K(C)$ *is the* unique *closed subset D satisfying the Frankowska property* (5.3) *(or its dual formulation* (5.4)*).*

We now define the following "viscosity property."

DEFINITION 5.4. *Let us consider a set-valued map $F : X \rightsquigarrow X$ and two subsets $C \subset K$ and K. We shall say that a subset D between C and K satisfies the* viscosity property *with respect to F if*

$$(5.8)\quad \begin{cases} \text{(i)} & \forall\, x \in \underline{D\backslash C},\ \ F(x) \cap T_D(x) \neq \emptyset, \\ \text{(ii)} & \forall\, x \in \overline{X\backslash D},\ \ F(x) \subset T_{\overline{X\backslash D}}(x), \end{cases}$$

and, in normal form,

$$(5.9)\quad \begin{cases} \text{(i)} & \forall\, x \in \underline{D\backslash C},\ \forall\, p \in N_D(x),\ \ \sigma(F(x), -p) \geq 0, \\ \text{(ii)} & \forall\, x \in \overline{X\backslash D},\ \forall\, p \in N_{\overline{X\backslash D}}(x),\ \ \sigma(F(x), p) \leq 0, \end{cases}$$

respectively.

When $C = \emptyset$, we recognize the definition of a discriminating kernel of K of the Hamiltonian $H(x,p) := \sigma(F(x), -p)$ given in [26], for instance.

THEOREM 5.5. *Let us assume that F is Marchaud and Lipschitz, that $C \subset K$ and K are closed, and that K is backward invariant. The viability kernel* $\mathrm{Viab}_F(K,C)$ *of the subset K with the target C under F is the* unique *closed subset $D \subset K$ satisfying the following:*

1. *the viscosity property* (5.8) *(or its dual formulation* (5.9)*);*
2. *$K \backslash D$ is a repeller.*

6. Stability properties. Consider two sequences of subsets $C_n \subset C$ and $K_n \subset X$ and their Painlevé–Kuratowski upper limits

$$C^\sharp := \mathrm{Limsup}_{n\to+\infty} C_n \;\&\; K^\sharp := \mathrm{Limsup}_{n\to+\infty} K_n.$$

Recall that the upper limit of a sequence of constant subsets C is the closure of C.

DEFINITION 6.1. *We define the hypolimit* $\lim^\sharp_{\downarrow n\to\infty} \tau^\sharp_{K_n}$ *of upper exit functions* $\tau^\sharp_{K_n}$ *whose hypograph is the upper limit of the hypographs of the functions* $\tau^\sharp_{K_n}$

$$\mathcal{H}p\left(\lim^\sharp_{\downarrow n\to\infty}\left(\tau^\sharp_{K_n}\right)\right) := \mathrm{Limsup}_{n\to\infty} \mathcal{H}p(\tau^\sharp_{K_n}).$$

It is the upper hypolimit *of the functions $\tau^\sharp_{K_n}$, equal to*

$$\left(\lim^\sharp_{\downarrow n\to\infty} \tau^\sharp_{K_n}\right)(x_0) = \limsup_{n\to\infty,\, x_n \to_{K_n} x_0} \tau^\sharp_{K_n}(x_n).$$

In the same way, we define the upper epilimit $\lim^\sharp_{\uparrow n\to\infty} \varpi^\flat_{(K_n,C_n)}$ *whose epigraph is the upper limit of the epigraphs of the functions* $\varpi^\flat_{(K_n,C_n)}$

$$\mathcal{E}p\left(\lim^\sharp_{\uparrow n\to\infty} \varpi^\flat_{(K_n,C_n)}\right) := \mathrm{Limsup}_{n\to\infty} \mathcal{E}p\left(\varpi^\flat_{(K_n,C_n)}\right).$$

It is the upper epilimit *of the functions $\varpi^\flat_{(K_n,C_n)}$, equal to*

$$\left(\lim^\sharp_{\uparrow n\to\infty} \varpi^\flat_{(K_n,C_n)}\right)(x_0) = \liminf_{n\to\infty,\, x_n \to_{K_n} x_0} \varpi^\flat_{(K_n,C_n)}(x_n).$$

We have to prove this very useful stability result.

THEOREM 6.2. *Let $\mathcal{S} : X \rightsquigarrow \mathcal{C}(0,+\infty;X)$ be a strict upper semicompact map. Consider two sequences of subsets $C_n \subset C$ and $K_n \subset X$ and their Painlevé–Kuratowski upper limits*

$$C^\sharp := \mathrm{Limsup}_{n\to+\infty} C_n \;\&\; K^\sharp := \mathrm{Limsup}_{n\to+\infty} K_n.$$

Then

1. *the upper hypolimit of the upper exit functions of a sequence of subsets K_n is smaller than or equal to the upper exit function of their upper limit:*

$$\left(\lim^\sharp_{\downarrow n\to\infty} \tau^\sharp_{K_n}\right)(x) \le \tau^\sharp_{K^\sharp}(x);$$

2. *the upper epilimit of the lower constrained hitting functions of a sequence of subsets $C_n \subset K_n$ is larger than or equal to the lower constrained hitting function of their upper limit:*

$$\left(\liminf^{\sharp}_{\uparrow n\to\infty} \varpi^{\flat}_{(K_n,C_n)}\right)(x) \;\geq\; \varpi^{\flat}_{(K^\sharp,C^\sharp)}(x).$$

Proof. Let us begin by proving the first inequality, which can be translated in the form of the inclusion

$$\mathrm{Limsup}_{n\to\infty} \mathcal{H}p\left(\tau^{\sharp}_{K_n}\right) \;\subset\; \mathcal{H}p\left(\tau^{\sharp}_{K^\sharp}\right).$$

For that purpose, let us take a sequence $(T_n, x_n) \in \mathcal{H}p(\tau^{\sharp}_{K_n})$ converging to (T, x) and check that this limit belongs to the hypograph of $\tau^{\sharp}_{K^\sharp}$. By definition, there exists a solution $x_n(\cdot) \in \mathcal{S}(x_n)$ starting at x_n such that, for every $t \in [0, T_n[$, $x_n(t)$ belongs to K_n. Since $\mathcal{S}$ is upper semicompact, a subsequence (again denoted by) $x_n(\cdot)$ converges uniformly on compact intervals to some solution $x(\cdot) \in \mathcal{S}(x)$ starting at x. Take $t < T$ and n large enough for having $t < T_n$. In this case, $x_n(t)$ belongs to K_n and, passing to the limit, $x(t)$ belongs to $K^\sharp$. This implies that

$$T \;\leq\; \tau_K(x(\cdot)) \;\leq\; \tau^{\sharp}_{K^\sharp}(x).$$

Taking $K_n := K$, $x_n := x \in K$, and $T_n < \tau^{\sharp}_K(x)$ converging to $\tau^{\sharp}_K(x)$, we infer that the solution $x(\cdot)$ obtained above achieves the supremum.

Let us prove now the second inequality, which can be translated in the form of the inclusion

$$\mathrm{Limsup}_{n\to\infty} \mathcal{E}p\left(\varpi^{\flat}_{(K_n,C_n)}\right) \;\subset\; \mathcal{E}p\left(\varpi^{\flat}_{(K_n,C_n)}\right).$$

For that purpose, let us take sequences $(T_n, x_n) \in \mathcal{E}p(\varpi^{\flat}_{(K_n,C_n)})$ converging to (T, x) and check that this limit belongs to the epigraph of $\varpi_{(K^\sharp,C^\sharp)}$.

For every $\varepsilon > 0$, there exist N such that for $n \geq N$, there exists a solution $x_n(\cdot) \in \mathcal{S}(x_n)$ and $t_n \leq T_n + \frac{\varepsilon}{2} \leq T + \varepsilon$ such that $x_n(t_n) \in C_n$, and for every $s < t_n$, $x_n(s) \in K_n$. Since $\mathcal{S}$ is upper semicompact, a subsequence (again denoted by) $x_n(\cdot)$ converges uniformly on compact intervals to some solution $x(\cdot) \in \mathcal{S}(x)$. Let us consider also a subsequence (again denoted by) t_n converging to some $T^\star \leq T + \varepsilon$. By passing to the limit, we infer that $x(T^\star)$ belongs to $C^\sharp$ and that, for any $s < T^\star$, $x(s)$ belongs to $K^\sharp$. This implies that

$$\varpi^{\flat}_{(K^\sharp,C^\sharp)}(x) \;\leq\; \varpi_{(K^\sharp,C^\sharp)}(x(\cdot)) \;\leq\; T^\star \;\leq\; T + \varepsilon.$$

We conclude by letting ε converge to 0. Taking $K_n := K$, $x_n := x \in K$, and $T_n < \tau^{\sharp}_K(x)$ converging to $\tau^{\sharp}_K(x)$, we infer that the solution $x(\cdot)$ obtained above achieves the supremum.

Taking $K_n := K$, $C_n := C$, $x_n := x \in K$, and $T_n \geq \varpi^{\flat}_{(K,C)}(x)$ converging to $\varpi^{\flat}_{(K,C)}(x)$, we infer that the solution $x(\cdot)$ obtained above achieves the infimum. □

We derive stability properties of the viability kernels with targets.

THEOREM 6.3. *Let us assume that the map $\mathcal{S}$ is upper semicompact.*

If a subset K captures a subset $C \subset K$ under $\mathcal{S}$, then its closure $\overline{K}$ also captures the closure $\overline{C}$ of the target C.

More generally, let us consider a sequence of subsets K_n and of targets $C_n \subset K_n$. If K_n captures C_n for every $n \geq 0$, then the upper limit $\mathrm{Limsup}_{n\to+\infty} K_n$ *captures the upper limit* $\mathrm{Limsup}_{n\to+\infty} C_n$ *of the targets C_n.*

Proof. Let us set

$$C^\sharp := \mathrm{Limsup}_{n\to+\infty} C_n \text{ and } K^\sharp := \mathrm{Limsup}_{n\to+\infty} K_n.$$

Let us consider the limit $x := \lim_{n\to+\infty} x_n \in K^\sharp$ of elements $x_n \in K_n$. Since C_n captures K_n, there exists a solution $x_n(\cdot) \in \mathcal{S}(x_n)$ viable in K_n until it possibly reaches C_n at time $t_n := \varpi_{C_n}(x_n(\cdot))$, finite or infinite.

Since $\mathcal{S}$ is upper semicompact, a subsequence (again denoted by) $x_n(\cdot)$ converges to some $x(\cdot) \in \mathcal{S}(x)$ uniformly on compact intervals.

Since $x_n(\cdot)$ is viable in K_n until it reaches C_n, we know that

$$\varpi_{C_n}(x_n(\cdot)) \leq \tau_{K_n}(x_n(\cdot)).$$

Either the limit x belongs to the viability kernel $\mathrm{Viab}(K^\sharp)$ of the upper limit $K^\sharp$, or else this limit x does not belong to the viability kernel $\mathrm{Viab}(K^\sharp)$ and we have to check that it belongs to the viable-capture basin $\mathrm{Capt}_{K^\sharp}(C^\sharp)$. This means that $\tau^\sharp_{K^\sharp}(x)$ is finite. Since $\mathcal{S}$ is upper semicompact, Theorem 6.2 implies that

$$\limsup_{n\to+\infty, x_n \to_{K_n} x} \tau^\sharp_{K_n}(x_n) \leq \tau^\sharp_{K^\sharp}(x).$$

For n large enough, there exists $T_n < +\infty$ satisfying

$$\varpi_{C_n}(x_n(\cdot)) \leq T_n \leq \tau_{K_n}(x_n(\cdot)) \leq \tau^\sharp_{K_n}(x_n) \leq \tau^\sharp_{K^\sharp}(x) + 1 < +\infty.$$

Therefore, a subsequence (again denoted by) T_n converges to some $T^\star \leq \tau^\sharp_{K^\sharp}(x) + 1$. Theorem 6.2 implies that

$$\varpi^\flat_{(K^\sharp, C^\sharp)}(x) \leq \varpi_{(K^\sharp, C^\sharp)}(x(\cdot)) \leq T^\star \leq \tau_{K^\sharp}(x(\cdot)) \leq \tau_{K^\sharp}(x).$$

Hence from every $x \in K^\sharp$ starts a solution viable in $K^\sharp$ until it possibly reaches C, so that $K^\sharp$ captures $C^\sharp$. □

As a consequence, we obtain the following theorem.

THEOREM 6.4. *Let us assume that the map $\mathcal{S}$ is upper semicompact. Then*

$$\overline{\mathrm{Viab}(K, C)} \subset \mathrm{Viab}(\overline{K}, \overline{C}),$$

and thus, if $C \subset K$ and K are closed, so is the viability kernel $\mathrm{Viab}(K, C)$ *of K with target C.*

More generally, let us consider a sequence of subsets K_n and of targets $C_n \subset K_n$ and their upper limits $K^\sharp$ and $C^\sharp$. Then

$$\mathrm{Limsup}_{n\to+\infty} \mathrm{Viab}(K_n, C_n) \subset \mathrm{Viab}(\mathrm{Limsup}_{n\to+\infty} K_n, \mathrm{Limsup}_{n\to+\infty} C_n). \tag{6.1}$$

THEOREM 6.5. *If the set-valued map $\mathcal{S}_-$ is lower semicontinuous, then for any sequence of closed subsets C_n,*

$$\mathrm{Capt}(\mathrm{Liminf}_{n\to+\infty} C_n) \subset \mathrm{Liminf}_{n\to+\infty} \mathrm{Capt}(C_n). \tag{6.2}$$

Proof. For proving that

$$\mathrm{Capt}(\mathrm{Liminf}_{n\to+\infty} C_n) \subset \mathrm{Liminf}_{n\to+\infty} \mathrm{Capt}(C_n),$$

let $C^\flat$ denote the lower limit of the subsets C_n. Let us take $x \in \mathrm{Capt}(C^\flat)$ and a solution $x(\cdot) \in \mathcal{S}(x)$ viable in K until it reaches the target $C^\flat$ at time $T < +\infty$ at $c := x(T) \in C^\flat$. Hence the function $t \mapsto y(t) := x(T-t)$ is a solution $y(\cdot) \in \mathcal{S}_-(c)$. Let us consider a sequence of elements $c_n \in C_n$ converging to c.

Since $\mathcal{S}_-$ is lower semicontinuous, there exist solutions $y_n(\cdot) \in \mathcal{S}_-(c_n)$ converging uniformly over compact intervals to $x(\cdot)$. Therefore, $x_n := y_n(T)$ converges to x. It is enough to observe that x_n belongs to $\mathrm{Capt}(C_n)$ to conclude. □

As a consequence, we obtain the following theorem.

THEOREM 6.6. *Let us consider a sequence of closed subsets C_n satisfying* $\mathrm{Viab}(K) \subset C_n \subset K$ *and*

$$\mathrm{Lim}_{n\to+\infty} C_n \;:=\; \mathrm{Limsup}_{n\to+\infty} C_n \;=\; \mathrm{Liminf}_{n\to+\infty} C_n.$$

If the set-valued map $\mathcal{S}$ is upper semicompact, if $\mathcal{S}_-$ is lower semicontinuous, and if K is closed and backward invariant under $\mathcal{S}$, then

$$\mathrm{Lim}_{n\to+\infty} \mathrm{Capt}^K(C_n) \;=\; \mathrm{Capt}^K(\mathrm{Lim}_{n\to+\infty} C_n). \tag{6.3}$$

7. Optimal evolutionary control system. We devote this section to statements of applications to optimal control and stopping time problems. We refer to [8, 9] for applications to systems of first-order partial differential equations and inclusions.

7.1. Control evolutionary systems. We denote by $L^1(0,\infty;\mathcal{U})$ the space of measurable integrable functions from $[0,+\infty[$ to a finite dimensional vector-space $\mathcal{U}$, the *control space*. We shall supply it with the weakened topology.

DEFINITION 7.1. *Let us consider topological vector spaces X (the* state space*) and $\mathcal{U}$ (the control space). A control evolutionary system is a set-valued map $\mathcal{C} : X \rightsquigarrow \mathcal{C}(0,\infty;X) \times L^1(0,\infty;\mathcal{U})$ associating with any x a set of state-control pairs $(x(\cdot),u(\cdot))$ satisfying the following.*

1. *The* translation property. *Let $(x(\cdot),u(\cdot)) \in \mathcal{C}(x)$. Then for all $T \geq 0$, the function $(y(\cdot),v(\cdot))$ defined by $y(t) := x(t+T)$ and $v(t) := u(t+T)$ is a solution $(y(\cdot),v(\cdot)) \in \mathcal{C}(x(T))$ starting at $x(T)$.*
2. *The* concatenation property. *Let $(x(\cdot),u(\cdot)) \in \mathcal{C}(x)$, and $T \geq 0$. Then for every $(y(\cdot),v(\cdot)) \in \mathcal{C}(x(T))$, the pair $(z(\cdot),w(\cdot))$ of functions defined by*

$$z(t) \;:=\; \begin{cases} x(t) & \text{if} \quad t \in [0,T], \\ y(t-T) & \text{if} \quad t \geq T \end{cases}$$

and

$$w(t) \;:=\; \begin{cases} u(t) & \text{if} \quad t \in [0,T], \\ v(t-T) & \text{if} \quad t \geq T \end{cases}$$

belongs to $\mathcal{C}(x)$.

We shall say that the control evolutionary system $\mathcal{C}$ is upper semicompact if the set-valued map $x \rightsquigarrow \mathcal{C}(x)$ is upper semicompact from X to $\mathcal{C}(0,\infty;X) \times L_1(0,\infty;X)$.

Control evolutionary systems provide examples of evolutionary systems by setting

$$\mathcal{S}(x) \;:=\; \bigcup_{\{u(\cdot)\;|\;(x(\cdot),u(\cdot))\in\mathcal{C}(x)\}} \{x(\cdot)\}.$$

Usual control problems provide examples of control evolutionary systems.

7.2. Control systems. Let us consider a control problem (P, f) with a priori feedback map $P : X \rightsquigarrow \mathcal{U}$ from X to some finite dimensional vector space $\mathcal{U}$ governing the evolution of $(x(\cdot), u(\cdot))$ according the system

$$\left\{ \begin{array}{ll} \text{(i)} & x'(t) \ = \ f(x(t), u(t)), \\ \text{(ii)} & u(t) \ \in \ P(x(t)). \end{array} \right. \tag{7.1}$$

Starting from x, we define $\mathcal{C}_{(P,F)}(x)$ as the set of pairs $(x(\cdot), u(\cdot)) \in \mathcal{C}(0, \infty; X) \times L^1(0, \infty; \mathcal{U})$ satisfying (7.1) for almost all $t \geq 0$ such that $x(0) = 0$.

DEFINITION 7.2. *We shall say that the* control system (P, f) *is*

1. Marchaud *if the set-valued map $P : X \rightsquigarrow \mathcal{U}$ is Marchaud, if $f : X \times \mathcal{U} \mapsto X$ is continuous and affine with respect to the control, and if f satisfies the growth condition*

$$\forall\ (x, u) \in \text{Graph}(P), \ \ \|f(x, u)\| \ \leq \ c(\|x\| + \|u\| + 1);$$

2. Lipschitz *if the set-valued map $P : X \rightsquigarrow \mathcal{U}$ is Lipschitz and if $f : X \times \mathcal{U} \mapsto X$ is Lipschitz.*

Therefore, a control system (P, f) provides an example of upper semicompact evolutionary systems $\mathcal{S}$ if the control system (P, f) is Marchaud and an example of a lower semicontinuous evolutionary systems $\mathcal{S}$ if the control system (P, f) is Lipschitz.

7.3. Optimal evolutionary control. Let us introduce the following two features:

1. a discount factor

$$\mathbf{m} : (x, u) \in X \times \mathcal{U} \mapsto \mathbf{m}(x, u) \in \mathbf{R},$$

2. an extended "Lagrangian"

$$\mathbf{l} : (x, u) \in X \times \mathcal{U} \mapsto \mathbf{l}(x, u) \in \mathbf{R},$$

used to measure a cumulated cost over time.

We associate with them the auxiliary evolutionary control system $\mathcal{R}$ defined by

$$\mathcal{R}(T, x, y) \ = \ \{(T - \cdot, x(\cdot), u(\cdot), y(\cdot))\}_{(x(\cdot),u(\cdot))\in\mathcal{C}(x)},$$

where

$$y(t) \ \leq \ e^{-\int_0^t \mathbf{m}(x(s),u(s))ds} \left(y - \int_0^t e^{\int_0^\tau \mathbf{m}(x(s),u(s))ds} \mathbf{l}(x(\tau), u(\tau))d\tau \right). \tag{7.2}$$

7.4. Objective and constraints. Let us consider two nonnegative extended cost functions $\mathbf{b}$ (constraint function) and $\mathbf{c}$ (objective function) satisfying

$$\forall\ (t, x) \in \mathbf{R}_+ \times \mathbf{R}_+^{n+1}, \ \ 0 \ \leq \ \mathbf{b}(t, x) \ \leq \ \mathbf{c}(t, x) \ \leq \ +\infty,$$

allowed to take infinite values in order to describe state constraints. We extend them as functions from $\mathbf{R} \times \mathbf{R}^{n+1}$ to $\mathbf{R}_+ \cup \{+\infty\}$ by setting

$$\forall\ (t, x) \notin \mathbf{R}_+ \times \mathbf{R}_+^{n+1}, \ \ \mathbf{b}(t, x) \ = \ \mathbf{c}(t, x) \ = \ +\infty,$$

so that nonnegativity constraints on time and on the state variables are automatically taken into account. In particular, we shall denote by $\mathbf{0}$ the function defined by

$$\mathbf{0}(t, x) \ = \left\{ \begin{array}{ll} 0 & \text{if} \qquad t \geq 0, \ x \in \mathbf{R}_+^{n+1}, \\ +\infty & \text{if not.} \end{array} \right.$$

Several control problems, in particular, financial problems such as the valuation of options, are stated in the following fashion.

DEFINITION 7.3. *The two nonnegative extended constraint and objective functions being given, and given also a horizon time $T > 0$, the problem is to*

1. *find the* valuation subset $\mathcal{V} \subset \mathbf{R}_+ \times \mathbf{R}^{n+1} \times \mathbf{R}$ *of triples* (T, x, y) *made of the horizon time* T*, the initial state* x*, and the cost* y *such that there exists a control* $t \in [0, T] \mapsto u(t) \in P(x(t))$ *and a time* $\varpi(T) \in [0, T]$ *such that the solution to the system* (7.3) *satisfying* $x(0) = x$*,* $y(0) = y$ *and*

$$\left\{ \begin{array}{ll} \text{(i)} & \forall\, t \in [0, \varpi(T)],\ \ y(t) \geq \mathbf{b}(T - t, x(t)), \\ \text{(ii)} & y(\varpi(T) \geq \mathbf{c}(T - \varpi(T), x(\varpi(T))); \end{array} \right. \tag{7.3}$$

2. *associate with any initial price* x *the smallest cost*

$$V(T, x) \ := \ \inf_{(T,x,y) \in \mathcal{V}} y. \tag{7.4}$$

The function $(T, x) \mapsto V(T, x)$ *is called the* value function of the problem*, i.e., the minimal initial cost* y *satisfying the two above constraints* (7.3)*.*

We observe at once the following property. *The value function satisfies the initial condition*

$$\forall\, x \in \mathbf{R}^n,\ \ V(0, x) \ = \ \mathbf{c}(0, x).$$

We observe that *the valuation subset* $\mathcal{V}$ *is the viable-capture basin of the epigraph of* $\mathbf{c}$ *viable in the epigraph of* $\mathbf{b}$ *under the auxiliary evolutionary control system* (7.3) because dynamical constraints (7.3) can be reformulated in the form

$$\left\{ \begin{array}{ll} \text{(i)} & \forall\, t \in [0, \varpi(T)],\ \ (T - t, x(t), y(t)) \ \in \ \mathcal{E}p(\mathbf{b}), \\ \text{(ii)} & (T - \varpi(T), x(\varpi(T)), y(\varpi(T))) \ \in \ \mathcal{E}p(\mathbf{c}). \end{array} \right. \tag{7.5}$$

Therefore, we can reformulate the definition of the valuation subset $\mathcal{V}$ and of the value function $(T, x) \mapsto V(T, x)$ in the following way.

PROPOSITION 7.4. *The valuation subset*

$$\mathcal{V} \ = \ \mathrm{Capt}_{\mathcal{R}}^{\mathcal{E}p(\mathbf{b})}(\mathcal{E}p(\mathbf{c}))$$

is the viable-basin capture of the epigraph of the cost function $\mathbf{c}$ *under the auxiliary evolutionary control system* (7.3) *viable in the epigraph of the cost function* $\mathbf{b}$*.*

We can prove that V is the concealed value function of an optimal evolutionary control system that we have to unearth. For that purpose, we associate with the function $\mathbf{c}$ the *cost functional*

$$\left\{ \begin{array}{l} J_{\mathbf{c}}(t, x; (x(\cdot), u(\cdot))) \\ \\ := \ e^{\int_0^t \mathbf{m}(x(s), u(s))ds} \mathbf{c}(T - t, x(t)) + \displaystyle\int_0^t e^{\int_0^\tau \mathbf{m}(x(s), u(s))ds} \mathbf{l}(x(\tau), u(\tau))d\tau \end{array} \right.$$

(where t ranges over $[0, T]$), constituted by the sum of the discounted spot cost and the cumulated costs at time t of a solution to the control problem starting at x at the initial time. The controls—most often prices or other regulees in economics, portfolio in finance—appear *both* in the discount factor $\mathbf{m}$ and the Lagrangian $\mathbf{l}$. In

the same way, we associate with the function $\mathbf{b}$ the *cost functional* $J_{\mathbf{b}}$ and the maximal cumulated cost up to the current time t:

$$K_{\mathbf{b}}(t,x;(x(\cdot),u(\cdot))) := \sup_{s\in[0,t]} J_{\mathbf{b}}(s,x;(x(\cdot),u(\cdot))).$$

We next integrate this cumulated cost together with the former cost $J_{\mathbf{c}}(t,x;(x(s),u(s)))$ by introducing the new cost function

$$L^{\mathbf{c}}_{\mathbf{b}}(t,x;(x(\cdot),u(\cdot))) := \max(K_{\mathbf{b}}(t,x;(x(\cdot),u(\cdot))), J_{\mathbf{c}}(t,x;(x(\cdot),u(\cdot)))).$$

The problem is now to minimize over all $t\in[0,T]$ and over all the solutions to the evolutionary control problem:

$$V_{\mathbf{b}}(\mathbf{c})(T,x) := \inf_{(x(\cdot),u(\cdot))\in\mathcal{C}(x)} \inf_{t\in[0,T]} L^{\mathbf{c}}_{\mathbf{b}}(t,x;(x(\cdot),u(\cdot))).$$

STATEMENT 6. *Let us assume that the extended functions* $\mathbf{b}$ *and* $\mathbf{c}$ *are nontrivial and nonnegative. The constrained discounted intertemporal value function* $V_{\mathbf{b}}(\mathbf{c})$ *is equal to the function* V *associated with the viable-capture basin* $\mathrm{Capt}^{\mathcal{E}p(\mathbf{b})}_{\mathcal{R}}(\mathcal{E}p(\mathbf{c}))$ *of* $\mathcal{E}p(\mathbf{c})$ *under* $\mathcal{R}$. *Furthermore, any solution* $(x(\cdot),u(\cdot))\in\mathcal{C}(x)$ *starting from* $x\in \mathrm{Dom}(V_{\mathbf{b}}(\mathbf{c}))$ *satisfying the inequality for every* $t\in[0,\varpi(T,x(\cdot))]$

$$(7.6)\quad \begin{cases} V_{\mathbf{b}}(\mathbf{c})(T,x) \\ \geq\ e^{\int_0^t \mathbf{m}(x(s),u(s))ds} V_{\mathbf{b}}(\mathbf{c})(T-t,x(t)) + \displaystyle\int_0^t e^{\int_0^\tau \mathbf{m}(x(s),u(s))ds}\mathbf{l}(x(\tau),u(\tau))d\tau \end{cases}$$

until the first time $\varpi(T,x(\cdot))$ *when*

$$V_{\mathbf{b}}(\mathbf{c})(T-\varpi(T,x(\cdot))),x(\varpi(T,x(\cdot))) = \mathbf{b}(T-\varpi(T,x(\cdot))),x(\varpi(T,x(\cdot)))$$

is an optimal solution for the optimal time $\varpi(T,x(\cdot))$ *and actually satisfies the equality*

$$(7.7)\quad \begin{cases} \forall\, t\in[0,\varpi(T,x(\cdot))],\ V_{\mathbf{b}}(\mathbf{c})(T,x) \\ \\ =\ e^{\int_0^t \mathbf{m}(x(s),u(s))ds} V_{\mathbf{b}}(\mathbf{c})(T-t,x(t)) + \displaystyle\int_0^t e^{\int_0^\tau \mathbf{m}(x(s),u(s))ds}\mathbf{l}(x(\tau),u(\tau))d\tau. \end{cases}$$

Finally, the value function is a solution $\mathbf{v}$ *to the two following functional equations stating that the functions* $L^{\mathbf{v}}_{\mathbf{b}}$ *and* $L^{\mathbf{c}}_{\mathbf{v}}$ *have the same infimum as* $L^{\mathbf{c}}_{\mathbf{b}}$:

$$(7.8)\quad \begin{cases} \inf_{(x(\cdot),u(\cdot))\in\mathcal{C}(x)} \inf_{t\in[0,T]} L^{\mathbf{c}}_{\mathbf{v}}(t,x;(x(\cdot),u(\cdot))) \\ =\ \mathbf{v}(T,x) \\ =\ \inf_{(x(\cdot),u(\cdot))\in\mathcal{C}(x)} \inf_{t\in[0,T]} L^{\mathbf{v}}_{\mathbf{b}}(t,x;(x(\cdot),u(\cdot))). \end{cases}$$

We list a manifold of examples in classical optimal control (in the case when $\mathbf{m}=0$), recalling that financial problems[6] (in the case when $\mathbf{l}=0$) also fit the above framework. Playing with the choice of the spot cost $\mathbf{c}$, we shall cover several examples.

[6]See [44] for a treatment of dynamical valuation of portfolios in the framework of dynamical games.

1. Taking $\mathbf{b} = 0$ and $\mathbf{c}$ defined by

$$\mathbf{c}(t,x) := \begin{cases} \mathbf{u}(x) & \text{if} \quad t = 0, \\ +\infty & \text{if} \quad t > 0, \end{cases}$$

the above problem boils down to

$$V(\mathbf{c})(T,x) := \inf_{(x(\cdot),u(\cdot))\in\mathcal{C}(x)} J_{\mathbf{c}}(T,x;(x(\cdot),u(\cdot))),$$

which is the *Bolza problem*

$$\inf_{(x(\cdot),u(\cdot))\in\mathcal{C}(x)} \left(\mathbf{u}(x(T)) + \int_0^T \mathbf{l}(x(\tau),u(\tau))d\tau\right)$$

and the *Mayer problem*

$$\inf_{(x(\cdot),u(\cdot))\in\mathcal{C}(x)} \mathbf{u}(x(T))$$

when, furthermore, $\mathbf{l} = 0$.

2. Taking $\mathbf{b} = 0$ and $\mathbf{c}(t,x) := \mathbf{u}(x)$, we find the classical *stopping time problem*

$$\mathbf{F}(u)(T,x) := \inf_{(x(\cdot),u(\cdot))\in\mathcal{C}(x)} \inf_{t\in[0,T]} \left(\mathbf{u}(x(t)) + \int_0^t \mathbf{l}(x(\tau),u(\tau))d\tau\right)$$

associated with $\mathbf{u}$ in control theory. The cost function $\mathbf{l}$ can be regarded as a density of Maslov measures, $\mathbf{F}(\mathbf{u})$ being then the *mathematical faith* of $\mathbf{u}$ introduced by Pierre Bernhard (with the minus sign, under the name of *mathematical fear*). They correspond to the *mathematical expectation* $\mathbf{E}(f)$ of densities f of probability measures.

3. Let us consider an extended function $\mathbf{b} : \mathbf{R}_+ \times X \mapsto \mathbf{R} \cup \{+\infty\}$ with which we associate the problem

$$W(\mathbf{b})(T,x) := \inf_{(x(\cdot),u(\cdot))\in\mathcal{C}(x)} \sup_{t\in[0,T]} J_{\mathbf{b}}(t,x;(x(\cdot),u(\cdot))).$$

$\mathbf{b} : \mathbf{R}_+ \times X \mapsto \mathbf{R} \cup \{+\infty\}$ is an extended cost function. We observe that $W(\mathbf{b}) = V_{\mathbf{b}}(\mathbf{c})$, where

$$\mathbf{c}(t,x) := \begin{cases} \mathbf{b}(0,x) & \text{if} \quad t = 0, \\ +\infty & \text{if} \quad t > 0. \end{cases}$$

Indeed, we see that $J_{\mathbf{c}}(t,x;(x(\cdot),u(\cdot))) = +\infty$ if $t < T$ and $J_{\mathbf{c}}(T,x;(x(\cdot),u(\cdot))) = J_{\mathbf{b}}(T,x;(x(\cdot),u(\cdot)))$. Therefore,

$$L^{\mathbf{c}}_{\mathbf{b}}(t,x;(x(\cdot),u(\cdot))) := \begin{cases} K_{\mathbf{b}}(T,x;(x(\cdot),u(\cdot))) & \text{if} \quad t = T, \\ +\infty & \text{if} \quad t < T, \end{cases}$$

and thus, $W(\mathbf{b}) = V_{\mathbf{b}}(\mathbf{c})$.

7.5. Episolutions to Hamilton–Jacobi–Bellman inequalities. Let us consider the case when the evolutionary control system is associated with the control system (7.1) and apply Theorem 5.3 characterizing capture basins in terms of the tangential conditions. This allows us to relate the value function with generalized solutions to Hamilton–Jacobi–Bellman partial differential variational inequalities

$$\left\{ \begin{array}{l} \forall\ (t,x) \in \Omega(\mathbf{v}), \\ -\dfrac{\partial \mathbf{v}(t,x)}{\partial t} + \inf\limits_{u \in P(x)} \left(\left\langle \dfrac{\partial \mathbf{v}(t,x)}{\partial x}, f(x,u) \right\rangle + \mathbf{l}(x,u) + \mathbf{m}(x,u)\mathbf{v}(t,x) \right) = 0 \end{array} \right.$$

on the subset

$$\Omega(\mathbf{v}) \ := \ \{(t,x) \in \mathbf{R}_+ \times X \mid \mathbf{b}(t,x) \ \leq \ \mathbf{v}(t,x) \ < \ \mathbf{c}(t,x)\}.$$

Let us recall that the *contingent epiderivative* $D_\uparrow \mathbf{v}(t,x)$ of $\mathbf{v}$ at (t,x) is defined by

$$D_\uparrow \mathbf{v}(t,x)(\lambda,v) \ := \ \liminf_{h \to 0+,\ u \to v} \frac{\mathbf{v}(t+h\lambda, x+hu)}{h}$$

and that

$$\mathcal{E}p(D_\uparrow \mathbf{v}(t,x)) \ = \ T_{\mathcal{E}p(\mathbf{v})}(t,x,\mathbf{v}(t,x)).$$

The first part of Theorem 5.3 implies a characterization of the value function as a solution of Hamilton–Jacobi variational inequalities.

STATEMENT 7 (Frankowska). *Let us assume that the control system* $(P,f,\mathbf{l},\mathbf{m})$ *is Marchaud and that the functions* $\mathbf{b}$ *and* $\mathbf{c}$ *are nontrivial, nonnegative, and lower semicontinuous.*

Then the value function $V_{\mathbf{b}}(\mathbf{c})$ *is characterized as the smallest of the nonnegative lower semicontinuous functions* $\mathbf{v} : \mathbf{R}_+ \times X \mapsto \mathbf{R}_+ \cup \{+\infty\}$ *satisfying for every* $(t,x) \in]0,\infty[\times X$

$$\left\{ \begin{array}{ll} \text{(i)} & \mathbf{b}(t,x) \ \leq \ \mathbf{v}(t,x) \ \leq \ \mathbf{c}(t,x), \\ \text{(ii)} & \text{if } (t,x) \in \Omega(\mathbf{v}), \\ & \inf_{u \in P(x)} \left(D_\uparrow \mathbf{v}(t,x)(-1, f(x,u)) + \mathbf{l}(x,u) + \mathbf{m}(x,u)\mathbf{v}(t,x) \right) \ \leq \ 0. \end{array} \right.$$

Let us set

$$\mathbf{R}(t,x) \ := \ \{u \in P(x) \mid D_\uparrow V_{\mathbf{b}}(\mathbf{c})(t,x)(-1,f(x,u)) + \mathbf{l}(x,u) + \mathbf{m}(x,u) V_{\mathbf{b}}(\mathbf{c})(t,x) \leq 0\}.$$

Knowing the value function, an optimal solution is obtained in the following way. Starting from x_0 *such that* $V_{\mathbf{b}}(\mathbf{c})(T,x_0) \ < \ \mathbf{c}(T,x_0)$*, any solution* $(x(\cdot),u(\cdot))$ *to the control system*

$$\left\{ \begin{array}{ll} \text{(i)} & x'(t) \ = \ f(x(t),u(t)), \\ \text{(ii)} & u(t) \ \in \ \mathbf{R}(t,x(t)), \end{array} \right. \tag{7.9}$$

is an optimal solution, and the first time $\varpi(T,x(\cdot)) \geq 0$ *when*

$$V_{\mathbf{b}}(\mathbf{c})(T - \varpi(T,x(\cdot)), x(\varpi(T,x(\cdot)))) \ = \ \mathbf{c}(T - \varpi(T,x(\cdot)), x(\varpi(T,x(\cdot))))$$

is the optimal time.

The second part of Theorem 5.3 implies the characterization of the value function $V_{\mathbf{b}}(\mathbf{c})$ as a unique Frankowska *episolution* to the Hamilton–Jacobi–Bellman variational inequality.

STATEMENT 8 (Frankowska). *Let us assume that the control system (P, f) is Marchaud and Lipschitz and that* $\mathbf{b}$ *and* $\mathbf{c}$ *are nontrivial, nonnegative, and lower semicontinuous.*

Then the value function $V_{\mathbf{b}}(\mathbf{c})$ is the unique lower semicontinuous episolution $\mathbf{v}$ *to the system of differential inequalities: for every $(t, x) \in \mathrm{Dom}(\mathbf{v})$,*

$$
(7.10)\quad \left\{ \begin{array}{ll}
\text{(i)} & \mathbf{b}(t,x) \le \mathbf{v}(t,x) \le \mathbf{c}(t,x), \\
\text{(ii)} & \text{if } \mathbf{v}(t,x) < \mathbf{c}(t,x), \\
 & \inf_{u \in P(x)} (D_\uparrow \mathbf{v}(t,x)(-1, f(x,u)) + \mathbf{l}(x,u) + \mathbf{m}(x,u)\mathbf{v}(t,x)) \le 0, \\
\text{(iii)} & \text{if } \mathbf{v}(t,x) > \mathbf{b}(t,x), \\
 & \sup_{u \in P(x)} (D_\uparrow \mathbf{v}(t,x)(1, -f(x,u)) - \mathbf{l}(x,u) - \mathbf{m}(x,u)\mathbf{v}(t,x)) \le 0, \\
\text{(iv)} & \text{if } \mathbf{v}(t,x) = \mathbf{b}(t,x), \\
 & \sup_{u \in P(x)} [\min(D_\uparrow \mathbf{v}(t,x)(1, -f(x,u)), D_\downarrow \mathbf{b}(t,x)(1, -f(x,u))) \\
 & -\mathbf{l}(x,u)) - \mathbf{m}(x,u)\mathbf{v}(t,x)))] \le 0.
\end{array} \right.
$$

Remark. Condition (7.10)(iv) is automatically satisfied whenever

$$
\sup_{u \in P(x)} (D_\downarrow \mathbf{b}(t,x)(1, -f(x,u)) - \mathbf{l}(x,u)) - \mathbf{m}(x,u)\mathbf{v}(t,x) \le 0,
$$

i.e., whenever the epigraph of $\mathbf{b}$ is locally backward invariant under the auxiliary system. □

7.6. Bilateral and viscosity solutions to Hamilton–Jacobi–Bellman variational inequalities. We obtain by duality equivalent statements involving subdifferential and/or superdifferentials, involving the Hamiltonian $H : X \times \mathbf{R}_+ \times X^\star \mapsto \mathbf{R} \cup \{+\infty\}$ associated with the control problem and the Lagrangian by

$$
H(x, y, p) := \inf_{u \in P(x)} (\langle p, f(x,u) \rangle + \mathbf{l}(x,u) + \mathbf{m}(x,u)y)
$$

and the horizon Hamiltonian $H^\infty : X \times X^\star \mapsto \mathbf{R} \cup \{+\infty\}$, by

$$
H^\infty(x, p) := \inf_{u \in P(x)} \langle p, f(x,u) \rangle.
$$

We recall the definition of the subdifferential $\partial_- \mathbf{v}(t,x)$ and the horizon subdifferential $\partial_-^\infty \mathbf{v}(t,x)$ of the function $\mathbf{v}$ at (t,x):

$$
\left\{ \begin{array}{ll}
\text{(i)} & (p_t, p_x) \in \partial_- \mathbf{v}(t,x) \text{ if } (p_t, p_x, -1) \in N_{\mathcal{E}p(\mathbf{v})}(t, x, \mathbf{v}(t,x)), \\
\text{(ii)} & (p_t, p_x) \in \partial_-^\infty \mathbf{v}(t,x) \text{ if } (p_t, p_x, 0) \in N_{\mathcal{E}p(\mathbf{v})}(t, x, \mathbf{v}(t,x)).
\end{array} \right.
$$

Let us recall that *the horizon subdifferential $\partial_-^\infty \mathbf{v}(t,x) = (0,0)$ whenever the domain of the contingent epiderivative $D_\uparrow \mathbf{v}(t,x)$ is dense in $\mathbf{R}_+ \times X$. This happens whenever* $\mathbf{v}$ *is Lipschitz in a neighborhood of* (t,x).

STATEMENT 9 (Frankowska). *Under the assumptions of Statement* 7*, the value function $V_{\mathbf{b}}(\mathbf{c})$ is the smallest lower semicontinuous nonnegative function $\mathbf{v} : X \mapsto \mathbf{R} \cup \{+\infty\}$ satisfying for every $(t,x) \in]0, \infty[\times X$*

$$
\left\{ \begin{array}{ll}
\text{(i)} & \mathbf{b}(t,x) \le \mathbf{v}(t,x) \le \mathbf{c}(t,x), \\
\text{(ii)} & \text{if } \mathbf{v}(t,x) < \mathbf{c}(t,x), \\
 & \forall\, (p_t, p_x) \in \partial_- \mathbf{v}(t,x), \quad -p_t + H(x, \mathbf{v}(t,x), p_x) \le 0, \\
 & \forall\, (p_t, p_x) \in \partial_-^\infty \mathbf{v}(t,x), \quad -p_t + H^\infty(x, p_x) \le 0.
\end{array} \right.
$$

Statement 8 can be stated in terms of subdifferentials, providing the existence and uniqueness of bilateral solutions proved independently by Barron and Jensen and Frankowska.

STATEMENT 10 (Barron–Jensen and Frankowska). *We posit the assumptions of Statement* 8. *Then the value function* $V_{\mathbf{b}}(\mathbf{c})$ *is the* unique lower semicontinuous solution $\mathbf{v}$—*also called* bilateral solution—*to the system of differential inequalities: for every* $(t,x) \in \mathrm{Dom}(\mathbf{v})$,

$$
(7.11)\quad \left\{\begin{array}{ll}
\text{(i)} & \mathbf{b}(t,x) \le \mathbf{v}(t,x) \le \mathbf{c}(t,x), \\
\text{(ii)} & \textit{if } \mathbf{b}(t,x) < \mathbf{v}(t,x) < \mathbf{c}(t,x), \textit{ the equations} \\
& \quad \forall\, (p_t,p_x) \in \partial_- \mathbf{v}(t,x), \quad -p_t + H(x,\mathbf{v}(t,x),p_x) = 0, \\
& \quad \forall\, (p_t,p_x) \in \partial_-^\infty \mathbf{v}(t,x), \quad -p_t + H^\infty(x,p_x) = 0, \\
\text{(iii)} & \textit{if } \mathbf{v}(t,x) = \mathbf{c}(t,x), \textit{ the boundary condition} \\
& \quad \forall\, (p_t,p_x) \in \partial_- \mathbf{v}(t,x), \quad -p_t + H(x,\mathbf{v}(t,x),p_x) \ge 0, \\
& \quad \forall\, (p_t,p_x) \in \partial_- \mathbf{v}^\infty(t,x), \quad -p_t + H^\infty(x,p_x) \ge 0, \\
\text{(iv)} & \textit{if } \mathbf{b}(t,x) = \mathbf{v}(t,x), \textit{ the boundary condition} \\
& \quad \forall\, (p_t,p_x) \in \partial_-^\infty \mathbf{v}(t,x) \cap -\partial_+^\infty \mathbf{b}(t,x), \quad -p_t + H^\infty(x,p_x) = 0.
\end{array}\right.
$$

See [20, 21, 39, 40].

REFERENCES

[1] J.-P. AUBIN, *Smooth and heavy solutions to control problems*, in Nonlinear and Convex Analysis, Lecture Notes in Pure and Appl. Math. 107, B.-L. Lin and S. Simons, eds., Marcel Dekker, New York, 1987.

[2] J.-P. AUBIN, *Viability Theory*, Birkhäuser, Boston, Basel, Berlin, 1991.

[3] J.-P. AUBIN, *Dynamic Economic Theory: A Viability Approach*, Springer-Verlag, New York, 1997.

[4] J.-P. AUBIN, *Mutational and Morphological Analysis: Tools for Shape Regulation and Morphogenesis*, Birkhäuser Boston, Boston, 1999.

[5] J.-P. AUBIN, *Impulse Differential Inclusions and Hybrid Systems: A Viability Approach*, Lecture notes, University of California Berkeley, Berkely, CA, 1999.

[6] J.-P. AUBIN, *Boundary-value problems for systems of first-order partial differential inclusions*, NoDEA Nonlinear Differential Equations Appl., 7 (2000), pp. 61–84.

[7] J.-P. AUBIN, N. BONNEUIL, AND F. MAURIN, *Non-linear structured population dynamics with co-variables*, Math. Popul. Studies, 91 (2000), pp. 1–31.

[8] J.-P. AUBIN, *Systems of first-order partial differential inclusions with constraints*, in Proceedings of Evolution Equations, Trento, Italy, 2000, to appear.

[9] J.-P. AUBIN, *Boundary-value problems for systems of Hamilton–Jacobi–Bellman inclusions with constraints*, SIAM J. Control Optim., submitted.

[10] J.-P. AUBIN AND H. FRANKOWSKA, *Set-Valued Analysis*, Birkhäuser, Boston, Basel, Berlin, 1990.

[11] J.-P. AUBIN AND G. DA PRATO, *Contingent solutions to the center manifold equation*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 9 (1992), pp. 13–28.

[12] J.-P. AUBIN AND H. FRANKOWSKA, *Inclusions aux dérivées partielles gouvernant des contrôles de rétroaction*, C. R. Acad. Sci. Paris Sér. I Math., 311 (1990), pp. 851–856.

[13] J.-P. AUBIN AND H. FRANKOWSKA, *Systèmes hyperboliques d'inclusions aux dérivées partielles*, C. R. Acad. Sci. Paris Sér. I Math., 312 (1991), pp. 271–276.

[14] J.-P. AUBIN AND H. FRANKOWSKA, *Hyperbolic systems of partial differential inclusions*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 18 (1992), pp. 541–562.

[15] J.-P. AUBIN AND H. FRANKOWSKA, *Partial differential inclusions governing feedback controls*, J. Convex Anal., 2 (1995), pp. 19–40.

[16] J.-P. AUBIN AND H. FRANKOWSKA, *Set-valued Solutions to the Cauchy problem for hyperbolic systems of partial differential inclusions*, NoDEA Nonlinear Differential Equations Appl., 4 (1997), pp. 149–168.

[17] J.-P. AUBIN, J. LYGEROS, M. QUINCAMPOIX, S. SASTRY, AND N. SEUBE, *Impulse differential inclusions: A viability approach to hybrid systems*, IEEE Trans. Automat. Control, to appear.

[18] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions to Hamilton-Jacobi-Bellman Equations*, Birkhäuser Boston, Boston, 1997.

[19] E. N. Barron and R. Jenson, *Semicontinuous viscosity solutions for Hamilton-Jacobi equations with convex Hamiltonians*, Comm. Partial Differential Equations, 15 (1990), pp. 1713–1742.

[20] E. N. Barron and R. Jensen, *Relaxed minimax control*, SIAM J. Control Optim., 33 (1995), pp. 1028–1039.

[21] E. N. Barron and R. Jensen, *Relaxation of constrained control problems*, SIAM J. Control Optim., 34 (1996), pp. 2077–2091.

[22] P. Cardaliaguet, *Conditions suffisantes de non vacuité du noyau de viabilité*, C. R. Acad. Sci. Paris Sér I Math., 314 (1992), pp. 797–800.

[23] P. Cardaliaguet, *Nonemptiness of the Viability Kernel of a Differential Inclusion*, Cahiers du CEREMADE 9363, Université Paris-Dauphine, Paris, France, 1992.

[24] P. Cardaliaguet, *Domaines dicriminants en jeux différentiels*, Thèse de l'Université de Paris-Dauphine, Paris, France, 1994.

[25] P. Cardaliaguet, *Nonsmooth semi-permeable barriers, Isaacs' equation, and application to a differential game with one target and two players*, Appl. Math. Optim., 36 (1997), pp. 125–146.

[26] P. Cardaliaguet, *On the regularity of semipermeable surfaces in control theory with application to the optimal exit-time problem (part* I*)*, SIAM J. Control Optim., 35 (1997), pp. 1638–1652.

[27] P. Cardaliaguet, *On the regularity of semipermeable surfaces in control theory with application to the optimal exit-time problem (part* II*)*, SIAM J. Control Optim., 35 (1997), pp. 1653–1671.

[28] P. Cardaliaguet, *On front propagation problems with nonlocal terms*, Adv. Differential Equations, 5 (2000), pp. 213–268.

[29] P. Cardaliaguet, M. Quincampoix, and P. Saint-Pierre, *Temps optimaux pour des problèmes avec contraintes et sans contrôlabilité locale*, C. R. Acad. Sci. Paris Sér I Math., 318 (1994), pp. 607-612

[30] P. Cardaliaguet, M. Quincampoix, and P. Saint-Pierre, *Contribution à l'étude des jeux différentiels quantitatifs et qualitatifs avec contrainte sur l'état*, C. R. Acad. Paris Sér I Math., 321 (1995), pp. 1543–1548.

[31] P. Cardaliaguet, M. Quincampoix, and P. Saint-Pierre, *Set-valued numerical methods for optimal control and differential games*, in Stochastic and Differential Games. Theory and Numerical Methods, Ann. Internat. Soc. Dynam. Games 4, Birkhäuser Boston, Boston, 1999, pp. 177–247.

[32] M. G. Crandall and P.-L. Lions, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[33] G. Dal Maso and H. Frankowska, *Value functions for Bolza problems with discontinuous Lagrangians and Hamilton-Jacobi inequalities*, ESAIM Control Optim. Calc. Var., to appear.

[34] L. Doyen, L. Najman, and J. Mattioli, *Mutational equations of morphological dilation tubes*, J. Math. Imaging Vision, 5 (1995), pp. 319–230.

[35] H. Frankowska, *L'équation d'Hamilton-Jacobi contingente*, C. R. Acad. Paris Sér. I Math., 304 (1987), pp. 295–298.

[36] H. Frankowska, *Optimal trajectories associated to a solution of contingent Hamilton-Jacobi equations*, in Proceedings of the 26th IEEE Conference on Decision and Control, Los Angeles, CA, 1987.

[37] H. Frankowska, *Optimal trajectories associated to a solution of contingent Hamilton-Jacobi equations*, Appl. Math. Optim., 19 (1989), pp. 291–311.

[38] H. Frankowska, *Hamilton-Jacobi equation: Viscosity solutions and generalized gradients*, J. Math. Anal. Appl., 141 (1989), pp. 21–26.

[39] H. Frankowska, *Lower semicontinuous solutions to Hamilton-Jacobi-Bellman equations*, in Proceedings of the 30th IEEE Conference on Decision and Control, Brighton, UK, 1991, pp. 11–13.

[40] H. Frankowska, *Lower semicontinuous solutions of Hamilton–Jacobi–Bellman equation*, SIAM J. Control Optim., 31 (1993), pp. 257–272.

[41] H. G. Guseinov, A. I. Subbotin, and V. N. Ushakov, *Derivatives for multivalued mappings with applications to game-theoretical problems of control*, Problems of Control and Information Theory, 14 (1985), pp. 155–168.

[42] G. Haddad, *Monotone trajectories of differential inclusions with memory*, Israel J. Math., 39 (1981), pp. 83–100.

[43] G. Haddad, *Monotone viable trajectories for functional differential inclusions*, J. Differential Equations, 42 (1981), pp. 1–24.
[44] D. Pujal, *Evaluation et gestion dynamique de portefeuilles*, Thesis, Université Paris-Dauphine, Paris, France, 2000.
[45] M. Quincampoix, *Enveloppes d'invariance pour des inclusions différentielles Lipschitziennes: Applications aux problèmes de cibles*, C. R. Acad. Sci. Paris Sér I. Math., 314 (1992), pp. 343–347.
[46] M. Quincampoix, *Frontières de domaines d'invariance et de viabilité pour des inclusions différentielles avec contraintes*, C. R. Acad. Sci. Paris Sér. I Math., 311 (1990), pp. 411–416.
[47] M. Quincampoix and P. Saint-Pierre, *An algorithm for viability kernels in Hölderian case: Approximation by discrete viability kernels*, J. Math. Syst. Estim. Control, 5 (1995), pp. 115–118.
[48] M. Quincampoix and V. Veliov, *Viability with a target: Theory and applications*, in Applications of Mathematics in Engineering, Heron, Sofia, 1998, pp. 47–54.
[49] R. T. Rockafellar and R. Wets, *Variational Analysis*, Springer-Verlag, New York, 1997.
[50] P. Saint-Pierre, *Approximation of the viability kernel*, Appl. Math. Optim., 29 (1994), pp. 187–209.
[51] S.-Z. Shi, *Théorèmes de viabilité pour les inclusions aux dérivées partielles*, C. R. Acad. Sci. Paris Sér. I Math., 303 (1986), pp. 11–14.
[52] S.-Z. Shi, *Viability Theory for Partial Differential Inclusions*, Cahier de MD 8601, Université Paris-Dauphine, Paris, France, 1986.
[53] S.-Z. Shi, *Nagumo type condition for partial differential inclusions*, Nonlinear Anal., 12 (1988), pp. 951–967.
[54] S.-Z. Shi, *Optimal control of strongly monotone variational inequalities*, SIAM J. Control Optim., 26 (1988), pp. 274–290.
[55] S.-Z. Shi, *Viability theorems for a class of differential-operator inclusions*, J. Differential Equations, 79 (1989), pp. 232–257.
[56] H. M. Soner and N. Touzi, *Stochastic target problems, dynamical programming and viscosity solutions*, ESAIM Control Optim. Calc. Var., to appear.
[57] V. M. Veliov, *Sufficient conditions for viability under imperfect measurement*, Set-Valued Anal., 1 (1993), pp. 305–317.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 3, pp. 882–897

ON PERSISTENT EXCITATION FOR LINEAR SYSTEMS WITH STOCHASTIC COEFFICIENTS*

DAVID LEVANONY† AND PETER E. CAINES‡

Abstract. A Wiener input process is shown to be persistently exciting (PE) for linear stochastic systems with time-varying, convergent, random coefficients, provided asymptotic noise controllability holds a.s. The PE result is in the sense that the minimum eigenvalue of the integrated outer product of the state process is of $O(t)$ (t being the upper time limit of the integral).

Key words. persistent excitation, linear stochastic systems, convergent coefficients, noise controllability

AMS subject classifications. 60G15, 60H30, 93B05, 93C05, 93E12, 93E35

PII. S0363012996300458

1. Introduction.

A central issue in the identification of stochastic systems by such methods as maximum likelihood estimation, recursive least squares, and, more generally, prediction error methods (see, e.g., Caines (1988) and the references therein) is whether the control input and the system disturbance input will excite the system state process sufficiently for consistent parameter estimation to occur. More specifically, many strong consistency results for the estimation of system parameters depend upon the satisfaction of some form of asymptotic condition on the minimum and, in some cases, the maximum and the minimum eigenvalues of the matrix process $\{P_t\}$ given by the integrated outer product of the system state process (or an equivalent observed regression process) with itself. Such a condition is called a *persistent excitation condition* on the state (respectively, regression) process with respect to the given system inputs.

The use of various forms of persistent excitation conditions is ubiquitous in the system identification field (see, e.g., Caines (1988) and the bibliography therein). To cite just a few examples, applications of such conditions appear in Duncan and Pasik-Duncan (1986), where an assumption implying that $\lambda_{max}(P_t)/\lambda_{min}(P_t) = O(1)$ is employed, and in Caines (1992), where a condition of the type $\lambda_{min}(P_t)/t = O(1)$ is assumed. The well-known persistent excitation condition developed by Lai and Wei (1982), which involves the ratio of the logarithm of the maximum eigenvalue to the value of the minimum eigenvalue, is discussed in the remark at the end of this paper. In this context, we note the persistent excitation results for linear, time-invariant, systems presented by Moore (1987), which show the value of stochastic excitation signals in the identification and adaptive control of otherwise deterministic systems.

In this paper we analyze time-varying stochastic linear systems that are driven solely by a Wiener process, and whose stochastic system matrix converges a.s. in such a way that the resulting system pair $[A_\infty, C]$ is a.s. controllable from the disturbance

*Received by the editors March 13, 1996; accepted for publication (in revised form) March 29, 2001; published electronically October 31, 2001. This research was partly supported by NSERC, Canada.

http://www.siam.org/journals/sicon/40-3/30045.html

†Department of Electrical and Computer Engineering, Ben-Gurion University, Beer Sheva 84105, Israel (levanony@ee.bgu.ac.il). This work was partly performed at the Department of Electrical Engineering, McGill University.

‡Department of Electrical Engineering, McGill University, Montreal, QC, H2A 2A7, Canada and Canadian Institute for Advanced Research (peterc@cim.mcgill.edu).

input. We show that the state process of such a system satisfies the persistent excitation condition that the minimum eigenvalue of P_t is of $O(t)$ a.s. as $t \to \infty$. Apart from its general applicability, this result plays a key role in the theory of stochastic adaptive control developed in Caines and Levanony (1993) and Levanony and Caines (1994, 1996).

2. Main result.

THEOREM 2.1. *Let $\{x_t\}$ be an $\mathbf{R}^n$ valued process, evolving according to*

$$dx_t = A_t x_t dt + C dw_t, \quad t \geq 0, \quad x_0 \text{ given,} \tag{2.1}$$

where $\{w_t\}$ is an $\mathbf{R}^m$ valued standard Brownian motion, and $\{A_t\}$ an $\mathbf{R}^{n\times n}$ valued $\{\mathbf{F}_t^x\}$ adapted process. Suppose that $\{A_t\}$ is a.s. continuous and that there exists an a.s. finite A_∞ such that

$$A_t \to A_\infty, \quad \text{a.s., as } t \to \infty. \tag{2.2}$$

Then if the pair $[A_\infty, C]$ is a.s. controllable, it follows that

$$\liminf_{t\to\infty} \frac{1}{t} \lambda_{min} \left\{ \int_0^t x_r x_r^T dr \right\} > 0, \quad \text{a.s.} \tag{2.3}$$

Proof. The proof consists of several steps, some of which are constructed to overcome a key problem, namely, how to handle the possibility of trajectories $\{x_t\}$ diverging to infinity.

2.1. A truncation mechanism. First, note that continuity and convergence imply that

$$\sup_{t\geq 0} \|A_t\| < \infty, \qquad \text{a.s.} \tag{2.4}$$

Hence it suffices to prove (2.3) for stopped versions $\{A_t^N\}$ of $\{A_t\}$ which satisfy

$$\sup_{t\geq 0} \|A_t^N\| \leq N, \qquad \text{a.s.} \tag{2.5}$$

for a deterministic arbitrary large $N < \infty$. This is proved by the following (standard) argument. Fix $\varepsilon > 0$ and let N be such that

$$\sup_{t\geq 0} \|A_t\| \leq N, \quad \text{with probability greater than } 1-\varepsilon.$$

Let $T_N = \inf\{t \geq 0 : \|A_t\| > N\}$ and define $\{x_t^N\}$ so as to satisfy

$$\begin{aligned} x_t^N &= x_t, \quad 0 \leq t < T_N, \\ dx_t^N &= A_{T_N} x_t^N dt + C dw_t, \quad T_N \leq t < \infty, \ x_{T_N}^N = x_{T_N}. \end{aligned}$$

Suppose now that (2.3) holds for $\{x_t^N\}$. Let $\Gamma_N \underline{\Delta} \{x_t = x_t^N, t \geq 0\}$, where, by the choice of N, $P(\Gamma_N) > 1-\varepsilon$, it then follows that (2.3) holds for $\{x_t\}$, with probability greater than $1-\varepsilon$. The arbitrary nature of $\varepsilon > 0$ then implies that if (2.4) holds and (2.3) holds for $\{x_t^N\}$ for all $N > 0$, then (2.3) holds for $\{x_t\}$.

At this point it is worth emphasizing that $A_t^N \to A_\infty^N$ a.s. (where $\{A_t^N\}$ denotes the dynamic matrix process for the truncated process $\{x_t^N\}$), where

$$A_\infty^N = \begin{cases} A_\infty & \text{a.e. on } \Gamma_N, \\ A_{T_N} & \text{otherwise.} \end{cases}$$

Therefore, while the pair $[A_\infty^N, C]$ is a.e. controllable on Γ_N, $[A_\infty^N, C]$ might be uncontrollable on Γ_N^c. (This point will reappear at the end of the proof.)

Henceforth we continue our investigation for the truncated process x^N. To avoid cumbersome notation, the superscript N will be omitted throughout, unless specifically required.

2.2. A consequence of the contradiction to the theorem. Assume that (2.3) does not hold. Then there exists an event $\widetilde{\Omega} \in \mathbf{F}_\infty^x$ with $P(\widetilde{\Omega}) > 0$ such that $\liminf_{t\to\infty} \frac{1}{t}\lambda_{min}\{\int_0^t x_r x_r^T dr\} = 0$ a.e. on $\widetilde{\Omega}$. Hence there exists an $\mathbf{R}^n$ vector process $\alpha_t \in \mathbf{F}_\infty^x, t \geq 0$, with

$$\|\alpha_t\| = \begin{cases} 1 & \text{a.e. on } \widetilde{\Omega}, \\ 0 & \text{a.e. on } \widetilde{\Omega}^c, \end{cases} \tag{2.6}$$

such that

$$\liminf_{t\to\infty} \frac{1}{t}\int_0^t (\alpha_t^T x_r)^2 dr = 0, \quad \text{a.s.} \tag{2.7}$$

The equality (2.7) in turn implies that

$$\liminf_{t\to\infty} \int_t^{t+1} (\alpha_{t+1}^T x_r)^2 dr = \liminf_{t\to\infty} \int_0^1 (\alpha_{t+1}^T x_{t+s})^2 ds = 0, \quad \text{a.s.} \tag{2.8}$$

2.3. Construction of a convergent subsequence. Equation (2.8) implies the existence of a subsequence $\{t_n\}_{n\geq 1}$ and an associated subsequential limit $\alpha \in \mathbf{R}^n$, $\|\alpha\| \leq 1$, such that $\alpha_{t_n+1} \to \alpha$ a.s. and

$$\lim_{n\to\infty} \int_0^1 (\alpha_{t_n+1}^T x_{t_n+s})^2 ds = 0, \quad \text{a.s.} \tag{2.9}$$

With $\{x_t\}$ being a strong solution of the linear SDE (2.1) (for which no finite explosion time exists), and due to (2.9), one may conclude that

$$\sup_{n\geq 1} \int_0^1 (\alpha_{t_n+1}^T x_{t_n+s})^2 ds < \infty, \quad \text{a.s.} \tag{2.10}$$

One now may assume that $E[\sup_{n\geq 1} \int_0^1 (\alpha_{t_n+1}^T x_{t_n+s})^2 ds] < \infty$, because if not, then by a truncation mechanism identical to the one described by (2.23)–(2.26) below, one can redefine the vector process $\{\alpha_{t_n+1}\}$ so as to make the left-hand side of (2.10) integrable (see below for details, omitted here to avoid repetition).

It now follows from the integrability above, (2.9), the a.s. convergence of $\{\alpha_{t_n+1}\}$, and Corollary II-2.4 in Revuz and Yor (1991) that

$$\lim_{n\to\infty} E\left[\int_0^1 (\alpha_{t_n+1}^T x_{t_n+s})^2 ds + \|\alpha_{t_n+1} - \alpha\| | \mathbf{F}_{t_n}^x\right] = 0, \quad \text{a.s.}, \tag{2.11}$$

which enables us to define a sequence of stopping times in the form:

$$\tau_n = \inf\left\{t > 0 : E\left[\int_0^1 (\alpha_{t+1}^T x_{t+s})^2 ds + \|\alpha_{t+1} - \alpha\| | \mathbf{F}_t^x\right] < 1/n\right\}. \tag{2.12}$$

Then, obviously, $\{\tau_n\}$ is an $\{\mathbf{F}_t^x\}$-adapted subsequence which satisfies

$$\lim_{n\to\infty} E\left[\int_0^1 (\alpha_n^T x_{\tau_n+s})^2 ds | \mathbf{F}_{\tau_n}^x\right] = 0, \quad \text{a.s.}, \tag{2.13}$$

where $\alpha_n \underline{\Delta} \alpha_{\tau_n+1}$. Note that

$$0 \le E\left[\lim_{n\to\infty}\int_0^1 (\alpha_n^T x_{\tau_n+s})^2 ds\right] = \lim_{n\to\infty} E\left[\int_0^1 (\alpha_n^T x_{\tau_n+s})^2 ds\right]$$
$$= \lim_{n\to\infty} E\left\{E\left[\int_0^1 (\alpha_n^T x_{\tau_n+s})^2 ds | \mathbf{F}_{\tau_n}^x\right]\right\} = E\left\{\lim_{n\to\infty} E\left[\int_0^1 (\alpha_n^T x_{\tau_n+s})^2 ds | \mathbf{F}_{\tau_n}^x\right]\right\} = 0.$$

The first equality is due to (2.9), (2.10) (together with its assumed integrability), and dominated convergence. The second equality is based on smoothing, while the third rests on (2.12) and dominated convergence. Finally, the last equality results from (2.13). This chain of calculations therefore leads one to conclude that

$$\lim_{n\to\infty}\int_0^1 (\alpha_n^T x_{\tau_n+s})^2 ds = 0, \quad \text{a.s.} \tag{2.14}$$

Similarly, (2.11) and (2.12) imply $\alpha_n \to \alpha$, a.s. (The replacement of the sequence $\{t_n\}$ with the *adapted* subsequence $\{\tau_n\}$ is made to enable the future use of conditional expectations, which, being $\mathbf{F}_{\tau_n}^x$-measurable approximations, will allow relatively simple calculations.)

We shall show that, under the hypotheses of the theorem, (2.14) yields a contradiction.

2.4. Calculation of bounds. Let $\phi_n(s,u)$ be the transition matrix generated by $\{A_{\tau_n+r}, u \le r \le s\}$, where $\dot{\phi}_n(r,u) = A_{\tau_n+r}\phi_n(r,u)$ and $\phi_n(r,r) = I, u \le r \le s$. Define

$$z_n(s) = \phi_n^T(s,0)\alpha_n,$$
$$m_n(s) = \int_0^s \phi_n(0,r) C dw_{\tau_n+r}.$$

Then the integrand in (2.14) can be rewritten as

$$(\alpha_n^T x_{\tau_n+s})^2 = [z_n^T(s)(x_{\tau_n} + m_n(s))]^2, \qquad 0 \le s \le 1. \tag{2.15}$$

Note now that the boundedness of $\{A_t\}$ given in (2.5) implies that

$$\|\phi_n(s,u)\| \le e^N \quad \forall n \ge 0,\ u,s \in [0,1], \quad \text{a.s.} \tag{2.16}$$

For the martingale $\{m_n(s), \mathbf{F}_{\tau_n+s}^x\}$ one therefore has

$$E \sup_{0\le s\le 1} \|m_n(s)\|^2 = E \sup_{0\le s\le 1} \left\|\int_0^s \phi_n(0,r) C dw_{\tau_n+r}\right\|^2$$
$$\le 4E\left\|\int_0^1 \phi_n(0,r) C dw_{\tau_n+r}\right\|^2$$
$$= 4Tr\left\{E\int_0^1 \phi_n(0,r) CC^T \phi_n^T(0,r) dr\right\}$$
$$\le 4e^{2N} TrCC^T \qquad \forall n \ge 0, \tag{2.17}$$

where the first inequality is Doob's inequality and the last inequality is due to (2.16).

It follows that

$$\sup_{n\geq 0}\int_0^1 \|m_n(s)\|^2 ds < \infty, \quad \text{a.s.} \tag{2.18}$$

Furthermore, (2.16) and the definition of α_n imply that

$$\sup_{n\geq 0}\sup_{0\leq s\leq 1} \|z_n(s)\| \leq e^N, \quad \text{a.s.}, \tag{2.19}$$

and hence, via (2.18),

$$\sup_{n\geq 0}\int_0^1 (z_n^T(s)m_n(s))^2 ds < \infty, \quad \text{a.s.} \tag{2.20}$$

Now, we rewrite (2.14) using (2.15) to obtain

$$\lim_{n\to 0}\int_0^1 (z_n^T(s)(x_{\tau_n}+m_n(s)))^2 ds = 0, \quad \text{a.s.} \tag{2.21}$$

Hence we conclude (by writing $x_{\tau_n} = (x_{\tau_n}+m_n)-m_n$, using (2.20), (2.21), and the Cauchy–Schwarz (C–S) inequality) that

$$\sup_{n\geq 0}\int_0^1 (z_n^T(s)x_{\tau_n})^2 ds < \infty, \quad \text{a.s.} \tag{2.22}$$

At this stage in the proof we have replaced the initial hypothesized contradiction (2.14) by the study of the consequences of (2.21) and (2.22), where the processes $\{x_{\tau_n+s}\ 0\leq s\leq 1\}, n\geq 0$, are generated by an A_t process with a.s. bounded trajectories. The next stage is to replace the integrals in (2.21) with integrals with past measurable integrands, and hence to approximate the basic controllability covariance (or Grammian) calculation, which will give a contradiction by virtue of the controllability of $[A_\infty, C]$.

2.5. A truncation to obtain uniform integrability. As will be seen below, the technique of the proof uses uniform integrability of the integral in (2.22). Since this is not guaranteed here, we resort to another truncation mechanism.

Let $\varepsilon > 0$ be such that $P(\widetilde{\Omega}) > 2\varepsilon$. Next let $\Omega_N \in \mathbf{F}_\infty^x$ be the set such that, by (2.22),

$$\sup_{n\geq 0}\int_0^1 (z_n^T(s)x_{\tau_n})^2 ds \leq N \quad \text{a.e. on } \Omega_N, \tag{2.23}$$

with $P(\Omega_N) > 1-\varepsilon$ for all sufficiently large N. Further, by (2.4) and (2.22), $P(\Gamma_N \cap \Omega_N) > 1-2\varepsilon$ for sufficiently large N, and hence $P(\widetilde{\Omega}\cap\Gamma_N\cap\Omega_N) > 0$. (Recall that $\Gamma_N = \{x_t = x_t^N, t\geq 0\}$.)

Define the $\mathbf{R}^n$ valued process $\{\alpha_n^N\}$ by

$$\alpha_n^N = \begin{cases} \alpha_n & \text{a.e. on } \Omega_N\cap\Gamma_N, \\ 0 & \text{a.e. on } \Omega_N^c\cup\Gamma_N^c \end{cases} \tag{2.24}$$

(with an associated limit α^N) and accordingly

$$z_n^N(s) = \phi_n^T(s,0)\alpha_n^N. \tag{2.25}$$

Then, by definition,

$$\sup_{n\geq 0}\int_0^1 (x_{\tau_n}^T z_n^N(s))^2 ds \leq N, \quad \text{a.s.} \tag{2.26}$$

Hence, with (2.26), the basic limiting relation (2.14) (or equivalently (2.21)) implies

$$\lim_{n\to\infty}\int_0^1 ((x_{\tau_n}+m_n(s))^T z_n^N(s))^2 ds = 0, \quad \text{a.s.} \tag{2.27}$$

Recall that z and hence z^N are uniformly bounded by e^N. Hence z^N is square integrable, and this enables us to define

$$\widehat{z}_n^N(s) = E[z_n^N(s)|\mathbf{F}_{\tau_n}^x], \quad s\in[0,1]. \tag{2.28}$$

2.6. The uniform consistency of $\widehat{z}^N$. Our objective now is to show that, in the formula (2.27), z^N can be replaced by $\widehat{z}^N$. Towards this end, we observe that the ODE for $z_n^N(s)$ (as defined in (2.25)) is

$$\frac{d}{ds}z_n^N(s) = A_{\tau_n+s}^T z_n^N(s), \quad z_n^N(0)=\alpha_n^N, \quad 0\leq s\leq 1. \tag{2.29}$$

In order to derive an ODE for $\widehat{z}_n^N$, note that

$$\begin{aligned}\|z_n(s+h)-z_n(s)\|/h &\leq \|\phi_n^T(s,0+h)-\phi_n^T(s,0)\|/h\\ &\leq 2\sup_{0\leq s\leq 1}\|A_{\tau_n+s}^T\phi_n(s,0)\|\leq 2Ne^N\end{aligned}$$

for any $n\geq 0, s\in[0,1]$, and small enough h.

Hence, by letting $h\to 0$, dominated convergence leads to

$$\begin{aligned}\frac{d}{ds}\widehat{z}_n^N(s) &= \lim_{h\to 0}E[z_n^N(s+h)-z_n^N(s)|\mathbf{F}_{\tau_n}^x]/h = E\left[\lim_{h\to 0}(z_n^N(s+h)-z_n^N(s))/h|\mathbf{F}_{\tau_n}^x\right]\\ &= E[A_{\tau_n+s}^T z_n^N(s)|\mathbf{F}_{\tau_n}^x], \qquad \widehat{z}_n^N(0)=\widehat{\alpha}_n^N \underline{\Delta} E[\alpha_n^N|\mathbf{F}_{\tau_n}^x].\end{aligned} \tag{2.30}$$

Let $e_n(s)=z_n^N(s)-\widehat{z}_n^N(s)$. Then, from (2.29) and (2.30),

$$\begin{aligned}\frac{d}{ds}e_n(s) &= A_{\tau_n+s}^T z_n^N(s) - E[A_{\tau_n+s}^T z_n^N(s)|\mathbf{F}_{\tau_n}^x]\\ &= A_{\tau_n+s}^T(z_n^N(s)-\widehat{z}_n^N(s)) + A_{\tau_n+s}^T\widehat{z}_n^N(s) - E[A_{\tau_n+s}^T z_n^N(s)|\mathbf{F}_{\tau_n}^x]\\ &= A_{\tau_n+s}^T e_n(s) + (A_{\tau_n+s}-A_{\tau_n})^T\widehat{z}_n^N(s) + A_{\tau_n}^T\widehat{z}_n^N(s) - E[A_{\tau_n+s}^T z_n^N(s)|\mathbf{F}_{\tau_n}^x]\\ &= A_{\tau_n+s}^T e_n(s) + (A_{\tau_n+s}-A_{\tau_n})^T\widehat{z}_n^N(s) + E[(A_{\tau_n}-A_{\tau_n+s})^T z_n^N(s)|\mathbf{F}_{\tau_n}^x],\end{aligned} \tag{2.31}$$

with the initial condition $e_n(0)=\alpha_n^N-\widehat{\alpha}_n^N$. (Note that in introducing $A_{\tau_n}^T\widehat{z}_n^N(s)$ into the conditional expectation, we have used the definition of $\widehat{z}^N$ and the measurability of A_{τ_n} w.r.t. $\mathbf{F}_{\tau_n}^x$.)

The solution of (2.31) is then

$$e_n(s) = \phi_n^T(s,0)(\alpha_n^N - \widehat{\alpha}_n^N) + \int_0^s \phi_n^T(r,s)[A_{\tau_n+r} - A_{\tau_n}]^T \widehat{z}_n^N(r)dr$$
$$+ \int_0^s \phi_n^T(r,s)E[(A_{\tau_n} - A_{\tau_n+r})^T z_n^N(r)|\mathbf{F}^x_{\tau_n}]dr. \tag{2.32}$$

Using the uniform bounds which hold on ϕ, z^N, and $\widehat{z}^N$, it can be deduced from (2.32) that

$$\sup_{0\le s\le 1}\|z_n^N(s) - \widehat{z}_n^N(s)\| = \sup_{0\le s\le 1}\|e_n(s)\| \le e^N\|\alpha_n^N - \widehat{\alpha}_n^N\| + e^{2N}\sup_{0\le s\le 1}\|A_{\tau_n+s} - A_{\tau_n}\|$$
$$+ e^{2N}\left(E\left[\sup_{0\le s\le 1}\|A_{\tau_n+s} - A_{\tau_n}\|^2|\mathbf{F}^x_{\tau_n}\right]\right)^{1/2}.$$

Let $u_n = \sup_{0\le s\le 1}\|A_{\tau_n+s} - A_{\tau_n}\|^2$ and note that, due to (2.2) and the definition of the stopped A_t process, $u_n \to 0$ a.s. as $n \to \infty$; hence the second term on the right-hand side (RHS) goes to zero as $n \to \infty$, a.s. Furthermore, since by (2.5) $0 \le u_n \le 2N$, the dominated convergence theorem implies that $Eu_n \to 0$, which makes the third term on the RHS decay to zero as $n \to \infty$. Finally, observing that $\widehat{\alpha}_n^N = E[\alpha_n^N|\mathbf{F}^x_{\tau_n}] \to E[\alpha^N|\mathbf{F}^x_\infty] = \alpha^N$ (as $\alpha^N \in \mathbf{F}^x_\infty$), one has

$$\limsup_{n\to\infty 0\le s\le 1}\|z_n^N(s) - \widehat{z}_n^N(s)\| = 0, \quad \text{a.s.} \tag{2.33}$$

Furthermore, using the same method with which (2.33) was obtained, we see that, with probability 1 (w.p.1)

$$\limsup_{n\to\infty 0\le s\le 1}\|\phi_n(s,0) - \exp A_{\tau_n}s\| = \limsup_{n\to\infty 0\le s\le 1}\sup_{\|y\|=1}\|\phi_n(s,0)y - [\exp A_{\tau_n}s]y\| = 0, \tag{2.34}$$

a fact used in what follows.

2.7. Main argument to obtain a contradiction. Returning to our main problem, it can be verified that the integrand in (2.27) may be rewritten as

$$\begin{aligned}[(x_{\tau_n} + m_n(s))^T z_n^N(s)]^2 &= [(x_{\tau_n} + m_n(s))^T \widehat{z}_n^N(s)]^2 + [(x_{\tau_n} + m_n(s))^T(\widehat{z}_n^N(s) - z_n^N(s))]^2\\ &- 2(\widehat{z}_n^N(s) - z_n^N(s))^T m_n(s)m_n^T(s)\widehat{z}_n^N(s)\\ &- 2(\widehat{z}_n^N(s) - z_n^N(s))^T m_n(s)x_{\tau_n}^T\widehat{z}_n^N(s)\\ &- 2(\widehat{z}_n^N(s) - z_n^N(s))^T x_{\tau_n}x_{\tau_n}^T\widehat{z}_n^N(s)\\ &- 2(\widehat{z}_n^N(s) - z_n^N(s))^T x_{\tau_n}m_n^T(s)\widehat{z}_n^N(s).\end{aligned} \tag{2.35}$$

Using (2.33), our next step is to show that the integrals over $[0,1]$ of the third and fourth terms on the RHS of (2.35) decay to zero as $n \to \infty$. Consider first the third term on the RHS of (2.35): Equipped with (2.18), (2.33), and the uniform bound on $\widehat{z}^N$, a C–S inequality leads to

$$\lim_{n\to\infty}\int_0^1 (\widehat{z}_n^N(s) - z_n^N(s))^T m_n(s)m_n^T(s)\widehat{z}_n^N(s)ds = 0, \quad \text{a.s.} \tag{2.36}$$

As for the fourth term on the RHS of (2.35), first note that (2.26), the fact that $x_{\tau_n} \in \mathbf{F}^x_{\tau_n}$, $n \geq 0$, and Jensen's inequality guarantee that $\sup_{n\geq 0} \int_0^1 (x_{\tau_n} \widehat{z}_n^N(s))^2 ds \leq N$, a.s. Hence, at this point, (2.18), (2.26), and a C–S inequality imply that

$$\lim_{n\to\infty} \int_0^1 (\widehat{z}_n^N(s) - z_n^N(s))^T m_n(s) x_{\tau_n}^T \widehat{z}_n^N(s) ds = 0, \quad \text{a.s.} \tag{2.37}$$

Let us consider the two last terms on the RHS of (2.35).

First, by the Burkholder–Davis–Gundy (B-D-G) inequality (e.g., Theorem IV-4.1, Revuz and Yor (1991)), together with the uniform bounds (2.16), (2.19), on ϕ_n and z_n (respectively), it follows that

$$\begin{aligned} E \sup_{s\in[0,1]} (m_n^T(s) z_n^N(s))^4 &\leq e^{4N} E \sup_{s\in[0,1]} \|m_n(s)\|^4 && \text{(due to (2.19))} \\ &\leq c_4 e^{4N} \left(Tr E \int_0^1 \phi_n(s,0) C C^T \phi_n^T(s,0) ds \right)^2 && \text{(due to the B-D-G inequality)} \\ &\leq c_4 e^{8N} (Tr C C^T)^2 && \text{(due to (2.16))} \end{aligned}$$

(where c_4 is the universal constant in the 4th moment, B-D-G inequality), and, by Theorem 2, section 3-VII of Shiryayev (1984) (for which use of the bound on the 4th moment above is required),

$$E \sup_{n\geq 0} \sup_{s\in[0,1]} (m_n^T(s) z_n^N(s))^2 \leq 9\sqrt{c_4} e^{4N} Tr C C^T.$$

This, together with (2.26), enables us to apply the dominated convergence theorem to (2.27). Recall that (2.35) is the integrand $((x_{\tau_n} + m_n(s))^T z_n^N(s))^2$ in (2.27), broken into the sum of six terms. After the elimination of the third and fourth terms (by (2.36) and (2.37), respectively) and the use of the dominated convergence theorem (justified above), (2.27) results in

$$\begin{aligned} 0 &= \lim_{n\to\infty} E \left[\int_0^1 \left((x_{\tau_n} + m_n(s))^T z_n^N(s) \right)^2 ds \right] \\ &\geq \liminf_{n\to\infty} E \left[\int_0^1 ((x_{\tau_n} + m_n(s))^T \widehat{z}_n^N(s))^2 ds \right] \\ &\quad - 2 \limsup_{n\to\infty} E \left[\int_0^1 \left(\widehat{z}_n^N(s) - z_n^N(s) \right)^T x_{\tau_n} \left(x_{\tau_n} + m_n(s) \right)^T \widehat{z}_n^N(s) ds \right]. \end{aligned} \tag{2.38}$$

On our way to obtaining a contradiction, the next step involves the elimination of the second term on the RHS of (2.38). First, note that the measurability of x_{τ_n} and $\widehat{z}_n^N$ w.r.t. $\mathbf{F}^x_{\tau_n}$, the definition of $\widehat{z}_n^N(s)$, and (2.26) together lead to

$$\begin{aligned} & E \left[\int_0^1 \left(\widehat{z}_n^N(s) - z_n^N(s) \right)^T x_{\tau_n} x_{\tau_n}^T \widehat{z}_n^N(s) ds \right] \\ &= E \left\{ E \left[\int_0^1 \left(\widehat{z}_n^N(s) - z_n^N(s) \right)^T x_{\tau_n} x_{\tau_n}^T \widehat{z}_n^N(s) ds | \mathbf{F}^x_{\tau_n} \right] \right\} \\ &= E \left\{ \int_0^1 E \left[\left(\widehat{z}_n^N(s) - z_n^N(s) \right)^T | \mathbf{F}^x_{\tau_n} \right] x_{\tau_n} x_{\tau_n}^T \widehat{z}_n^N(s) ds \right\} = 0 \quad \forall n \geq 0. \end{aligned} \tag{2.39}$$

To complete the elimination of the second term in the right-most expression in the chain (2.38), we claim that

$$\lim_{n\to\infty} E\left[\int_0^1 \left(\widehat{z}_n^N(s) - z_n^N(s)\right)^T x_{\tau_n} m_n^T(s)\widehat{z}_n^N(s)ds\right] = 0. \tag{2.40}$$

To see this, first note that, (i) since x_{τ_n} and $\widehat{z}_n^N(s)$ are $\mathbf{F}^x_{\tau_n}$-measurable and (ii) because $\{m_n(s), \mathbf{F}^x_{\tau_n+s}, s \in [0,1]\}$ is a zero mean martingale satisfying $E[m_n(s)|\mathbf{F}^x_{\tau_n}] = 0$, it follows that

$$\begin{aligned}
&E\left[\int_0^1 \left(\widehat{z}_n^N(s) - z_n^N(s)\right)^T x_{\tau_n} m_n^T(s)\widehat{z}_n^N(s)ds\right] \\
&= \int_0^1 E\left[\left(\widehat{z}_n^N(s) - z_n^N(s)\right)^T x_{\tau_n} m_n^T(s)\widehat{z}_n^N(s)\right] ds \\
&= \int_0^1 E\left\{\left(\widehat{z}_n^N(s)\right)^T x_{\tau_n} E\left[m_n(s)|\mathbf{F}^x_{\tau_n}\right]\widehat{z}_n^N(s)\right\} ds - \int_0^1 E\left[\left(z_n^N(s)\right)^T x_{\tau_n} m_n^T(s)\widehat{z}_n(s)\right] ds \\
&= \int_0^1 E\left[x_{\tau_n}^T z_n^N(s) m_n^T(s)\widehat{z}_n(s)\right] ds.
\end{aligned}$$

(A Fubini-type argument justifies the interchange between integration (over $[0,1]$) and expectation.)

Hence the claim (2.40) is equivalent to the following statement, for which a proof is given in the appendix.

PROPOSITION 2.2.

$$\lim_{n\to\infty} \int_0^1 E\left[x_{\tau_n}^T z_n^N(s) m_n^T(s)\widehat{z}_n^N(s)\right] ds = 0. \tag{2.41}$$

We now proceed assuming the validity of (2.41). Having shown that the second term on the RHS of (2.38) is equal to zero, it is clear that a contradiction to (2.38) will be established if it is shown that

$$\liminf_{n\to\infty} E\left[\int_0^1 \left((x_{\tau_n} + m_n(s))^T \widehat{z}_n^N(s)\right)^2 ds\right] > 0, \tag{2.42}$$

which in turn is implied by

$$\liminf_{n\to\infty} E\left[\int_0^1 \left((x_{\tau_n} + m_n(s))^T \widehat{z}_n^N(s)\right)^2 ds|\mathbf{F}^x_{\tau_n}\right] > 0, \tag{2.43}$$

with positive probability.

2.8. Approximations to the transition matrix. It remains to prove (2.43). To simplify the presentation, we first show that $m_n^T(s)\widehat{z}_n^N(s)$ may be replaced (in the formula (2.43)) by a conditionally Gaussian process. Let

$$\widetilde{m}_n(s) = \int_0^s [\exp -A_{\tau_n} r] C dw_{\tau_n+r}, \tag{2.44}$$

$$\widetilde{z}_n^N(s) = [\exp A_{\tau_n} s]^T \widehat{\alpha}_n^N. \tag{2.45}$$

Then, using the convergence of $\{A_{\tau_n}\}$ to A_∞, we follow an argument closely analogous to that which led to (2.34) (the difference being that the corresponding second term

on the RHS of (2.31), (2.32), etc., is zero in the current case) to obtain

$$\limsup_{n\to\infty 0\le s\le 1} \|\widehat{z}_n^N(s) - \widetilde{z}_n^N\| = 0, \quad \text{a.s.} \tag{2.46}$$

Hence, by dominated convergence,

$$\lim_{n\to\infty} E \sup_{0\le s\le 1} \|\widehat{z}_n^N(s) - \widetilde{z}_n^N(s)\|^2 = 0.$$

Furthermore, using calculations similar to (2.17),

$$\begin{aligned}
&\lim_{n\to\infty} E \sup_{0\le s\le 1} \|m_n(s) - \widetilde{m}_n(s)\|^2 \\
&\quad \le \lim_{n\to\infty} 4TrE \int_0^1 [\phi_n(0,r) - \exp(-A_{\tau_n} r)]CC^T[\phi_n(0,r) - \exp(-A_{\tau_n} r)]^T dr = 0,
\end{aligned} \tag{2.47}$$

which is obtained due to the uniform convergence of the integrand (see (2.34)) and dominated convergence.

Therefore, since $\|\widetilde{z}_n^N(s)\|^2 \le e^{2N}$ and (see (2.17))

$$E\left[\sup_{0\le s\le 1} \|m_n(s)\|^2 |\mathbf{F}^x_{\tau_n}\right] \le 4e^{2N} Tr\ CC^T,$$

one has

$$\begin{aligned}
&E \sup_{0\le s\le 1} |m_n^T(s)\widehat{z}_n^N(s) - \widetilde{m}_n^T(s)\widetilde{z}_n^N(s)|^2 \\
&\le 2E \sup_{0\le s\le 1} \|\widetilde{z}_n^N(s)\|^2 \sup_{0\le s\le 1} \|m_n(s) - \widetilde{m}_n(s)\|^2 \\
&\quad + 2E \sup_{0\le s\le 1} \|\widetilde{z}_n^N(s) - \widehat{z}_n^N(s)\|^2 E\left[\sup_{0\le s\le 1} \|m_n(s)\|^2 |\mathbf{F}^x_{\tau_n}\right] \to 0, \quad \text{a.s., } n\to\infty.
\end{aligned} \tag{2.48}$$

Proof of (2.43). Equipped with (2.48), it is obvious that proving (2.43) is equivalent to showing that

$$\liminf_{n\to\infty} E\left[\int_0^1 (x_{\tau_n}^T \widehat{z}_n^N(s) + \widetilde{m}_n^T(s)\widetilde{z}_n^N(s))^2 ds |\mathbf{F}^x_{\tau_n}\right] > 0, \quad \text{with positive probability.} \tag{2.49}$$

Towards this end we use standard (Gaussian) calculations. Let

$$\gamma_{n+1} = \int_0^1 (x_{\tau_n}^T \widehat{z}_n^N(s) + \widetilde{m}_n(s)\widetilde{z}_n^N(s))^2 \wedge 1 ds. \tag{2.50}$$

To establish (2.49), it suffices to show that

$$\liminf_{n\to\infty} E[\gamma_{n+1} |\mathbf{F}^x_{\tau_n}] > 0, \quad \text{a.e. on } \widetilde{\Omega}_N, \tag{2.51}$$

where $\widetilde{\Omega}_N \underline{\Delta} \widetilde{\Omega} \cap \Gamma_N \cap \Omega_N$, and, by the choice of N, $P(\widetilde{\Omega}_N) > 0$.

To simplify notation, we define

$$\mu_n(s) = x^T_{\tau_n} \widehat{z}^N_n(s), \tag{2.52}$$

$$\zeta_n(s) = \widetilde{m}^T_n(s)\widetilde{z}^N_n(s) = (\widehat{\alpha}^N_n)^T \int_0^s [\exp A_{\tau_n}(s-r)]C dw_{\tau_n+r}. \tag{2.53}$$

Note that $\mu_n(s) \in \mathbf{F}^x_{\tau_n}$, and that $\zeta_n(s)$, for any fixed s, is a conditionally centered Gaussian random variable (w.r.t. $\mathbf{F}^x_{\tau_n}$).

Define the conditional variance by

$$\begin{aligned}\sigma^2_n(s) &= E[\zeta^2_n(s)|\mathbf{F}^x_{\tau_n}] \\ &= (\widehat{\alpha}^N_n)^T \int_0^s [\exp A_{\tau_n}(s-r)]CC^T[\exp A_{\tau_n}(s-r)^T]dr\widehat{\alpha}^N_n. \end{aligned} \tag{2.54}$$

With these definitions and observations we turn to the calculation of $E[\gamma_{n+1}|\mathbf{F}^x_{\tau_n}]$:

$$\begin{aligned} E[\gamma_{n+1}|\mathbf{F}^x_{\tau_n}] &= E\left\{\int_0^1 (\zeta_n(s)+\mu_n(s))^2 \wedge 1 ds|\mathbf{F}^x_{\tau_n}\right\} \\ &= \int_0^1 E[(\zeta_n(s)+\mu_n(s))^2 \wedge 1|\mathbf{F}^x_{\tau_n}]ds \\ &\geq \int_0^1 P[(\zeta_n(s)+\mu_n(s))^2 \geq 1|\mathbf{F}^x_{\tau_n}]ds \\ &\geq \frac{1}{2}\inf_{\frac{1}{2}\leq s\leq 1}\left\{P\left(\zeta_n(s)\geq 1-\mu_n(s)|\mathbf{F}^x_{\tau_n}\right)+P\left(\zeta_n(s)\leq -1-\mu_n(s)|\mathbf{F}^x_{\tau_n}\right)\right\} \\ &\geq \inf_{\frac{1}{2}\leq s\leq 1} P(\zeta_n(s)\geq 1|\mathbf{F}^x_{\tau_n}) = \inf_{\frac{1}{2}\leq s\leq 1}\frac{1}{\sqrt{2\pi}}\int_{1/\sigma_n(s)}^{\infty}\exp\left(\frac{-u^2}{2}\right)du \\ &= \frac{1}{\sqrt{2\pi}}\int_{1/\sigma_n(\frac{1}{2})}^{\infty}\exp\left(\frac{-u^2}{2}\right)du \\ &\geq \frac{1}{\sqrt{2\pi}}\frac{\sigma_n(\frac{1}{2})}{1+\sigma^2_n(\frac{1}{2})}\exp\left(\frac{-1}{2\sigma^2_n(\frac{1}{2})}\right), \end{aligned} \tag{2.55}$$

where the last equality is due to the fact that $\sigma_n(\cdot)$ is nondecreasing, and the last inequality is a simple lower bound for the Gaussian integral.

Recall that

$$\widehat{\alpha}^N_n = E[\alpha^N_n|\mathbf{F}^x_{\tau_n}] \to E[\alpha^N|\mathbf{F}^x_{\infty}] = \alpha^N, \quad \text{a.s. as } n\to\infty.$$

Furthermore, the integrand in (2.54) converges uniformly. This leads to

$$\sigma^2_{\infty} \underline{\Delta} \lim_{n\to\infty}\sigma^2_n\left(\frac{1}{2}\right) = (\alpha^N)^T \int_0^{1/2}\left[\exp A_{\infty}\left(\frac{1}{2}-r\right)\right]CC^T\left[\exp A_{\infty}\left(\frac{1}{2}-r\right)\right]^T dr\alpha^N, \tag{2.56}$$

and, with the continuity of the last expression on the RHS of (2.55) (as a function of σ), one has

$$\liminf_{n\to\infty} E[\gamma_{n+1}|\mathbf{F}^x_{\tau_n}] \geq \frac{1}{\sqrt{2\pi}}\frac{\sigma_{\infty}}{1+\sigma^2_{\infty}}\exp\left(\frac{-1}{2\sigma^2_{\infty}}\right), \quad \text{a.s.} \tag{2.57}$$

Finally, recall that, by definition, $\|\alpha^N\| = 1$ a.e. on $\widetilde{\Omega}_N$. Hence, with the fact that the pair $[A_\infty, C]$ is controllable (a.s.), it follows that

$$\sigma_\infty^2 > 0 \quad \text{a.e. on } \widetilde{\Omega}_N, \tag{2.58}$$

which immediately leads to (2.51). (Note that for the truncated process x^N and its limiting dynamic matrix A_∞^N, the pair $[A_\infty^N, C]$ is a.e. controllable on Γ_N, while $[A_\infty^N, C]$ might be uncontrollable on Γ_N^C, for which $A_\infty^N = A_{T_N}, T_N < \infty$.)

Since N was chosen such that $P(\widetilde{\Omega}_N) > 0$, the required contradiction is finally established. □

3. Examples and conclusion. We conclude the paper with a discussion of potential applications of the main result. Let $\lambda_m(t) = \lambda_m\{\int_0^t x_r x_r^T dr\}, m = \min, \max$. Then it follows from (2.3) that at one extreme $\lambda_{min}(t) = O(t)$, a.s. On the other hand, it is known that with unstable dynamics $\lambda_{max}(t) = O(te^{\beta t})$, a.s., for some $\beta > 0$, as $t \to \infty$. It turns out, then, that in the context of parameter estimation, an input signal $\{x_t\}$ (satisfying the above growth rates for its $\lambda_m(t)$, m = min, max), lacks adequate excitation for a recursive least squares (RLS)-type algorithm, as it fails to satisfy the condition (from Lai and Wei (1982)) that $\log \lambda_{max}(t)/\lambda_{min}(t) \to 0$ as $t \to \infty$, a.s. (In the absence of the convergence of this ratio to zero, Lai and Wei provide a counterexample to consistency.) On the other hand, in a Bayesian setting, (2.3) is sufficient to ensure consistency, where only $\lim_{t\to\infty} \lambda_{min}(t) = \infty$ is required.

Example 1. As an application of the main result of the paper, consider the system (2.1), where $\{A_t\}$ is a random, time-varying, convergent matrix process, i.e., a process satisfying (2.2). In this case, the consistency of the Bayesian estimate $E[A_t|\mathbf{F}_t^x]$ is ensured by the condition $\lambda_{min}(t) = O(t)$, a.s., a property implied by (2.3) (see Levanony (2001)). For the system described by (2.1), one then has $E[A_t|\mathbf{F}_t^x] \to A_\infty$, a.s. Furthermore, suppose that instead of the Bayesian estimate $E[A_t|\mathbf{F}_t^x]$ one uses a least squares (LS) estimate, denoted here by $\widehat{A}_t$; and finally, suppose that the system (2.1) can be parameterized by a *deterministic* parameter, say $\theta \in \mathbf{R}^p$. Then, by applying the idea of Bayesian embedding (Kumar (1989); see also the discussion in Caines (1988, pp. 294–295)), we may identify $\widehat{A}_t$ with $E[A_t|\mathbf{F}_t^x]$ for almost all $\theta \in \mathbf{R}^p$, and so it follows that under (2.3),

$$\widehat{A}_t \to A_\infty, \quad \text{a.s., for almost all} \quad \theta \in \mathbf{R}^p. \tag{3.1}$$

Hence, we conclude that, subject to the a.s. convergence condition (2.2), both the Bayesian and the LS consistency properties hold.

Example 2. As a control-related example of the above, consider an adaptive linear quadratic Gaussian (LQG) control problem of the form described in Duncan, Guo, and Pasik-Duncan (1999), where the system

$$dx_t = Ax_t dt + Bu_t + Cdw_t, \quad t \geq 0, \tag{3.2}$$

is to be controlled by an input function $u = \{u_t\}$, which is chosen so as to minimize a (standard) long-term averaged linear quadratic (LQ) cost. It is well known that the optimal control u^0 takes the form $u_t^0 = -K(A, B)x_t$. With *unknown* system matrices (A, B), one can use a CE (certainty equivalence) approach by which $\theta = (A, B)$ is estimated by $\widehat{\theta}_t = \widehat{\theta}_t(x_s, 0 \leq s \leq t)$, generated by some specified parameter estimation algorithm. The resulting control input would then assume the form $u_t = -K(\widehat{\theta}_t)x_t$

and, under this scheme, (3.1) is rewritten as (2.1) with $A_t = A - BK(\widehat{\theta}_t)$. (See Kumar (1983) for the discrete time, finite parameter set case.)

Suppose that the chosen parameter estimation algorithm is self-convergent (not necessarily consistent) (Duncan, Guo, and Pasik-Duncan (1999)); that is, for almost all parameters $\theta \in \mathbf{R}^p$ there exists an a.s. finite random vector $\widehat{\theta}_\infty$ such that $\widehat{\theta}_t \to \widehat{\theta}_\infty$ (w.p.1). Then, the continuity of $K(\cdot)$ clearly leads to $A_t \to A_\infty = A - BK(\widehat{\theta}_\infty)$, which is the basic assumption (2.2) above. Hence, given (2.2), the main result implies the persistent excitation (PE) property (2.3).

Now consider a separate procedure to estimate the time varying, *closed-loop* system matrix A_t by a Bayesian or an LS estimate, depending, respectively, on whether $\theta = (A, B)$ is random and Gaussian, or deterministic. By Example 1 above, both procedures yield a strongly consistent limit, where, in the LS case, this property is obtained in terms of Bayesian embedding.

Although the estimation of $\theta = (A, B)$ and the estimation of the closed-loop matrix $A_t = A - BK(\widehat{\theta}_t)$ were conceptually taken as two separate procedures, the two can be combined within one estimation scheme, as is easily seen in the Bayesian setting. Suppose that $\theta = (A, B)$ is Gaussian and independent of x_0, $\{w_t\}$. Then the Bayesian estimate of A_t takes the form

$$E[A_t|\mathbf{F}_t^x] = E[A - BK(\widehat{\theta}_t)|\mathbf{F}_t^x] = E[A|\mathbf{F}_t^x] - E[B|\mathbf{F}_t^x]K(\widehat{\theta}_t), \tag{3.3}$$

where the last equality follows from the fact that $K(\widehat{\theta}_t) \in \mathbf{F}_t^x$. Now take $\widehat{\theta}_t = E[\theta|\mathbf{F}_t^x]$; then the closed-loop estimate (3.3) is immediately obtained without any further calculation. In the deterministic (A, B) setting, the same applies for the LS estimate by virtue of Bayesian embedding (BE). Further, note that the Bayesian estimate of θ converges to a finite limit and, due to BE, so does the LS estimate (the self-convergence property, leading to (2.2)).

As far as adaptive control is concerned, it is interesting to note that the strong consistency of the *closed-loop* Bayesian (equivalently, LS) estimates implies that the corresponding generated limits $\widehat{\theta}_\infty = \theta' = (A', B')$ are characterized by the property that they form closed-loop dynamics which are *indistinguishable* from the actual (closed-loop) dynamics, that is to say, $A' - B'K(\theta') = A - BK(\theta')$ (Caines and Levanony (1993); Levanony and Caines (1996)).

This property, obtained from (2.3), may form the foundation of modifications to the (original) adaptive schemes, which would lead to the desired property of long-run, optimal performance (Caines and Levanony (1993)).

Appendix: Proof of Proposition 2.2. Recall (2.53), i.e., that $\zeta_n(s) = \widetilde{m}_n^T(s)\widetilde{z}_n^N(s)$ (where $\widetilde{m}$ and $\widetilde{z}$ are defined in (2.44) and (2.45), respectively), and let

$$\zeta(s) = (\alpha^N)^T \int_0^s \big[\exp A_\infty(s-r)\big] C d\beta_r, \tag{A.1}$$

where $A_\infty \in \mathbf{F}_\infty^x (= \sigma\{\bigcup_{n\geq 0} \mathbf{F}_{\tau_n}^x\})$ is the a.s. limit of $\{A_{\tau_n} \in \mathbf{F}_{\tau_n}^x\}$ and β is a (vector-valued) Brownian motion, *independent of* $\mathbf{F}_\infty^x$. (It is assumed that the probability space has been properly extended to support this Brownian motion.) Note that this definition makes $\zeta(s)$ conditionally Gaussian w.r.t. $\mathbf{F}_\infty^x$ (as is ζ_n, w.r.t. $\mathbf{F}_{\tau_n}^x$).

We now show that ζ induces a unique measure on $\mathbf{R}$ such that

$$\zeta_n(s) \longrightarrow \zeta(s) \quad \forall s \in [0, 1] \tag{A.2}$$

in distribution. For (A.2) to hold, it suffices that (i) all moments of ζ_n converge to those of ζ and that (ii) the set of moments form a separating function class, a fact which guarantees uniqueness; see, e.g., Billingsley (1979) and Breiman (1968).

The $2k$th moment of $\zeta_n(s)$ is calculated by utilizing its conditionally Gaussian distribution (note that odd moments of ζ_n and ζ are identically zero):

(A.3)
$$E|\zeta_n(s)|^{2k} = E\big\{E\big[|\zeta_n(s)|^{2k}|\mathbf{F}^x_{\tau_n}\big]\big\} = 1\cdot 3\cdot 5\cdots(2k-1)E\left\{E\big[|\zeta_n(s)|^2|\mathbf{F}^x_{\tau_n}\big]^k\right\},$$

where, w.p.1,

$$E\big[|\zeta_n(s)|^2|\mathbf{F}^x_{\tau_n}\big] = E\big[|\widetilde{m}^T_n(s)\widetilde{z}^N_n(s)|^2|\mathbf{F}^x_{\tau_n}\big]$$

$$= (\widehat{\alpha}^N_n)^T\int_0^s \big[\exp A_{\tau_n}(s-r)\big]CC^T\big[\exp A_{\tau_n}(s-r)\big]^T dr\widehat{\alpha}^N_n \le ||C||^2e^{2N}. \tag{A.4}$$

A similar calculation leads to

$$E|\zeta(s)|^{2k} = 1\cdot 3\cdot 5\cdots(2k-1)E\left\{E\big[|\zeta(s)|^2|\mathbf{F}^x_\infty\big]^k\right\}, \tag{A.5}$$

where

$$E\big[|\zeta(s)|^2|\mathbf{F}^x_\infty\big] = (\alpha^N)^T\int_0^s \big[\exp A_\infty(s-r)\big]CC^T\big[\exp A_\infty(s-r)\big]^T dr\alpha^N. \tag{A.6}$$

Now as $\widehat{\alpha}^N_n \to \alpha^N$, $A_{\tau_n}\to A_\infty$ (a.s.) with α^N and A_∞ being $\mathbf{F}^x_\infty$-measurable, it follows (with the aid of dominated convergence) that

$$E|\zeta_n(s)|^{2k} \to E|\zeta(s)|^{2k} \tag{A.7}$$

as $n\to\infty$, $k=1,2,\ldots$, $s\in[0,1]$, which in turn implies (A.2) (Breiman (1968), Theorem 8.48). Uniqueness follows from the fact that

$$\limsup_{k\to\infty}\frac{(E|\zeta(s)|^{2k})^{1/2k}}{2k} \le \limsup_{k\to\infty}\frac{(1\cdot 3\cdot 5\cdots(2k-1))^{1/2k}}{2k}||C||^2e^N \le ||C||^2e^N < \infty,$$

which makes the set of moments a separating function class (Breiman (1968), Proposition 8.49). (Another uniqueness condition, namely that $\sum_{k=1}^\infty E|\zeta(s)|^{2k}r^{2k}/(2k)! < \infty$ for some $r>0$, is satisfied here with $0<r<(||C||^2e^N)^{-1}$ (Billingsley (1979, Theorem 30.1)).)

Now, combining (2.27) with (2.33), (2.46), and (2.47) leads to

$$\lim_{n\to\infty}\int_0^1 (x^T_{\tau_n}z^N_n(s)+\zeta_n(s))^2ds = 0, \quad \text{a.s.}, \tag{A.8}$$

and integrability and dominated convergence imply that

$$\lim_{n\to\infty}\int_0^1 E(x^T_{\tau_n}z^N_n(s)+\zeta_n(s))^2ds = 0. \tag{A.9}$$

Define $\eta_n(s)\underline{\Delta}x^T_{\tau_n}z^N_n(s)$, fix $s\in[0,1]$, and write

$$\delta_n(s) = \zeta_n(s)+\eta_n(s). \tag{A.10}$$

Then by (A.9) one obviously has

$$\delta_n(s) \longrightarrow 0 \tag{A.11}$$

in L_2, a.e. on [0,1]. We now claim the following.

LEMMA A.1.

$$(\zeta_n(s), \delta_n(s)) \longrightarrow (\zeta(s), 0), \tag{A.12}$$

in distribution, a.e. on [0,1].

Proof. Based on (A.2) and (A.11),

$$\begin{aligned} P(\zeta_n(s) \le a, \delta_n(s) \le b) &= P(\zeta_n(s) \le a) - P(\zeta_n(s) \le a, \delta_n(s) > b) \\ &\to \begin{cases} P(\zeta(s) \le a), & b \ge 0, \\ 0, & b < 0. \end{cases} \end{aligned}$$

This is due to the fact that

$$P(\zeta_n(s) \le a, \delta_n(s) > b) \le \begin{cases} P(\delta_n(s) > b) \to 0, & b \ge 0, \\ P(\zeta_n(s) \le a) \to P(\zeta(s) \le a), & b < 0, \end{cases}$$

while on the other hand,

$$P(\zeta_n(s) \le a, \delta_n(s) > b) \ge \begin{cases} 0, & b \ge 0, \\ P(\zeta_n(s) \le a) - P(\delta_n(s) \le b) \to P(\zeta(s) \le a), & b < 0. \end{cases}$$

Thus the claim is proved. □

It follows from (A.10) and (A.12) that there exists a random process $\{\eta(s), s \in [0,1]\}$, defined on the *underlying probability space* $(\Omega, \mathbf{F}, P)$, such that

$$(\zeta_n(s), \eta_n(s)) \longrightarrow (\zeta(s), \eta(s)), \tag{A.13}$$

in distribution, a.e. on [0, 1], where $\zeta(s)$ has been previously constructed as a random variable on the underlying probability space $(\Omega, \mathbf{F}, P)$, and, by (A.10) and (A.12), the aforementioned limit $\eta(s)$ is also defined on $(\Omega, \mathbf{F}, P)$. Furthermore, recall that $\eta_n(s) = x_{\tau_n}^T z_n^N(s) = x_{\tau_n}^T \phi_n(s,0)\alpha_n^N$, where $\alpha_n^N \in \mathbf{F}_\infty^x$, $x_{\tau_n} \in \mathbf{F}_{\tau_n}^x \subset \mathbf{F}_\infty^x$, and $\phi_n \in \mathbf{F}_{\tau_n+1}^x \subset \mathbf{F}_\infty^x$. Next recall that, by (2.34),

$$z_n^N(s) \longrightarrow z^N(s) = [\exp A_\infty^T s]\alpha^N, \tag{A.14}$$

uniformly over $[0,1]$, a.s. Since (i) $z^N(s)$ is $\mathbf{F}_\infty^x$-measurable (by definition) and (ii) $\{x_{\tau_n}\} \in \mathbf{F}_\infty^x$, it follows from (A.14) and Skorohod's theorem (Billigsley (1979, Theorem 25.6)) that $\{\eta(s), s \in [0,1]\}$, being the limit in distribution of $\eta_n(s) = x_{\tau_n}^T z_n^N(s) \in \mathbf{F}_\infty^x$, $s \in [0,1]$, is also $\mathbf{F}_\infty^x$-measurable (recall that η is defined on $(\Omega, \mathbf{F}, P)$).

In light of the definition of ζ (see (A.1)), this fact therefore implies that η and ζ are conditionally independent w.r.t. $\mathbf{F}_\infty^x$; i.e., the limit joint distribution of $\eta_n(s)$ and $\zeta_n(s)$ is that of two $\mathbf{F}_\infty^x$-conditionally independent random processes. This, together with the fact that ζ is conditionally centered, yields

$$\begin{aligned} \lim_{n\to\infty} E\big[x_{\tau_n}^T z_n^N(s) \widetilde{m}_n^T(s) \widetilde{z}_n(s)\big] &= \lim_{n\to\infty} E\big[\eta_n(s)\zeta_n(s)\big] = E\big[\eta(s)\zeta(s)\big] \\ &= E\big\{\eta(s) E\big[\zeta(s) | \mathbf{F}_\infty^x\big]\big\} = 0, \end{aligned} \tag{A.15}$$

a.e. on [0, 1]. Hence, utilizing dominated convergence yields

$$\lim_{n\to\infty}\int_0^1 E\big[x_{\tau_n}^T z_n^N(s)\widetilde{m}_n^T(s)\widetilde{z}_n(s)\big]\,ds = 0,$$

and, with (2.26) and (2.48), equation (2.41) is finally obtained. □

Remarks. 1. Recall that the construction of the limit measure for ζ_n (namely, that of ζ) has been carried out by the use of a Brownian motion β, assumed to be *independent* of $\mathbf{F}_\infty^x$. While the uniqueness of the resulting limit law has already been proved, it is worth noting that this independence, which plays a key role in the proof (see (A.15)), is crucial. This is due to the fact that if β were not independent of $\mathbf{F}_\infty^x$, then (A.5) and (A.6) would no longer hold, thus ruling out the validity of (A.7) and hence that of the claim (A.2). By introducing this particular Brownian motion β, we have (implicitly) constructed the conditional orthogonality between ζ and η (as part of their joint distribution), which, in turn, enabled the calculation made in (A.15).

2. We note the following additional contradiction, which was generated in the proof of the contradiction to (2.27): while (A.10) and (A.12) imply that $E[\eta(s)\zeta(s)] = -E[\zeta^2(s)] < 0$, this is not the case, since η and ζ were found to be mutually orthogonal.

Acknowledgments. The authors gratefully acknowledge the constructive comments of Ofer Zeitouni and those of the anonymous reviewers.

REFERENCES

P. Billingsley (1979), *Probability and Measure*, John Wiley, New York.
L. Breiman (1968), *Probability*, Addison–Wesley, Reading, MA.
P. E. Caines (1988), *Linear Stochastic Systems*, John Wiley, New York.
P. E. Caines (1992), *Continuous-time stochastic adaptive control: Non-explosion, ε-consistency and stability*, Systems Control Lett., 19, pp. 169–176.
P. E. Caines and D. Levanony (1993), *Performance monitored continuous-time LQ stochastic adaptive control*, in Proceedings of the 32nd IEEE Conference on Decision and Control, pp. 3539–3543.
T. E. Duncan and B. Pasik-Duncan (1986), *A parameter estimate associated with the adaptive control of stochastic systems*, in Analysis and Optimization of Systems, Lectures Notes in Control and Inform. Sci., 83, Springer-Verlag, New York, pp. 508–514.
T. E. Duncan, L. Guo, and B. Pasik-Duncan (1999), *Adaptive continuous-time linear quadratic Gaussian control*, IEEE Trans. Automat. Control, 44, pp. 1653–1662.
P. R. Kumar (1983), *Optimal adaptive control of linear-quadratic-Gaussian systems*, SIAM J. Control Optim., 21, pp. 163–178.
P. R. Kumar (1989), *Convergence of adaptive control schemes using least-squares estimates*, in Proceedings of the 28th IEEE Conference on Decision and Control, pp. 727–731.
T. L. Lai and C. Z. Wei (1982), *Least squares estimation in stochastic regression models with applications to identification and control of dynamical systems*, Ann. Statist., 10, pp. 154–166.
D. Levanony and P. E. Caines (1994), *Lagrangian stochastic LQ adaptation*, in Proceedings of the SIAM Annual Meeting, San Diego, CA, p. A4.
D. Levanony and P. E. Caines (1996), *Lagrangian Stochastic Adaptation*, Research Report, Dept. Electrical Engineering, McGill University, Montreal, QC, Canada.
D. Levanony (2001), *On the Consistent Filtering of Convergent Semimartingales*, preprint.
J. B. Moore (1987), *A universality advantage of stochastic exitation signals for adaptive control*, Systems Control Lett., 9, pp. 55–58.
D. Revuz and M. Yor (1991), *Continuous Martingales and Brownian Motion*, Springer-Verlag, New York.
A. N. Shiryayev (1984), *Probability*, Springer-Verlag, New York.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 3, pp. 898–924

EXISTENCE AND LIMITING BEHAVIOR OF A NON–INTERIOR-POINT TRAJECTORY FOR NONLINEAR COMPLEMENTARITY PROBLEMS WITHOUT STRICT FEASIBILITY CONDITION*

YUN-BIN ZHAO† AND DUAN LI‡

Abstract. For P_0-complementarity problems, most existing non–interior-point path-following methods require the existence of a strictly feasible point. (For a P_*-complementarity problem, the existence of a strictly feasible point is equivalent to the nonemptyness and the boundedness of the solution set.) In this paper, we propose a new homotopy formulation for complementarity problems by which a new non–interior-point continuation trajectory is generated. The existence and the boundedness of this non–interior-point trajectory for P_0-complementarity problems are proved under a very mild condition that is weaker than most conditions used in the literature. One prominent feature of this condition is that it may hold even when the often-assumed strict feasibility condition fails to hold. In particular, for a P_*-problem it turns out that the new non–interior-point trajectory exists and is bounded if and only if the problem has a solution. We also study the convergence of this trajectory and characterize its limiting point as the parameter approaches zero.

Key words. complementarity problems, non–interior-point methods, homotopy continuation trajectories, P_0-functions, P_*-functions

AMS subject classifications. 90C30, 90C33, 65K10

PII. S0363012900372477

1. Introduction.

The standard complementarity problem (CP) is to find a pair $(x, y) \in R^n \times R^n$ such that

$$y = f(x), \quad (x, y) \geq 0, \quad \text{and } x^T y = 0,$$

where $f : R^n \to R^n$ is a continuous function. This problem has many applications in optimization, economics, and engineering. See, for example, Cottle, Pang, and Stone [8], Harker and Pang [14], Heemels, Schumacher, and Weiland [15], van der Schaft and Schumacher [37], and Lötstedt [23].

The first non–interior-point method for the CP was proposed by Chen and Harker [5], and was based on the use of a Chen–Harker–Kanzow–Smale smooth function. Due to the impressive numerical performance of the algorithm, as well as its ideal convenience for application to those CPs in which interiority restriction on the iterates is quite severe, there is a growing interest in non–interior-point methods for the CP, which have yielded many fruitful results; see, e.g., Kanzow [18], Burke and Xu [1, 2, 3, 4], Xu [35], Xu and Burke [36], Chen and Chen [6], Hotta and Yoshise [16], Hotta, Inaba, and Yoshise [17], Qi and Sun [26], and Tseng [32]. In the setting of P_0-CPs, a common feature of the above-mentioned non–interior-point methods is the assumption of the strict feasibility condition (or the nonemptyness and the boundedness conditions

*Received by the editors May 16, 2000; accepted for publication (in revised form) April 10, 2001; published electronically October 31, 2001. This work was partially supported by grant CUHK4392/99E, research Grants Council, Hong Kong.

http://www.siam.org/journals/sicon/40-3/37247.html

†Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, NT, Hong Kong and Institute of Applied Mathematics, Chinese Academy of Sciences, Beijing, 100080, China (ybzhao@se.cuhk.edu.hk).

‡Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, NT, Hong Kong (dli@se.cuhk.edu.hk).

on the solution set) and a properness condition. For instance, Hotta and Yoshise [16] utilized the following condition.

CONDITION 1.1. (i) *f is a P_0-function, i.e., for any distinct vectors x, y in R^n*

$$\max_{x_i \neq y_i} (x_i - y_i)(f_i(x) - f_i(y)) \geq 0.$$

(ii) *There exists a strictly feasible point (x^0, y^0), i.e., $x^0 > 0$ and $y^0 = f(x^0) > 0$.*

(iii) *The set*

$$U^{-1}(D) = \{(u, x, y) \in R^n_+ \times R^{2n} : U(u, x, y) \in D\}$$

is bounded for every compact subset D of $R^n_+ \times V(R^n_{++} \times R^{2n})$, where $V : R^n_+ \times R^{2n} \to R^n$ and $U : R^n_+ \times R^{2n} \to R^n_+ \times R^{2n}$ are given by

$$V(u, x, y) = x + y - \sqrt{(x - y)^2 + 4u}$$

and

$$U(u, x, y) = \begin{pmatrix} u \\ x + y - \sqrt{(x - y)^2 + 4u} \\ y - f(x) \end{pmatrix} = \begin{pmatrix} u \\ V(u, x, y) \\ y - f(x) \end{pmatrix}, \tag{1.1}$$

respectively. All the above algebraic operations are performed componentwise.

The following standard condition was widely used in interior-point methods and non–interior-point methods. See, for example, [2, 3, 6, 16, 17, 19, 20, 21, 26, 38].

CONDITION 1.2. (i) *f is monotone, i.e., $(x - y)^T(f(x) - f(y)) \geq 0$ for any $(x, y) \in R^{2n}$.*

(ii) *There exists a strictly feasible point (x^0, y^0), i.e., $x^0 > 0$ and $y^0 = f(x^0) > 0$.*

Condition 1.2 implies Condition 1.1 (see [16, 26]). Hotta and Yoshise [16] pointed out that Condition 1.1 implies the well known Condition 1.5 in Kojima, Megiddo, and Noma [19]. As observed by Zhao and Li [42] (see also section 3 of this paper), Condition 1.5 in [19] implies that the solution set of the CP is nonempty and bounded. Thus, the above-mentioned Conditions 1.1 and 1.2 imply that the solution set of the CP is nonempty and bounded. Ravindran and Gowda (Corollary 5 in [27]) showed that a P_0-CP with a nonempty and bounded solution set must have a strictly feasible point. Moreover, for monotone CPs the converse is also true, i.e., the solution set of the monotone CP is nonempty and bounded if and only if it has a strictly feasible point. (See also Chen, Chen, and Kanzow [7].) This property of the monotone problem can be extended to the case of P_*-CPs. We recall that a map $f : R^n \to R^n$ is said to be a P_*-function if there exists a constant $\tau \geq 0$ such that

$$(1 + \tau) \sum_{i \in I_+} (x_i - y_i)(f_i(x) - f_i(y)) + \sum_{i \in I_-} (x_i - y_i)(f_i(x) - f_i(y)) \geq 0$$

for all distinct vectors x, y in R^n, where $I_+ = \{i : (x_i - y_i)(f_i(x) - f_i(y)) > 0\}$ and $I_- = \{1, \ldots, n\} \setminus I_+$. (See, Cottle, Pang, and Venkateswaran [9], Kojima et al. [20], Väliaho [33], Zhao and Han [38], and Zhao and Isac [39, 40].) Clearly a monotone function is a P_*-function, but the converse is not true. Zhao and Li [41, 42] pointed out that for a P_*-CP the following three conditions are equivalent:

(i) There exists a strictly feasible point.

(ii) The solution set of the CP is nonempty and bounded.

(iii) The central path of the CP exists.

Since most existing (interior-point and) non–interior-point path-following algorithms for CPs are based on the use of a certain continuation trajectory such as the central path, whose existence is closely related to the existence of a strictly feasible point, we conclude that for P_0-CPs these (interior-point and) non–interior-point algorithms are, in fact, confined to solving a class of strictly feasible problems. Other non-interior-point algorithms in the literature also suffer from the same restriction. For instance, the algorithms developed by Chen and Harker [5], Burke and Xu [1], and Chen and Chen [6] require the P_0 and R_0 assumption, which also implies that the solution set of the CP is nonempty and bounded, and hence the problem is strictly feasible. The strict feasibility condition plays an indispensable role in these known non–interior-point methods. In section 3, we give an example to show that Hotta and Yoshise's non–interior-point trajectory [16] does not necessarily exist when the problem has no strictly feasible point, in which case the solution set of the P_0-CP is unbounded (provided that it is nonempty). An interesting question is how to circumvent this difficulty so that a non–interior-point path-following method can be designed to solve a CP even when there is no strictly feasible point.

In this paper, we shall propose a new homotopy formulation of the CP. Based on this formulation, a new non–interior-point continuation trajectory for the CP can be generated. This new continuation trajectory possesses a desirable feature: For P_0-CPs, the existence and the boundedness of the continuation trajectory can be ensured under a mild condition that is weaker than most existing conditions like Conditions 1.1 and 1.2. The often assumed strict feasibility condition is not required here. In particular, for P_*-CPs, the proposed continuation trajectory exists and is bounded if and only if the problem has a solution. In other words, the existence and the boundedness of the trajectory for P_*-CPs do not require the strict feasibility condition (which is equivalent to the nonemptiness and boundedness of the solution set). We also (i) provide some sufficient conditions for the convergence of the entire trajectory as the parameter approaches zero and (ii) identify the properties of the limiting point of this trajectory. The results presented in the paper provide us with a theoretical basis for devising a new non–interior-point path-following method for CPs. This method can be expected to solve a more general class of complementarity problems than those to which most existing methods can be applied.

This paper is organized as follows. In section 2, we define a new homotopy formulation for the CP. In section 3, we specify a new properness condition that will be used to prove the existence and boundedness of a new continuation trajectory in section 4. We also compare this condition with several others known in the literature. The limiting behavior of the trajectory is studied in section 5. Final remarks are given in section 6.

Notation. We denote by R^n the space of n-dimensional real vectors, and by $R^n_+(R^n_{++}$, respectively) the nonnegative orthant (positive orthant, respectively). If $x \in R^n_+$ (R^n_{++}), we write $x \geq 0$ $(x > 0)$ for simplicity. All vectors, unless otherwise stated, are column vectors. T denotes the transpose of a vector. The symbol e denotes the vector in R^n with all of its components equal to one. For given vectors u, w, v in R^n, the triplet (u, w, v) (the pair (x, y)) denotes the column vector $(u^T, w^T, v^T)^T$ ($(x^T, y^T)^T$). For any $u \in R^n_+$, the symbol u^p denotes the pth power of the vector u, i.e., the vector $(u_1^p, \ldots, u_n^p)^T$, where $p > 0$ is a positive scalar. In particular, when $p = 1/2$, $\sqrt{u}$ denotes the vector $(\sqrt{u_1}, \ldots, \sqrt{u_n})^T$. The symbol diag$(x)$ denotes the $n \times n$ diagonal matrix whose (i, i)th entry is x_i. For any $x, y \in R^n$ with $x \leq y$, we

define the rectangular box $[x_1, y_1] \times \cdots \times [x_n, y_n]$ as $[x, y]$.

2. A new homotopy formulation for CPs. Let $(\bar{u}, \bar{v}, \bar{r})$ be a fixed point in $R^n_{++} \times R^{2n}$ and let

$$\bar{w} = \{\theta(\bar{u}, \bar{v}, \bar{r}) \in R^n_{++} \times R^{2n} : \theta \in (0, 1]\}.$$

Let $U : R^n_+ \times R^{2n} \to R^{3n}$ be defined by (1.1). Set

$$U^{-1}(\bar{w}) = \{z = (u, x, y) \in R^n_{++} \times R^{2n} : U(z) = \theta(\bar{u}, \bar{v}, \bar{r}) \text{ for some } \theta \in (0, 1]\}.$$

Under Condition 1.1, Hotta and Yoshise [16] showed that the above set forms a continuous trajectory leading to a solution of the CP. Based on this fact, they designed a globally convergent path-following method for the CP. However, it is easy to see that the strict feasibility condition plays an essential role in the existence of the Hotta and Yoshise trajectory. In fact, it is impossible to remove the strict feasibility condition from Conditions 1.1 and 1.2 without destroying the existence of their trajectory, as we see in the following example.

Example 2.1. Let $f(x) = Mx + q$, where

$$M = \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix}, \quad q = \begin{pmatrix} -1 \\ 0 \end{pmatrix}.$$

This function is a P_0-function and there exists no strictly feasible point. The solution set of the corresponding CP is unbounded. Let $\bar{u} = (\bar{u}_1, \bar{u}_2)^T \in R^2_{++}$, $\bar{v} = (\bar{v}_1, \bar{v}_2)^T \in R^2$, and $\bar{r} = (\bar{r}_1, \bar{r}_2)^T \in R^2$. From Lemma 1.1 in [16], the system $U(u, x, y) = \theta(\bar{u}, \bar{v}, \bar{r})$ can be written as follows:

$$u = \theta\bar{u}, \quad y = f(x) + \theta\bar{r}, \quad x - \theta\bar{v}/2 > 0, \quad y - \theta\bar{v}/2 > 0,$$
$$\operatorname{diag}(x - \theta\bar{v}/2)(y - \theta\bar{v}/2) = \theta\bar{u}.$$

Note that $y = f(x) + \theta\bar{r} = \binom{-x_2 - 1 + \theta\bar{r}_1}{\theta\bar{r}_2}$. The last equation above can be rewritten as

$$(x_1 - \theta\bar{v}_1/2)(-x_2 - 1 + \theta\bar{r}_1 - \theta\bar{v}_1/2) = \theta\bar{u}_1, \tag{2.1}$$

$$(x_2 - \theta\bar{v}_2/2)(\theta\bar{r}_2 - \theta\bar{v}_2/2) = \theta\bar{u}_2.$$

Since $\theta > 0$, the second equation above reduces to

$$(x_2 - \theta\bar{v}_2/2)(\bar{r}_2 - \bar{v}_2/2) = \bar{u}_2. \tag{2.2}$$

Case 1: $\bar{r}_2 \le \bar{v}_2/2$. Since $x_2 - \theta\bar{v}_2/2 > 0$ and $\bar{u}_2 > 0$, the above equation has no solution, and thus the system $U(u, x, y) = \theta(\bar{u}, \bar{v}, \bar{r})$ has no solution.

Case 2: $\bar{r}_2 > \bar{v}_2/2$. In this case, (2.2) can be written as

$$x_2 = (\bar{r}_2 - \bar{v}_2/2)^{-1}\bar{u}_2 + \theta\bar{v}_2/2.$$

Hence, for all sufficiently small $\theta > 0$, we have

$$-x_2 - 1 + \theta\bar{r}_1 - \theta\bar{v}_1/2 = -(\bar{r}_2 - \bar{v}_2/2)^{-1}\bar{u}_2 - \theta\bar{v}_2/2 - 1 + \theta\bar{r}_1 - \theta\bar{v}_1/2 < 0.$$

Since $x_1 - \theta\bar{v}_1/2 > 0$, we deduce from the above that (2.1) has no solution for all sufficiently small $\theta > 0$. Thus, the Hotta–Yoshise trajectory [16] does not exist.

Motivated by the above example, we now introduce a new homotopy formulation for the CP. Let $p \in (0,\infty)$ and $q \in [1,\infty)$ be two fixed numbers throughout the paper. Define the homotopy map $H : R^n_+ \times R^{2n} \to R^{3n}$ as follows:

$$H(u,x,y) = \begin{pmatrix} u \\ x+y-\sqrt{(x-y)^2+4u^q} \\ y-(f(x)+\mathrm{diag}\,(u^p)\,x) \end{pmatrix}, \quad (u,x,y) \in R^n_+ \times R^{2n}.$$

The above homotopy map is the focus of our study. It is worth mentioning that for each fixed vector $u > 0$, the function $f(x) + \mathrm{diag}\,(u^p)\,x$ can be viewed as a form of the renowned Tikhonov regularization of f, which was originally utilized to handle ill-posed problems. Recently, more attention has been paid to such techniques; see, e.g., Venkateswaran [34], Facchinei [10], Facchinei and Kanzow [11], Facchinei and Pang [12], Ravindran and Gowda [27], Gowda and Tawhid [13], Sznajder and Gowda [29], Qi [25], Sun [28], Tseng [31], and Zhao and Li [42]. To deal with the case of nonexistence of a strictly feasible point (or unboundedness of the solution set), we will see from the later discussion that it is a judicious choice to use the above new homotopy formulation of the CP.

The above homotopy map encompasses several extra variants. For instance, when $q = 1$ and $p \to \infty$, the above homotopy map, as u varies within the open rectangular box $(0, e)$, reduces to the one proposed by Hotta and Yoshise [16]. When $q = 2$ and $p \to \infty$, the above homotopy map, as u varies within $(0, e)$, is precisely the one studied by Burke and Xu [2, 3, 4], and Qi and Sun [26].

It is not difficult to see that if $H(u,x,y) = 0$, then (x,y) is a solution to the CP; conversely, if (x,y) is a solution to the CP, then $(0,x,y)$ is a solution to the equation $H(u,x,y) = 0$. Thus, a CP can be solved by locating a solution to the nonlinear equation $H(u,x,y) = 0$. The most widely used continuation method for solving this equation is the path-following algorithm that traces certain continuation trajectories leading to the solution set. We do not study such a numerical algorithm in this paper. The purpose here is to establish a theoretical basis for constructing a new non–interior-point path-following algorithm. Such a method can be used to solve a class of problems that is broader than those to which most existing path-following methods can be applied.

Given $(a,b,c) \in R^n_{++} \times R^{2n}$, we consider the system

$$H(u,x,y) = \theta(a,b,c), \tag{2.3}$$

where $\theta \in (0,1]$. Define $\bar{Z} = \{\theta(a,b,c) : \theta \in (0,1]\}$. In section 4, we will show that the set

$$H^{-1}(\bar{Z}) = \{(u,x,y) \in R^n_{++} \times R^{2n} : H(u,x,y) = \theta(a,b,c),\ \theta \in (0,1]\}$$

forms a unique, continuous curve leading to a solution of the CP under certain mild conditions. We now give two basic results that will be used later. The first result below gives an equivalent formulation of the system (2.3). This result plays a critical role in the analysis throughout the paper. For the given $(a,b,c) \in R^n_{++} \times R^{2n}$, we define the map $\mathcal{Y} : R^n \times (0,1] \to R^n$ by

$$\begin{aligned} \mathcal{Y}(x,\theta) := {} & x + f(x) + \theta^p \mathrm{diag}\,(a^p)\,x + \theta c \\ & + \sqrt{[x-(f(x)+\theta^p \mathrm{diag}\,(a^p)\,x+\theta c)]^2 + 4\theta^q a^q} - \theta b. \end{aligned} \tag{2.4}$$

LEMMA 2.1. *The solutions of* $\mathcal{Y}(x,\theta) = 0$ *are in one-to-one correspondence with those of* $H(u,x,y) = \theta(a,b,c)$. *Specifically, for the given scalar* $\theta \in (0,1]$, *if* (u,x,y) *is a solution to the system* $H(u,x,y) = \theta(a,b,c)$, *then* x *is a solution to the equation* $\mathcal{Y}(x,\theta) = 0$; *conversely, if* x *is a solution to the equation* $\mathcal{Y}(x,\theta) = 0$, *then* (u,x,y), *where* $u = \theta a$ *and* $y = f(x) + diag\{u^p\}x + \theta c$, *is a solution to the system* $H(u,x,y) = \theta(a,b,c)$.

Proof. The result is easy to show. Indeed, the equation $H(u,x,y) = \theta(a,b,c)$ is equivalent to the following system:

$$u = \theta a, \tag{2.5}$$

$$x + y - \sqrt{(x-y)^2 + 4u^q} = \theta b, \tag{2.6}$$

$$y = f(x) + \mathrm{diag}\{u^p\}x + \theta c. \tag{2.7}$$

Substituting (2.5) and (2.7) into (2.6) yields $\mathcal{Y}(x,\theta) = 0$. □

It is well known (see Lemma 1.1 in [16]) that for every nonnegative number $\mu \geq 0$, a triplet $(\alpha,\beta,\gamma) \in R^3$ satisfies

$$\phi(\mu,\alpha,\beta) = \alpha + \beta - \sqrt{(\alpha-\beta)^2 + 4\mu} = \gamma$$

if and only if $(\alpha - \gamma/2, \beta - \gamma/2) \geq 0$ and $(\alpha - \gamma/2)(\beta - \gamma/2) = \mu \geq 0$. Moreover, if $\mu > 0$, then $(\alpha - \gamma/2, \beta - \gamma/2) > 0$. By this fact, from (2.5)–(2.7) we have the following lemma.

LEMMA 2.2. *Let* $(a,b,c) \in R^n_{++} \times R^{2n}$ *be a fixed vector. Then for any* $\theta \in (0,1]$, *the vector* $(u(\theta), x(\theta), y(\theta))$ *is a solution to the system* (2.3) *if and only if it satisfies the following system:*

$$u(\theta) = \theta a, \tag{2.8}$$

$$y(\theta) = f(x(\theta)) + \theta^p \mathrm{diag}\,(a^p)\, x(\theta) + \theta c, \tag{2.9}$$

$$x(\theta) - \theta b/2 > 0, \quad y(\theta) - \theta b/2 > 0, \tag{2.10}$$

$$\mathrm{diag}\,(x(\theta) - \theta b/2)\,(y(\theta) - \theta b/2) = \theta^q a^q. \tag{2.11}$$

Remark 2.1. Since $\theta \in (0,1]$ and f is continuous, it follows from (2.8) and (2.9) that a sequence $\{(u(\theta_k), x(\theta_k), y(\theta_k))\}$, where $\theta_k \in (0,1]$, is unbounded if and only if $\{x(\theta_k)\}$ is unbounded. This fact will be frequently used in the later sections.

3. A new properness condition. In this section, we specify a new condition that is used to prove the existence and boundedness of the trajectory in the next section. To understand this condition better, we show first some properties of a semimonotone function. A map $f : R^n \to R^n$ is said to be semimonotone if, for any distinct vectors x, y in R^n with $x \geq y$, there exists a component i such that $x_i > y_i$ and $f_i(x) \geq f_i(y)$. It is evident that each P_0-function is semimonotone. The following result is a generalization of Lemma 1 in Ravindran and Gowda [27]. The proof is similar to the ones in such works as Tseng [30], Gowda and Tawhid [13], and Facchinei and Kanzow [11].

LEMMA 3.1. *Let* $f : R^n \to R^n$ *be a continuous semimonotone function. Let* $\{z^k\}$ *be an arbitrary sequence with* $\|z^k\| \to \infty$ *and* $z^k \geq \bar{z}$ *for all* k, *where* $\bar{z} \in R^n$ *is a fixed*

vector. Then there exist a subsequence of $\{z^k\}$*, denoted by* $\{z^{k_j}\}$*, and a fixed index* i_0 *such that* $z_{i_0}^{k_j} \to \infty$ *and* $f_{i_0}(z^{k_j})$ *is bounded from below.*

Proof. Passing through a subsequence, we may assume that there exists an index set I such that $z_i^k \to \infty$ for all $i \in I$, and $\{z_i^k\}$ is bounded for all $i \notin I$. Construct $\{y^k\}$ as follows: $y_i^k = \bar{z}_i$ if $i \in I$, and $y_i^k = z_i^k$ if $i \notin I$. Then we have that $z^k \neq y^k$ and $z^k \geq y^k$ for all sufficiently large k. By the semimonotone property of f, for each sufficiently large k there exists at least one index i such that $z_i^k > y_i^k$ and $f_i(z^k) \geq f_i(y^k)$. Thus, there exist an index $i_0 \in I$ and a subsequence of $\{z^k\}$, denoted by $\{z^{k_j}\}$, such that $z_{i_0}^{k_j} > y_{i_0}^{k_j}$ and $f_{i_0}(z^{k_j}) \geq f_{i_0}(y^{k_j})$ for all j. By this construction, $\{y^{k_j}\}$ is bounded and so is $\{f_{i_0}(y^{k_j})\}$. Hence $\{f_{i_0}(z^{k_j})\}$ is bounded from below. □

Given $(a, b, c) \in R^n_{++} \times R^{2n}$ and $\theta \in (0, 1]$, we define a function $\mathcal{F}_{(a,b,c,\theta)} : R^{2n} \to R^{2n}$ as follows:

$$\mathcal{F}_{(a,b,c,\theta)}(x, y) = \begin{pmatrix} Xy \\ y - f(x + \theta b/2) - \theta^p \mathrm{diag}\,(a^p)\, x - \theta c \end{pmatrix}, \tag{3.1}$$

where $X = \mathrm{diag}(x)$. The next property of semimonotone functions is one of the motivations for our new properness condition.

PROPOSITION 3.1. *Let* $f : R^n \to R^n$ *be a continuous semimonotone function. Then for any* $(a, b, c, \theta) \in R^n_{++} \times R^{2n} \times (0, 1]$*, the set*

$$\mathcal{F}^{-1}_{(a,b,c,\theta)}(D) = \{(x, y) \in R^n_{++} : \mathcal{F}_{(a,b,c,\theta)}(x, y) \in D\}$$

is bounded for any compact set D *in* $R^n_+ \times R^n$*.*

Proof. Assume the contrary: there exist certain $(a', b', c', \theta') \in R^n_{++} \times R^{2n} \times (0, 1]$ and a compact set $D' \subseteq R^n_+ \times R^n$ such that $\mathcal{F}^{-1}_{(a',b',c',\theta')}(D')$ is unbounded. Let $\{(x^k, y^k)\} \subseteq \mathcal{F}^{-1}_{(a',b',c',\theta')}(D')$ such that $\|(x^k, y^k)\| \to \infty$. Notice that

$$\mathcal{F}_{(a',b',c',\theta')}(x^k, y^k) = \begin{pmatrix} X^k y^k \\ y^k - f(x^k + \theta' b'/2) - (\theta')^p \mathrm{diag}\,((a')^p)\, x^k - \theta' c' \end{pmatrix} \in D'.$$

There is a sequence $\{(u^k, v^k)\} \subseteq D'$, where $u^k \in R^n_+$ and $v^k \in R^n$, such that for all k we have

$$X^k y^k = u^k \geq 0, \tag{3.2}$$

$$y^k = f(x^k + \theta' b'/2) + (\theta')^p \mathrm{diag}\,((a')^p)\, x^k + \theta' c' + v^k. \tag{3.3}$$

Since $\{(u^k, v^k)\}$ is bounded and since $\|(x^k, y^k)\| \to \infty$, by continuity we conclude that the sequence $\{x^k\}$ is unbounded. Thus, we may assume that $\|x^k\| \to \infty$. Since $x^k \in R^n_+$ for all k, passing through a subsequence we may assume that there exists an index set I such that $x_i^k \to \infty$ for all $i \in I$, and $\{x_i^k\}$ is bounded for all $i \notin I$. Since $x_i^k \to \infty$ for all $i \in I$, it follows from (3.2) that $y_i^k \to 0$ for all $i \in I$. Hence from (3.3) we have

$$f_i(x^k + \theta' b'/2) = y_i^k - (\theta')^p \mathrm{diag}\,((a')^p)\, x_i^k - \theta' c'_i - v_i^k \to -\infty$$

for all $i \in I$. However, by Lemma 3.1 there exists an index $i \in I$ such that $\{f_i(x^k + \theta' b'/2)\}$ is bounded from below. This is a contradiction. □

As a particular case, let $D_r := [0, rv_1] \times [-rv_2, rv_3] \subseteq R^n_+ \times R^n$, where $v_i \in R^n_+ (i = 1, 2, 3)$ and r is a nonnegative number. We deduce from the above proposition that the

set $\mathcal{F}^{-1}_{(a,b,c,r)}(D_r)$ is bounded for any $0 < r < \infty$ if f is continuous and semimonotone. Inspired by this observation, we impose the following properness condition on the CP.

CONDITION 3.1. *For any given $(a, b, c) \in R^n_{++} \times R^{2n}$ and scalar $\hat{t} \geq 0$ there exists a scalar $1 \geq \theta^* > 0$ such that*

$$\bigcup_{\theta \in (0,\theta^*]} \mathcal{F}^{-1}_{(a,b,c,\theta)}(D_\theta)$$

is bounded, where

$$\mathcal{F}^{-1}_{(a,b,c,\theta)}(D_\theta) := \{(x, y) \in R^{2n}_{++} : \mathcal{F}_{(a,b,c,\theta)}(x, y) \in D_\theta\}$$

and

$$D_\theta := [0, \theta a^q] \times [-\theta \hat{t} e, \theta \hat{t} e] \subseteq R^n_+ \times R^n.$$

Notice that for a fixed $\bar{\theta} \in (0, 1)$, the above set $D_\theta \subseteq D_{\bar{\theta}} := [0, \bar{\theta} e] \times [-\bar{\theta} e, \bar{\theta} e]$ for all sufficiently small θ. Thus, we can see that Condition 3.1 holds if the following condition is satisfied.

CONDITION 3.2. *For any given $(a, b, c) \in R^n_{++} \times R^{2n}$ there exists a scalar $1 > \bar{\theta} > 0$ such that*

$$\bigcup_{\theta \in (0,\bar{\theta}]} \mathcal{F}^{-1}_{(a,b,c,\theta)}(D_{\bar{\theta}})$$

is bounded, where $D_{\bar{\theta}} = [0, \bar{\theta} e] \times [-\bar{\theta} e, \bar{\theta} e]$ and

$$\mathcal{F}^{-1}_{(a,b,c,\theta)}(D_{\bar{\theta}}) = \{(x, y) \in R^n_{++} : \mathcal{F}_{(a,b,c,\theta)}(x, y) \in D_{\bar{\theta}}\}.$$

At first glance, Condition 3.1 may seem to be a little unusual. As we will see subsequently, this condition is actually quite weak. A prominent feature of Condition 3.1 is that it may hold even when the solution set of the CP is unbounded or the strict feasibility condition fails to hold. Specifically, for P_0-complementarity problems we will show that previously known conditions such as Condition 1.1, Condition 1.2, the nonemptiness and boundedness assumption of the solution set (in particular, the P_0 together with R_0 property), and Condition 1.5 in [19] all imply Condition 3.1. However, the converse is not true (see Theorem 3.1 below). Before we prove this fact, we list some helpful results. The following result is easy to prove by using the compactness of S and continuity of f. Its proof is omitted.

LEMMA 3.2. *Let S be a compact set in R^n and $(a, b, c) \in R^n_{++} \times R^{2n}$ be a fixed triplet. Let f be a continuous function from R^n into itself.*

(i) *Let*

$$G(x) := x + f(x) - \sqrt{(x - f(x))^2}.$$

Define $\bar{G} : R^n \times (0, 1] \times R^n_+ \times R^n \to R^n$ by

$$\bar{G}(x, \theta, w, v) = x + f(x) + \theta^p \mathrm{diag}\,(a^p)\,(x - \theta b/2) + \theta(c + b/2) + v$$
$$- \sqrt{[x - (f(x) + \theta^p \mathrm{diag}\,(a^p)\,(x - \theta b/2) + \theta(c + b/2) + v)]^2 + 4w} - \theta b.$$

Then for any $\delta > 0$ there exists a scalar $\bar{\theta} \in (0, 1]$ such that for all $\theta \in (0, \bar{\theta}]$ and $(w, v) \in [0, \bar{\theta} e] \times [-\bar{\theta} e, \bar{\theta} e] \subseteq R^n_+ \times R^n$ we have

$$\sup_{x \in S} \|\bar{G}(x, \theta, w, v) - G(x)\| < \delta.$$

(ii) *Given any* $\hat{\theta} \in (0,1)$*, then for any* $\delta > 0$ *there exists a sufficiently small scalar* $\beta > 0$ *such that*

$$\sup_{x \in S} \|\mathcal{Y}(x,\theta) - \mathcal{Y}(x,\hat{\theta})\| < \delta \quad \text{for all } \theta \text{ such that } |\theta - \hat{\theta}| < \beta,$$

where $\mathcal{Y}(x,\theta)$ *is defined by* (2.4).

The next result, which was pointed out by Gowda and Tawhid [13], is very useful for the subsequent analysis.

LEMMA 3.3. *Let* $\Phi(x,v) = x + f(x) - \sqrt{(x - f(x))^2 + v^2}$*, where* $v \in R^n$.

(i) *If* f *is a* P_0*-function, then* $\Phi(x,v)$ *is a* P_0*-function in* x*. Moreover, if* $v^2 \in R^n_{++}$*, then* $\Phi(x,v)$ *is a P-function in* x.

(ii) *If* f *is a P-function, then* $\Phi(x,v)$ *is a P-function in* x.

The following upper-semicontinuity property of a P_0-function is due to Ravindran and Gowda [27].

LEMMA 3.4. *Let* $g : R^n \to R^n$ *be a* P_0*-function. Suppose that* $g^{-1}(0)$ *is nonempty and compact. Then for any given* $\varepsilon > 0$ *there exists a scalar* $\gamma > 0$ *such that for any* P_0*-function* h *with*

$$\sup_{\bar{\Omega}} \|h(x) - g(x)\| < \gamma$$

we have

$$\emptyset \neq h^{-1}(0) \subseteq g^{-1}(0) + \varepsilon B,$$

where B *denotes the open unit ball in* R^n *and* $\bar{\Omega}$ *is the closure of the set* $\Omega = g^{-1}(0) + \varepsilon B$.

We now show that several well-known existing conditions used in the literature of interior-point and non–interior-point methods imply Condition 3.1. However, the converse is not true, since Condition 3.1 may hold for P_0-CPs in the absence of the strict feasibility condition.

THEOREM 3.1. *Let* f *be a* P_0*-function. If one of the following condition holds,*

(i) *Condition* 1.1*,*

(ii) *Condition* 1.2*,*

(iii) *Condition* 1.5 *in Kojima, Magiddo, and Noma* [19]*,*

(iv) *the solution set of the CP is nonempty and bounded,*

(v) f *is a* P_0 *and* R_0 *function* [1, 6]*,*

then Condition 3.1 *holds. However, the converse is not true, i.e., Condition* 3.1 *does not imply any one of the above conditions.*

Proof. The implication (ii) $\Rightarrow$ (i) is pointed out in [16, 26]. It is easy to verify that (i) $\Rightarrow$ (iv). In fact, if (i) holds, Hotta and Yoshise [16] showed that their non–interior-point trajectory exists and a subtrajectory is bounded, and hence each of the accumulation points of the subtrajectory is a solution to the CP. Hence the solution set of the CP is not empty. We further demonstrate that it is bounded. Indeed, by Condition 1.1, there is a point $x^0 > 0$ such that $f(x^0) > 0$. It follows from Lemma 2.1 in [16] that $R^n_- \times R^n_+ \subseteq V(R^n_{++} \times R^{2n})$. Thus, by Condition 1.1 again, the set $U^{-1}(D)$ is bounded for every compact subset D of $R^n_+ \times R^n_- \times R^n_+$. In particular, set

$$D := \{(0,0,0)\} \subseteq R^n_+ \times R^n_- \times R^n_+.$$

Then the set $U^{-1}(0)$ is bounded. The set $U^{-1}(0)$ coincides with the solution set of the CP. Hence (i) $\Rightarrow$ (iv).

By a proof similar to the above, we can show that (iii)⇒ (iv). The implication of (v) ⇒ (iv) is a known result.

Therefore, to show that each condition of (i)–(v) implies Condition 3.1, it is sufficient to prove that (iv) implies Condition 3.1. Indeed, assume that f is a P_0-function and the solution set of the CP is nonempty and bounded. We show that Condition 3.2 holds (and hence Condition 3.1 holds). Let $G : R^n \to R^n$ be given by

$$G(x) := x + f(x) - \sqrt{(x - f(x))^2},$$

which is a P_0-function (Lemma 3.3). Since $G^{-1}(0) = \{x \in R^n : G(x) = 0\}$ coincides with the solution set of the CP, by the assumption, the set $G^{-1}(0)$ is nonempty and bounded; in fact, it is a compact set by the continuity of f. For any scalar $\varepsilon > 0$, by Lemma 3.4 there is a scalar $\delta > 0$ such that for any P_0-function $h : R^n \to R^n$ with

$$\sup_{x \in \bar{\Omega}} \|h(x) - G(x)\| < \delta, \tag{3.4}$$

where $\Omega = G^{-1}(0) + \varepsilon B$,

$$0 \neq h^{-1}(0) \subseteq G^{-1}(0) + \varepsilon B. \tag{3.5}$$

Let (a, b, c) be a fixed triplet in $R^n_{++} \times R^{2n}$, and let $\bar{G}(x, \theta, w, v)$ be given as in Lemma 3.2, where $\theta \in (0, 1]$ and $(w, v) \in R^n_+ \times R^n$. Clearly, the function

$$f(x) + \operatorname{diag}(\theta^p a^p)(x - \theta b/2) + \theta(c + b/2) + v$$

is a P-function in x. Since $w \in R^n_+$, it follows from (ii) of Lemma 3.3 that $\bar{G}(x, \theta, w, v)$ is a P-function in x. By (i) of Lemma 3.2, there exists a sufficiently small number $\bar{\theta} \in (0, 1)$ such that for all $\theta \in (0, \bar{\theta}]$ and $(w, v) \in [0, \bar{\theta}e] \times [-\bar{\theta}e, \bar{\theta}e]$ we have

$$\sup_{x \in \bar{\Omega}} \|\bar{G}(x, \theta, w, v) - G(x)\| < \delta.$$

Thus, setting $h(x) := \bar{G}(x, \theta, w, v)$ in (3.4), we have from (3.5) that

$$\emptyset \neq \bar{G}^{-1}_{(\theta,w,v)}(0) \subseteq G^{-1}(0) + \varepsilon B$$

for all $\theta \in (0, \bar{\theta}]$ and $(w, v) \in [0, \bar{\theta}] \times [-\bar{\theta}e, \bar{\theta}e]$, where

$$\bar{G}^{-1}_{(\theta,w,v)}(0) = \{x \in R^n : \bar{G}(x, \theta, w, v) = 0\}.$$

Hence,

$$\bigcup_{(\theta,w,v) \in (0,\bar{\theta}] \times D_{\bar{\theta}}} \bar{G}^{-1}_{(\theta,w,v)}(0) \subseteq G^{-1}(0) + \varepsilon B,$$

where $D_{\bar{\theta}} = [0, \bar{\theta}e] \times [-\bar{\theta}e, \bar{\theta}e]$. On the other hand, it is easy to verify that

$$\bar{G}(x, \theta, w, v) = 0, \qquad \theta \in (0, \bar{\theta}],\ (w, v) \in D_{\bar{\theta}},$$

if and only if

$$\begin{aligned}
&x - \theta b/2 \geq 0, \qquad y - \theta b/2 \geq 0,\\
&\operatorname{diag}(x - \theta b/2)(y - \theta b/2) = w,\\
&y - \theta b/2 - f(x) - \theta^p \operatorname{diag}(a^p)(x - \theta b/2) - \theta c = v,\\
&\theta \in (0, \bar{\theta}], \qquad (w, v) \in D_{\bar{\theta}}.
\end{aligned}$$

Define $\bar{x} = x - \theta b/2$ and $\bar{y} = y - \theta b/2$. The above system can be rewritten as

$$\mathcal{F}_{(a,b,c,\theta)}(\bar{x}, \bar{y}) \in D_{\bar{\theta}}, \quad (\bar{x}, \bar{y}) \geq 0, \ \theta \in (0, \bar{\theta}],$$

where $\mathcal{F}_{(a,b,c,\theta)}$ is defined by (3.1). Denote

$$\mathcal{F}^{-1}_{(a,b,c,\theta)}(D_{\bar{\theta}}) = \{(u, v) \in R^{2n}_{++} : \mathcal{F}_{(a,b,c,\theta)}(u, v) \in D_{\bar{\theta}}\}.$$

Then from the above discussion, we deduce that

$$\begin{aligned}
&\{x \in R^n : x = \bar{x} + \theta b/2, (\bar{x}, \bar{y}) \in \mathcal{F}^{-1}_{(a,b,c,\theta)}(D_{\bar{\theta}}), \ \theta \in (0, \bar{\theta}]\} \\
&\subseteq \{x \in R^n : x = \bar{x} + \theta b/2, \mathcal{F}_{(a,b,c,\theta)}(\bar{x}, \bar{y}) \in D_{\bar{\theta}}, (\bar{x}, \bar{y}) \geq 0, \ \theta \in (0, \bar{\theta}]\} \\
&= \bigcup_{(\theta,w,v) \in (0,\bar{\theta}] \times D_{\bar{\theta}}} \bar{G}^{-1}_{(\theta,w,v)}(0) \\
&\subseteq G^{-1}(0) + \varepsilon B.
\end{aligned}$$

Since $G^{-1}(0) + \varepsilon B$ is bounded, we deduce from the above that

$$\bigcup_{\theta \in (0,\bar{\theta}]} \mathcal{F}^{-1}_{(a,b,c,\theta)}(D_{\bar{\theta}})$$

is bounded. Hence, Condition 3.2 is satisfied, and thus Condition 3.1 holds.

Since each of the conditions listed in the theorem implies the existence of a strictly feasible point, to show that Condition 3.1 does not imply each of these conditions it suffices to prove that Condition 3.1 may hold even when there is no strictly feasible point. Now consider, in R^2, the following example:

$$f(x) = \begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 0 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 2x_2 - 1 \end{pmatrix},$$

which is a P_0-function. Clearly, f has no strictly feasible point, and the corresponding complementarity problem has an unbounded solution set. However, this example does satisfy Condition 3.1. Indeed, choose $p \in (0, 1]$ and $q \in [1, \infty]$ and let $(a, b, c) \in R^2_{++} \times R^2 \times R^2$ be a fixed vector. We show for each scalar $0 < \theta^* < 1$ that the set $\cup_{\theta \in (0,\theta^*]} \mathcal{F}^{-1}_{(a,b,c,\theta)}(D_\theta)$ is a bounded set, where all symbols are defined as in Condition 3.1. Assume that $\{(x^k, y^k)\}$ is an arbitrary sequence contained in the set. Then $(x^k, y^k) > 0$, and for each (x^k, y^k) there is a scalar $\theta_k \in (0, \theta^*]$ such that $\mathcal{F}_{(a,b,c,\theta_k)}(x^k, y^k) \in D_{\theta_k}$. By the definitions of D_{θ_k} and $\mathcal{F}_{(a,b,c,\theta)}$, there exist two vectors $d^k \in [0, a^q]$ and $\bar{d}^k \in [-\hat{t}e, \hat{t}e]$ such that

$$\begin{aligned}
&X^k y^k = \theta_k d^k \in [0, \theta_k a^q], \\
&y^k - f(x^k + \theta_k b/2) - \theta_k^p \mathrm{diag}(a^p) x^k - \theta_k c = \theta_k \bar{d}^k \in \theta_k [-\hat{t}e, \hat{t}e],
\end{aligned}$$

where $X^k = \mathrm{diag}(x^k)$. For this example, the second equation above can be rewritten as

$$y^k_1 = (\theta_k a_1)^p x^k_1 + \theta_k c_1 + \theta_k \bar{d}^k_1,$$

$$y^k_2 = 2(x^k_2 + \theta_k b_2/2) - 1 + (\theta_k a_2)^p x^k_2 + \theta_k c_2 + \theta_k \bar{d}^k_2.$$

Thus, from $X^k y^k = \theta_k d^k$ we have

$$\theta_k d_1^k = x_1^k y_1^k = (\theta_k a_1)^p (x_1^k)^2 + \theta_k x_1^k c_1 + \theta_k x_1^k \bar{d}_1^k,$$

i.e.,

$$\theta_k^{1-p} d_1^k = a_1^p (x_1^k)^2 + \theta_k^{1-p} (c_1 + \bar{d}_1^k) x_1^k,$$

and

$$\theta_k d_2^k = x_2^k y_2^k = (2 + (\theta_k a_2)^p)(x_2^k)^2 + [\theta_k (b_2 + c_2 + \bar{d}_2^k) - 1] x_2^k.$$

From the above two relations, we conclude that the sequence $\{x^k\}$ is bounded, and by continuity so is $\{y^k\}$. Therefore, the set $\cup_{\theta \in (0,\theta^*]} \mathcal{F}^{-1}_{(a,b,c,\theta)}(D_\theta)$ is bounded, i.e., Condition 3.1 is satisfied. □

4. Existence and boundedness of the trajectory. The purpose of this section is to show the existence and the boundedness of the proposed continuation trajectory for P_0-CPs under Condition 3.1. To begin with, we recall a useful result on the degree of a continuous function. Let Ω be a bounded open set in R^n. The symbols $\bar{\Omega}$ and $\partial\Omega$ denote the closure and boundary of Ω, respectively. Let h be a continuous function from $\bar{\Omega}$ into R^n. For any vector $y \in R^n$ such that $y \notin h(\partial\Omega)$, the degree of h at y with respect to Ω is defined by $\deg(h, \Omega, y)$. The following result can be found in Lloyd [22].

LEMMA 4.1. (i) *If h is injective on R^n, then for any $y \in h(\Omega)$, $|deg(h, \Omega, y)| = 1$.*
(ii) *If $deg(h, \Omega, y) \neq 0$, then the equation $h(x) = y$ has a solution in Ω.*
(iii) *Let g be a continuous function from $\bar{\Omega} \to R^n$. Let*

$$\mathcal{H}(x,t) = tg(x) + (1-t)h(x), \quad 0 \leq t \leq 1.$$

If $y \notin \{\mathcal{H}(x,t) : x \in \partial\Omega, t \in [0,1]\}$, then $deg(g, \Omega, y) = deg(h, \Omega, y)$.

We are ready to prove a general and essential result.

THEOREM 4.1. *Let (a,b,c) be a fixed vector in $R^n_{++} \times R^{2n}$. Let $f : R^n \to R^n$ be a continuous semimonotone function.*
(i) *For each $\theta \in (0,1]$, the system (2.3) has a solution.*
(ii) *If Condition 3.1 holds, then the set*

$$\{(u,x,y) \in \mathcal{T}_\theta : \ \theta \in (0,1]\} := \bigcup_{\theta \in (0,1]} \mathcal{T}_\theta \tag{4.1}$$

is bounded, where

$$\mathcal{T}_\theta := \{(u,x,y) : \ H(u,x,y) = \theta(a,b,c)\}.$$

Proof. Let f be a continuous semimonotone function. We show the result by contradiction. Assume that there is a scalar $\hat{\theta} \in (0,1]$ such that system (2.3) has no solution.

Let $\mathcal{H} : R^n \times [0,1] \to R^n$ be defined by

$$\mathcal{H}(x,t) = t(x - \hat{\theta}b/2) + (1-t)\mathcal{Y}(x,\hat{\theta}), \quad t \in [0,1],$$

where $\mathcal{Y}$ is given by (2.4). We first show that the set

$$\mathcal{S} = \{x \in R^n : \mathcal{H}(x,t) = 0 \ \text{ for some } t \in [0,1]\}$$

is unbounded. Indeed, assume the contrary: $\mathcal{S}$ is bounded. Then for any fixed $\varepsilon > 0$, the set $D := (\{\hat{\theta}b/2\} \cup \mathcal{S}) + \varepsilon B$ is a bounded open set in R^n, where B is the open unit ball in R^n. Clearly, the intersection of $\mathcal{S}$ and the boundary of D is empty, i.e., for all $x \in \partial D, \mathcal{H}(x,t) \neq 0$ for all $t \in [0,1]$. Therefore,

$$|\deg(\mathcal{Y}(\cdot,\hat{\theta}), D, 0)| = |\deg(g, D, 0)| = 1,$$

where $g(x) := x - \hat{\theta}b/2$. The first part of the equation above follows from (iii) of Lemma 4.1, and the second part of the equation follows from (i) of Lemma 4.1 since g is an injective mapping. Thus, it follows from (ii) of Lemma 4.1 that the equation $\mathcal{Y}(x,\hat{\theta}) = 0$ has a solution (in D). Thus, by Lemma 2.1, $H(u,x,y) = \hat{\theta}(a,b,c)$ has a solution. This contradicts our assumption at the beginning of the proof. Therefore, the set $\mathcal{S}$ is unbounded.

Since $\mathcal{S}$ is unbounded, there is a sequence $\{x^k\}$ contained in $\mathcal{S}$ such that $\|x^k\| \to \infty$. We now show that $\{x^k\}$ satisfies the relations

$$x^k - \hat{\theta}b/2 > 0$$

and

$$f(x^k) \leq -\hat{\theta}^p \mathrm{diag}\,(a^p)\, x^k + 2\hat{\theta}^q [\mathrm{diag}(x^k - \hat{\theta}b/2)]^{-1} a^q + \hat{\theta}(b/2 - c)$$

for all sufficiently large k. Indeed, since $\|x^k\| \to \infty$, there is a $k_0 > 0$ such that $\|x^k - \hat{\theta}b/2\| > 0$ for all $k > k_0$. Since $x^k \in \mathcal{S}$, by the definition of $\mathcal{S}$ there is a scalar $t^k \in [0,1]$ such that

$$\mathcal{H}(x^k, t^k) = t^k(x^k - \hat{\theta}b/2) + (1 - t^k)\mathcal{Y}(x^k, \hat{\theta}) = 0.$$

Since $x^k \neq \hat{\theta}b/2$ for all $k > k_0$, we deduce from the above that $t^k \neq 1$ for all $k > k_0$. Since system (2.3) has no solution (by the assumption), i.e., $\mathcal{Y}(x,\hat{\theta}) \neq 0$ for all $x \in R^n$ (Lemma 2.1), it follows from the above equation that $t^k \neq 0$. Therefore, we have that $t^k \in (0,1)$ for all $k > k_0$. The above equation can be written as

$$\begin{aligned} &t^k(x^k - \hat{\theta}b/2) + (1 - t^k)[(x^k - \hat{\theta}b/2) + (y^k - \hat{\theta}b/2)] \\ &- (1 - t^k)\sqrt{[(x^k - \hat{\theta}b/2) - (y^k - \hat{\theta}b/2)]^2 + 4\hat{\theta}^q a^q} = 0, \end{aligned}$$

where

$$y^k = f(x^k) + \hat{\theta}^p \mathrm{diag}\,(a^p)\, x^k + \hat{\theta}c.$$

Define

$$\hat{x}^k = x^k - \hat{\theta}b/2, \qquad \hat{y}^k = y^k - \hat{\theta}b/2.$$

The above equation can be further written as

$$\hat{x}^k + (1 - t^k)\hat{y}^k = (1 - t^k)\sqrt{(\hat{x}^k - \hat{y}^k)^2 + 4\hat{\theta}^q a^q}. \tag{4.2}$$

Squaring both sides of this equation and simplifying (all the algebraic operations are performed componentwise), we obtain

$$t^k(2 - t^k)(\hat{x}^k)^2 + 2(1 - t^k)(2 - t^k)\hat{X}^k\hat{y}^k = 4(1 - t^k)^2\hat{\theta}^q a^q,$$

where $\hat{X}^k = \mathrm{diag}(\hat{x}^k)$. It follows from the above that $\hat{x}_i^k \neq 0$ for all $i = 1, 2, \ldots, n$. Multiplying both sides of the above by $(\hat{X}^k)^{-1}$ and dividing both sides by $2(1-t^k)(2-t^k)$ yield

$$\hat{y}^k = -\frac{t^k}{2(1-t^k)}\hat{x}^k + \frac{2(1-t^k)\hat{\theta}^q}{2-t^k}(\hat{X}^k)^{-1}a^q.$$

Thus, we have

$$\hat{x}^k + (1-t^k)\hat{y}^k = \left(1 - \frac{t^k}{2}\right)\hat{x}^k + \frac{2(1-t^k)^2\hat{\theta}^q}{2-t^k}(\hat{X}^k)^{-1}a^q.$$

If $\hat{x}_i^k \leq 0$ for some i, then we have from the above that $\hat{x}_i^k + (1-t^k)\hat{y}_i^k \leq 0$, which contradicts the right-hand side of (4.2). Thus, $\{\hat{x}^k\} \subseteq R^n_{++}$ and

$$\hat{y}^k \leq \frac{2(1-t^k)\hat{\theta}^q}{2-t^k}(\hat{X}^k)^{-1}a^q \leq 2\hat{\theta}^q(\hat{X}^k)^{-1}a^q.$$

That is, $x^k - \hat{\theta}b/2 > 0$ and

$$f(x^k) \leq -\hat{\theta}^p \mathrm{diag}\,(a^p)\,x^k + 2\hat{\theta}^q\left[\mathrm{diag}\left(x^k - \hat{\theta}b/2\right)\right]^{-1}a^q + \hat{\theta}(b/2 - c) \tag{4.3}$$

for all $k > k_0$. Passing through a subsequence, we may suppose that there is an index set I such that $x_i^k \to \infty$ for all $i \in I$ and $\{x_i^k\}$ is bounded for $i \notin I$. It follows from (4.3) that $f_i(x^k) \to -\infty$ for all $i \in I$. This contradicts the consequence of Lemma 3.1, which states that there exists an index $i \in I$ such that $f_i(x^k)$ is bounded from below. Hence, item (i) of the theorem is shown.

We now prove item (ii) of the theorem, i.e., the boundedness of set (4.1). Assume the contrary: the set $\{(u, x, y) \in \mathcal{T}_\theta : \theta \in (0, 1]\}$ is unbounded, i.e., there exists an unbounded sequence $\{(u(\theta_k), x(\theta_k), y(\theta_k))\}$ contained in the set, where $0 < \theta_k \leq 1$. Thus, the sequence $\{x(\theta_k)\}$ is unbounded (Remark 2.1). Without loss of generality, we assume that $\|x(\theta_k)\| \to \infty$ as $k \to \infty$. Note that $\{(u(\theta_k), x(\theta_k), y(\theta_k))\}$ satisfies the system (2.8)–(2.11), where θ is replaced by θ_k. By the unboundedness of $\{x(\theta_k)\}$ and $x(\theta_k) \geq \hat{\theta}b/2$, it follows from Lemma 3.1 that there exist a subsequence of $\{x(\theta_k)\}$ denoted also by $\{x(\theta_k)\}$ and an index m such that $x_m(\theta_k) \to \infty$ and $f_m(x(\theta_k))$ is bounded from below. From (2.11) we have

$$y_m(\theta_k) - \theta_k b_m/2 = \frac{\theta_k^q a_m^q}{x_m(\theta_k) - \theta_k b_m/2},$$

and by using (2.9) we obtain

$$f_m(x(\theta_k)) = \theta_k b_m/2 + \frac{\theta_k^q a_m^q}{x_m(\theta_k) - \theta_k b_m/2} - \theta_k c_m - (\theta_k a_m)^p x_m(\theta_k).$$

Since $x_m(\theta_k) \to \infty$ and $f_m(x(\theta_k))$ is bounded from below, we deduce from the above that $\theta_k^p \to 0$, and thus $\theta_k \to 0$. Define

$$\hat{x}(\theta_k) = x(\theta_k) - \theta_k b/2, \qquad \hat{y}(\theta_k) = y(\theta_k) - \theta_k b/2.$$

By (2.10) and (2.11), we have

$$(\hat{x}(\theta_k), \hat{y}(\theta_k)) > 0, \qquad \hat{X}(\theta_k)\hat{x}(\theta_k) = \theta_k^q a^q.$$

By using (2.9) again, we have

$$\begin{aligned}
&\hat{y}(\theta_k) - f(\hat{x}(\theta_k) + \theta_k b/2) - \theta_k^p \mathrm{diag}\,(a^p)\,\hat{x}(\theta_k) - \theta_k c \\
&= y(\theta_k) - \theta_k b/2 - f(x(\theta_k)) - \theta_k^p \mathrm{diag}\,(a^p)\,x(\theta_k) + \theta_k^p \mathrm{diag}\,(a^p)\,(\theta_k b/2) - \theta_k c \\
&= -\theta_k b/2 + \theta_k^p \mathrm{diag}\,(a^p)\,(\theta_k b/2) \\
&= \theta_k[-b/2 + \theta_k^p \mathrm{diag}\,(a^p)\,b/2].
\end{aligned}$$

Let $\hat{t} = \| - b/2\|_\infty + \|\mathrm{diag}(a^p)b/2\|_\infty$. Then, for any $\theta^k \in (0,1]$, we have $-\hat{t}e \le -b/2 + \theta_k^p \mathrm{diag}(a^p)b/2 \le \hat{t}e$. Therefore,

$$\theta_k^q a^q = \theta_k(\theta_k^{q-1} a^q) \in [0, \theta_k a^q],$$

$$\theta_k(-b/2 + \theta_k^p \mathrm{diag}\,(a^p)\,b/2) \in \theta_k[-\hat{t}e, \hat{t}e].$$

Therefore,

$$\begin{aligned}
\mathcal{F}_{(a,b,c,\theta_k)}(\hat{x}(\theta_k), \hat{y}(\theta_k)) &= \begin{pmatrix} \hat{X}(\theta_k)\hat{y}(\theta_k) \\ \hat{y}(\theta_k) - f(\hat{x}(\theta_k) + \theta_k b/2) - \theta_k^p \mathrm{diag}\,(a^p)\,\hat{x}(\theta_k) - \theta_k c \end{pmatrix} \\
&\in [0, \theta_k a^q] \times \theta_k[-\hat{t}e, \hat{t}e] =: D_{\theta_k}
\end{aligned}$$

for all $\theta_k \in (0,1]$. That is, $(\hat{x}(\theta_k), \hat{y}(\theta_k)) \in \mathcal{F}^{-1}_{(a,b,c,\theta_k)}(D_{\theta_k})$ for all $\theta_k \in (0,1]$. Hence, for any $1 \ge \theta^* > 0$, the sequence

$$\{(\hat{x}(\theta_k), \hat{y}(\theta_k)) : \theta_k \in (0, \theta^*]\} \subseteq \bigcup_{\theta_k \in (0,\theta^*]} \mathcal{F}^{-1}_{(a,b,c,\theta_k)}(D_{\theta_k}) \subseteq \bigcup_{\theta \in (0,\theta^*]} \mathcal{F}^{-1}_{(a,b,c,\theta)}(D_\theta).$$

By Condition 3.1, there exists a $\theta^* \in (0,1]$ such that the right-hand side of the above is bounded. However, the left-hand side is an unbounded sequence. This contradiction shows that the set (4.1) is indeed bounded. The proof is thus complete. □

Since Condition 3.2 implies Condition 3.1, the following result is an immediate consequence of Theorem 4.1.

COROLLARY 4.1. *Let $f : R^n \to R^n$ be a continuous semimonotone function. If Condition* 3.2 *holds, then the set* (4.1) *is bounded.*

While system (2.3) has a solution for a continuous semimonotone function if Condition 3.1 holds, it is not clear whether the solution of system (2.3) is unique for each $\theta \in (0,1]$. However, for continuous P_0-functions, which are special cases of continuous semimonotone functions, it is easy to prove that for each $\theta \in (0,1]$ system (2.3) has a unique solution which is also continuous in θ. We summarize the result as follows.

THEOREM 4.2. *Let $f : R^n \to R^n$ be a continuous P_0-function.*

(i) *For each $\theta \in (0,1]$, system* (2.3) *has a unique solution $(u(\theta), x(\theta), y(\theta))$ which is continuous in θ.*

(ii) *If Condition* 3.1 *(in particular, Condition* 3.2*) is satisfied, then the entire trajectory $\{(u(\theta), x(\theta), y(\theta)) : \theta \in (0,1]\}$ is bounded. Hence there exists at least a convergence subsequence $(u(\theta_k), x(\theta_k), y(\theta_k))$ converging, as $\theta_k \to 0$, to $(0, x^*, y^*)$, where x^* is a solution to the CP.*

(iii) *If f is continuously differentiable, then $(u(\theta), x(\theta), y(\theta))$ is also continuously differentiable in θ. In this case, the set $\{(u(\theta), x(\theta), y(\theta)) : \theta \in (0,1]\}$ forms a smooth trajectory.*

Proof. Since each P_0-function is a semimonotone function, by Theorem 4.1, system (2.3) has at least one solution. It is sufficient to show that the system has at most one solution. Let

$$g(x,a,c,\theta) = f(x) + \theta^p \mathrm{diag}\,(a^p)\,x + \theta c.$$

Since f is a P_0-function and $a \in R^n_{++}$, the function $g(x,a,c,\theta)$ is a P-function in x. Thus, by Lemma 3.3, the map

$$\mathcal{Y}(x,\theta) = x + g(x,a,c,\theta) - \sqrt{(x - g(x,a,c,\theta))^2 + 4\theta^q a^q} - \theta b$$

is a P-function in x. Since every P-function is univalent (one-to-one), the equation $\mathcal{Y}(x,\theta) = 0$ has at most one solution. Hence system (2.3) has at most one solution, by Lemma 2.1.

The continuity of $(u(\theta), x(\theta), y(\theta))$ follows easily from Lemma 3.4. Indeed, given $\hat{\theta} \in (0,1)$, in order to show the continuity of $(u(\theta), x(\theta), y(\theta))$ at $\hat{\theta}$ it is sufficient to prove the continuity of $x(\theta)$ at $\hat{\theta}$. Since $\mathcal{Y}(x,\theta)$ is a P-function in x, $x(\hat{\theta})$ is the unique element in $\mathcal{Y}^{-1}_{\hat{\theta}}(0) = \{x : \mathcal{Y}(x,\hat{\theta}) = 0\}$. By Lemma 3.4, for any $\varepsilon > 0$ there exists a scalar $\delta > 0$ such that for any P_0-function h satisfying

$$\sup_{x \in \bar{\Omega}} \|h(x) - \mathcal{Y}(x,\hat{\theta})\| < \delta, \tag{4.4}$$

where $\Omega = \mathcal{Y}^{-1}_{\hat{\theta}}(0) + \varepsilon B$, we have

$$\emptyset \neq h^{-1}(0) \subseteq \mathcal{Y}^{-1}_{\hat{\theta}}(0) + \varepsilon B = x(\hat{\theta}) + \varepsilon B. \tag{4.5}$$

For this given δ, it follows from (ii) of Lemma 3.2 that there is a scalar $\beta > 0$ such that

$$\sup_{x \in \bar{\Omega}} \|\mathcal{Y}(x,\theta) - \mathcal{Y}(x,\hat{\theta})\| < \delta$$

for all $\theta > 0$ such that $|\theta - \hat{\theta}| < \beta$. Setting $h(x) := \mathcal{Y}(x,\theta)$ in (4.4), we deduce from (4.5) that $\mathcal{Y}^{-1}_{\theta}(0) = \{x : \mathcal{Y}(x,\theta) = 0\} \subseteq x(\hat{\theta}) + \varepsilon B$ for all θ with $|\theta - \hat{\theta}| < \beta$. By the P-property of $\mathcal{Y}$, $x(\theta)$ is a unique element in $\mathcal{Y}^{-1}_{\theta}(0)$. Thus, $\|x(\theta) - x(\hat{\theta})\| < \varepsilon$ for all $\theta > 0$ such that $|\theta - \hat{\theta}| < \beta$, i.e., $x(\theta)$ is continuous at $\hat{\theta}$. Item (i) of the theorem follows.

Item (ii) follows immediately from Theorem 4.1, since P_0-functions are semimonotone. We now prove Item (iii). Consider the following $3n \times 3n$ matrix

$$A := \begin{pmatrix} I & 0 & 0 \\ -2qU^{q-1}D & I - (X-Y)D & I + (X-Y)D \\ pU^{p-1}X & -(f'(x) + \mathrm{diag}(u^p)) & I \end{pmatrix},$$

where $U = \mathrm{diag}(u), X = \mathrm{diag}(x), Y = \mathrm{diag}(y)$, and $D = \mathrm{diag}(d)$ with $d = (d_1, \ldots, d_n)^T$, where

$$d_i = 1/\sqrt{(x_i - y_i)^2 + 4u_i^q}, \quad i = 1, 2, \ldots, n.$$

If $u \in R^n_{++}$, then it is easy to see that $I - (X-Y)D$ and $I + (X-Y)D$ are positive diagonal matrices for every $(x,y) \in R^{2n}$. Thus, by Lemma 5.4 in Kojima, Megiddo, and Noma [19], the matrix

$$\begin{pmatrix} I - (X-Y)D & I + (X-Y)D \\ -(f'(x) + \mathrm{diag}(u^p)) & I \end{pmatrix}$$

is nonsingular when f is a P_0-function. Hence A is a nonsingular matrix for every $(u, x, y) \in R^n_{++} \times R^{2n}$. Since the matrix A coincides with the Jacobian matrix (with respect to (u, x, y)) of the equation

$$H(u, x, y) - \theta(a, b, c) = 0,$$

by the implicit function theorem, there is a unique smooth (i.e., continuously differentiable) curve $(u(t), x(t), y(t))$ such that

$$H(u(t), x(t), y(t)) = t(a, b, c)$$

for all t sufficiently close to θ, and

$$(u(t), x(t), y(t))|_{t=\theta} = (u(\theta), x(\theta), y(\theta)).$$

In particular, $(u(\cdot), x(\cdot), y(\cdot))$ is continuously differentiable at θ. □

Furthermore, if f is a P_*-function we can obtain a much stronger result. We now consider this important situation and show that for a P_*-function the proposed trajectory exists and is bounded, provided that the solution set of the CP is nonempty. For simplicity, we consider the case of $(a, b, c) \in R^n_{++} \times R^n_- \times R^n$, i.e., the vector b is confined to R^n_-. We also consider the case of $c \in R^n_{++}$ when it is necessary. The stipulation that $b \in R^n_-$ has also been used in some non–interior-point algorithms; see Burke and Xu [3] and Hotta, Inabar, and Yoshise [17], where the iterate $\{(x^k, y^k)\}$ is required to satisfy

$$x^k + y^k - \sqrt{(x^k - y^k)^2 + 4\mu^k} \le 0,$$

which is equivalent to the requirement of "$b \in R^n_-$."

LEMMA 4.2. *Let v^* be an arbitrary solution of the CP, and $(a, b, c) \in R^n_{++} \times R^n_- \times R^n$ be a fixed vector. Let $(u(\theta), x(\theta), y(\theta))$ satisfy the system (2.8)–(2.11) for each $\theta \in (0, 1]$. Then the following inequality holds:*

$$(x_i(\theta) - v_i^*)(f_i(x(\theta)) - f_i(v^*)) \le \theta^q e^T a^q - \theta b^T f(v^*)/2 - \theta^p \min_{1 \le i \le n} M_i,$$

where

$$M_i = a_i^p x_i(\theta)(x_i(\theta) - v_i^*) + \theta^{1-p}(c_i - b_i/2)(x_i(\theta) - v_i^*).$$

Proof. Define $\bar{y}_i(\theta) = y_i(\theta) - \theta b_i/2$ and $\bar{x}_i(\theta) = x_i(\theta) - \theta b_i/2$. Let v^* be an arbitrary solution to the CP. By (2.11) and noting that $(\bar{x}(\theta), \bar{y}(\theta)) > 0$, we have for each i,

$$\begin{aligned}(\bar{y}_i(\theta) - f_i(v^*))(\bar{x}_i(\theta) - v_i^*) &= \bar{y}_i(\theta)\bar{x}_i(\theta) - f_i(v^*)\bar{x}_i(\theta) - \bar{y}_i(\theta)v_i^* \\ &\le \bar{y}_i(\theta)\bar{x}_i(\theta) = \theta^q a_i^q. \end{aligned} \tag{4.6}$$

We also note that

$$\begin{aligned}&(x_i(\theta) - v_i^*)(f_i(x(\theta)) - f_i(v^*)) \\ &= (x_i(\theta) - v_i^*)(y_i(\theta) - (\theta a_i)^p x_i(\theta) - \theta c_i - f_i(v^*)) \\ &= (\bar{x}_i(\theta) - v_i^* + \theta b_i/2)(\bar{y}_i(\theta) + \theta b_i/2 - (\theta a_i)^p x_i(\theta) - \theta c_i - f_i(v^*)) \\ &= (\bar{x}_i(\theta) - v_i^*)(\bar{y}_i(\theta) - f_i(v^*)) + (\bar{x}_i(\theta) - v_i^*)(\theta b_i/2 - (\theta a_i)^p x_i(\theta) - \theta c_i) \\ &\quad + (\theta b_i/2)(\bar{y}_i(\theta) + \theta b_i/2 - (\theta a_i)^p x_i(\theta) - \theta c_i - f_i(v^*)).\end{aligned}$$

Thus, by (4.6) and noting that $b \le 0$ and $\bar{y}(\theta) > 0$, we have the following for all i:

$$\begin{aligned}
&(x_i(\theta) - v_i^*)(f_i(x(\theta) - f_i(v^*)) \\
&\le \theta^q a_i^q + (x_i(\theta) - v_i^* - \theta b_i/2)(\theta b_i/2 - (\theta a_i)^p x_i(\theta) - \theta c_i) \\
&\quad + (\theta b_i/2)(\theta b_i/2 - (\theta a_i)^p x_i(\theta) - \theta c_i - f_i(v^*)) \\
&= \theta^q a_i^q - (\theta a_i)^p x_i(\theta)(x_i(\theta) - v_i^*) \\
&\quad + (\theta b_i/2 - \theta c_i)(x_i(\theta) - v_i^*) - \theta b_i f_i(v^*)/2 \\
&\le \theta^q e^T a^q - \theta b^T f(v^*)/2 \\
&\quad - \theta^p \min_{1\le i\le n} [a_i^p x_i(\theta)(x_i(\theta) - v_i^*) + \theta^{1-p}(c_i - b_i/2)(x_i(\theta) - v_i^*)].
\end{aligned}$$

The last inequality follows from the fact that $a_i^q \le e^T a^q$ and $-b_i f_i(v^*) \le -b^T f(v^*)$ since $b \le 0$ and $f(v^*) \ge 0$. The proof is thus complete. □

We are ready to prove the following result.

THEOREM 4.3. *Let f be a continuous P_*-function and $(a, b, c) \in R^n_{++} \times R^n_- \times R^n$ be a fixed vector. Assume that the solution set of the CP is nonempty.*

(i) *If $p \le 1$ and $q \in [1, \infty)$, then the trajectory $\{(u(\theta), x(\theta), y(\theta)) : \theta \in (0, 1]\}$ generated by* (2.3) *is bounded.*

(ii) *If $p > 1$, $q \in [1, \infty)$, and $c \in R^n_{++}$, then the trajectory $\{(u(\theta), x(\theta), y(\theta)) : \theta \in (0, 1]\}$ generated by* (2.3) *is bounded.*

Proof. We still use the notation

$$(\bar{x}(\theta), \bar{y}(\theta)) = (x(\theta) - \theta b/2,\ y(\theta) - \theta b/2).$$

By (2.11), and noting that $b \le 0$ and $(\bar{x}(\theta), \bar{y}(\theta)) > 0$, we have

$$\begin{aligned}
x(\theta)^T y(\theta) &= (\bar{x}(\theta) + \theta b/2)^T (\bar{y}(\theta) + \theta b/2) \\
&\le \theta^q e^T a^q + \theta^2 \|b\|^2/4. \qquad (4.7)
\end{aligned}$$

Let v^* be an arbitrary solution of the CP. By the P_*-property of f and by Lemma 4.2 we have

$$\begin{aligned}
&(v^*)^T y(\theta) + f(v^*)^T x(\theta) \\
&= -(x(\theta) - v^*)^T (y(\theta) - f(v^*)) + x(\theta)^T y(\theta) \\
&= -(x(\theta) - v^*)^T (f(x(\theta)) + \theta^p \mathrm{diag}(a^p) x(\theta) + \theta c - f(v^*)) + x(\theta)^T y(\theta) \\
&= -(x(\theta) - v^*)^T (f(x(\theta)) - f(v^*)) - \theta^p [\mathrm{diag}(a^p) x(\theta)]^T (x(\theta) - v^*) \\
&\quad - \theta c^T (x(\theta) - v^*) + x(\theta)^T y(\theta) \\
&\le \tau \sum_{i \in I_+} (x_i(\theta) - v_i^*)^T (f_i(x(\theta)) - f_i(v^*)) \\
&\quad - \theta^p [\mathrm{diag}(a^p) x(\theta)]^T (x(\theta) - v^*) - \theta c^T (x(\theta) - v^*) + x(\theta)^T y(\theta) \\
&\le \tau n \max_{1\le i\le n} (x_i(\theta) - v_i^*)^T (f_i(x(\theta)) - f_i(v^*)) \\
&\quad - \theta^p [\ \mathrm{diag}(a^p) x(\theta)]^T (x(\theta) - v^*) - \theta c^T (x(\theta) - v^*) + x(\theta)^T y(\theta) \\
&\le \tau n \left(\theta^q e^T a^q - \theta b^T f(v^*)/2 - \theta^p \min_{1\le i\le n} M_i \right) \\
&\quad - \theta^p [\mathrm{diag}(a^p) x(\theta)]^T (x(\theta) - v^*) - \theta c^T (x(\theta) - v^*) + x(\theta)^T y(\theta).
\end{aligned}$$

The last inequality follows from Lemma 4.2, and M_i is given as in Lemma 4.2. By (4.7) and the above inequality, we have

$$\begin{aligned}
&(v^*)^T\bar{y}(\theta) + f(v^*)^T\bar{x}(\theta)\\
&= (v^*)^T y(\theta) + f(v^*)^T x(\theta) - \theta b^T(v^* + f(v^*))/2\\
&\le \theta^q(1+\tau n)e^T a^q + \theta^2\|b\|^2/4 - \theta\tau n b^T f(v^*)/2 - \theta^p\tau n \min_{1\le i\le n} M_i\\
&\quad -\theta^p[\ \mathrm{diag}\ (a^p)x(\theta)]^T(x(\theta)-v^*) - \theta c^T(x(\theta)-v^*) - \theta b^T(v^*+f(v^*))/2. \qquad (4.8)
\end{aligned}$$

(i) We now consider the case of $p \le 1$. Notice that the left-hand side is nonnegative. Dividing both sides of the above inequality by θ^p and rearranging terms, we have

$$\begin{aligned}
&[\mathrm{diag}(a^p)x(\theta)]^T(x(\theta)-v^*) + \theta^{1-p}c^T(x(\theta)-v^*) + \tau n \min_{1\le i\le n} M_i\\
&\le \theta^{q-p}(1+\tau n)e^T a^q + \theta^{2-p}\|b\|^2/4 - \theta^{1-p}\tau n b^T f(v^*)/2\\
&\quad -\theta^{1-p}b^T(v^*+f(v^*))/2. \qquad (4.9)
\end{aligned}$$

Since $a^p \in R^n_{++}$, $p \le 1$, and $q \in [1,\infty)$, we conclude from the above inequality that the set $\{x(\theta) : \theta \in (0,1]\}$ is bounded, and by continuity the set $\{y(\theta) : \theta \in (0,1]\}$ is also bounded. Item (i) of the theorem is proved.

(ii) If $p > 1$ and $c \in R^n_{++}$, then, since $(\bar{x}(\theta), \bar{y}(\theta)) > 0$ and $b \le 0$, we have

$$\begin{aligned}
M_i &\ge -a_i^p v_i^* x_i(\theta) + \theta^{1-p}(c_i - b_i/2)(\bar{x}_i(\theta) + \theta b_i/2 - v_i^*)\\
&\ge -a_i^p v_i^* \bar{x}_i(\theta) + \theta^{1-p}(c_i - b_i/2)(\theta b_i/2 - v_i^*)\\
&\ge -[\mathrm{diag}(a^p)\bar{x}(\theta)]^T v^* + \theta^{1-p}\min_{1\le i\le n}(c_i - b_i/2)(\theta b_i/2 - v_i^*). \qquad (4.10)
\end{aligned}$$

Since the left-hand side of (4.8) is nonnegative, by (4.10) and (4.8), we have

$$\begin{aligned}
0 &\le \theta^q(1+\tau n)e^T a^q + \theta^2\|b\|^2/4 - \theta\tau n b^T f(v^*)/2\\
&\quad + \theta^p\tau n[\mathrm{diag}(a^p)\bar{x}(\theta)]^T v^* - \theta\tau n \min_{1\le i\le n}(c_i - b_i/2)(\theta b_i/2 - v_i^*)\\
&\quad + \theta^p[\mathrm{diag}(a^p)x(\theta)]^T v^* - \theta c^T x(\theta) + \theta c^T v^* - \theta b^T(v^* + f(v^*))/2\\
&\le \theta^q(1+\tau n)e^T a^q + \theta^2\|b\|^2/4 - \theta\tau n b^T f(v^*)/2\\
&\quad + \theta^p(1+\tau n)[\mathrm{diag}(a^p)\bar{x}(\theta)]^T v^* - \theta\tau n \min_{1\le i\le n}(c_i - b_i/2)(\theta b_i/2 - v_i^*)\\
&\quad + \theta^p[\mathrm{diag}(a^p)v^*]^T b/2 - \theta c^T\bar{x}(\theta) - \theta c^T b/2 + \theta c^T v^* - \theta b^T(v^* + f(v^*))/2.
\end{aligned}$$

Dividing both sides of the above by θ and rearranging terms, we have

$$\begin{aligned}
&\left(c - \theta^{p-1}(1+\tau n)\mathrm{diag}(a^p)v^*\right)^T\bar{x}(\theta)\\
&\le \theta^{q-1}(1+\tau n)e^T a^q + \theta\|b\|^2/4 - \tau n b^T f(v^*)/2\\
&\quad - \tau n \min_{1\le i\le n}(c_i - b_i/2)(\theta b_i/2 - v_i^*)\\
&\quad + \theta^{p-1}[\mathrm{diag}\ (a^p)v^*]^T b/2 - c^T b/2 + c^T v^* - b^T(v^* + f(v^*))/2.
\end{aligned}$$

Since $p > 1$ and $c \in R^n_{++}$, there must exist a $\delta \in (0,1)$ such that for all $\theta \in (0,\delta]$ we have that $c - \theta^{p-1}(1+\tau n)\mathrm{diag}(a^p)v^* \ge c/2 > 0$. Thus we can see from the above inequality that the set $\{\bar{x}(\theta) : \theta \in (0,\delta]\}$ is bounded. Thus, the set $\{x(\theta) : \theta \in (0,\delta]\}$

is bounded. The boundedness of the set $\{x(\theta) : \theta \in [\delta, 1]\}$ can be obtained by (4.8) again. Indeed, if a subsequence in $\{x(\theta) : \theta \in [\delta, 1]\}$ is unbounded, then there exists a subsequence denoted by $\{x(\theta_k)\}$ such that $\|x(\theta_k)\| \to \infty$ as $k \to \infty$, where $\theta_k \in [\delta, 1]$. Applying (4.8) to this sequence, the left-hand side of it is nonnegative. The right-hand side of the inequality (4.8), however, tends to $-\infty$. This is a contradiction. Therefore, we conclude that the entire set $\{x(\theta) : \theta \in (0, 1]\}$ is bounded. So is $\{y(\theta) : \theta \in (0, 1]\}$, by continuity. □

Remark 4.1. The above result shows that the nonemptiness of the solution set implies the boundedness of the entire trajectory $\{(x(\theta), y(\theta)) : \theta \in (0, 1]\}$. Notice that the boundedness of this trajectory in turn implies the nonemptiness of the solution set. Therefore, we may conclude that the boundedness of this trajectory is equivalent to the solvability of the problem.

5. Limiting behavior of the trajectory. We have shown that Condition 3.1 (and hence most of the known conditions used in interior-point and non–interior-point methods) can guarantee the boundedness of the proposed continuation trajectory. Thus, there exists at least one convergent subsequence $\{(u(\theta_k), x(\theta_k), y(\theta_k))\}$ whose limiting point is a solution to the CP. Two natural questions arise: (i) When is the entire trajectory convergent? (ii) What can be said about the limiting point of it? This section is devoted to these questions. For $0 < p < 1$ and $(a, b, c) \in R^n_{++} \times R^n_{-} \times R^n$, or $0 < p \leq q$ and $(a, b, c) \in R^n_{++} \times \{0\} \times \{0\}$, we show (in Theorem 5.1) that if f is a P_*-function and the CP has a least element solution, then the entire trajectory $\{(u(\theta), x(\theta), y(\theta))\}$ generated by (2.3) converges, as $\theta \to 0$, to the unique least element solution, and that if f is monotone, then the entire trajectory is convergent as $\theta \to 0$, and the limiting point is the N-norm least solution, where $N = \mathrm{diag}(a^p)$. For $p > q$ and $(a, b, c) \in R^n_{++} \times \{0\} \times \{0\}$, we show, among other things, that any limiting point of the sequence $\{(u(\theta_k), x(\theta_k), y(\theta_k))\}$ as $\theta \to 0$ is a maximal complementarity solution (see Theorem 5.2).

To begin, we review some concepts that will be used in this section. An element x^* of the set S is said to be the N-norm least element, where N is a positive definite matrix, if $\|N^{1/2}x^*\| \leq \|N^{1/2}u\|$ for all $u \in S$. In particular, if $N = I$, the solution x^* is called the least 2-norm element of S. An element x^* of the set S is said to be a least element of S if $x^* \leq u$ for all $u \in S$ (Pang [24]). An element x^* of the set S is said to be a weak Pareto minimal element if there is no element u in S such that $u < x^*$ (Sznajder and Gowda [29]). It is evident that the (unique) least element is a weak Pareto minimal element, but the converse is not true. If the solution set $\mathrm{SOL}_{cp}(f)$ is convex, it is known that there exists a unique partition of the index set $\{1, \ldots, n\}$ denoted by I, J, and O such that $\{1, \ldots, n\} = I \cup J \cup O$, and the intersection of each pair of them is empty. In fact,

$$I = \{i : x^*_i > 0 \ \text{ for some } x^* \in \mathrm{SOL}_{cp}(f)\},$$

$$J = \{j : f_j(x^*) > 0 \ \text{ for some } x^* \in \mathrm{SOL}_{cp}(f)\},$$

$$O = \{k : x^*_k = f_k(x^*) = 0 \ \text{ for all } x^* \in \mathrm{SOL}_{cp}(f)\}.$$

Since the solution set is convex, there must exist a solution x^* satisfying $x^*_i > 0$ for all $i \in I$, $f_i(x^*) > 0$ for all $i \in J$, and $x^*_i = f_i(x^*) = 0$ for all $i \in O$. Such a solution is called a maximal complementarity solution. When $O = \emptyset$, i.e., $x^* + f(x^*) > 0$, x^* is called a strict complementarity solution. We now prove the following result.

THEOREM 5.1. *Assume that the solution set $SOL_{cp}(f)$ is nonempty. Let p, q, and (a, b, c) satisfy one of the following conditions:*

(C1) $0 < p < 1$, $q \in [1, \infty)$, *and* $(a, b, c) \in R^n_{++} \times R^n_- \times R^n$.
(C2) $0 < p < q$, $q \in [1, \infty)$, *and* $(a, b, c) \in R^n_{++} \times \{0\} \times \{0\}$.

Then the following results hold:

(i) *If f is a continuous P_*-function and the least element solution of the CP exists, then the entire trajectory $\{x(\theta) : \theta \in (0, 1]\}$ generated by (2.3) converges, as $\theta \to 0$, to the unique least element solution.*

(ii) *If f is a continuous monotone mapping, then the entire continuation trajectory $\{x(\theta) : \theta \in (0, 1]\}$ generated by system (2.3) converges, as $\theta \to 0$, to a solution of the CP. This solution, denoted by x^*, is an N-norm least solution, i.e.,*

$$\|N^{1/2}x^*\| \le \|N^{1/2}v^*\| \quad \text{for all } v^* \in SOL_{cp}(f),$$

where $N = \mathrm{diag}(a^p)$. In particular, if $a = \alpha e$, where $\alpha > 0$ is a positive scalar, then this solution is the (unique) least two-norm solution.

Proof. We show first that the result holds under condition (C1). Let v^* be an arbitrary solution of the CP. Then (4.9) holds. By (i) of Theorem 4.3 the entire continuation trajectory $\{(u(\theta), x(\theta), y(\theta))\}$ is bounded, provided that the solution set of the CP is nonempty. Let x^* be an arbitrary accumulation point of $\{x(\theta)\}$ as $\theta \to 0$. Since $0 < p < 1$ and $q \in [1, \infty)$, letting $\theta \to 0$ in (4.9), we have

$$\tau n \min_{1 \le i \le n} a_i^p x_i^*(x_i^* - v_i^*) + [\mathrm{diag}(a^p)(x^*)]^T(x^* - v^*) \le 0. \tag{5.1}$$

Note that v^* is an arbitrary solution of the CP. If the problem has a least element solution u^*, setting $v^* = u^*$ in the above inequality, we deduce that $x^* = u^*$. Since the least element solution is unique, we conclude that the entire trajectory converges to the solution.

Since each monotone map is a P_*-function with the constant $\tau = 0$, the inequality (5.1), in this case, reduces to

$$[\mathrm{diag}(a^p)(x^*)]^T(x^* - v^*) \le 0, \tag{5.2}$$

where v^* is an arbitrary solution of the CP. To show the convergence of the entire trajectory, it is sufficient to show that x^* is unique. Indeed, if there exists another vector u^* such that u^* is also an accumulation point to the trajectory, then we have

$$[\mathrm{diag}(a^p)(u^*)]^T(u^* - v^*) \le 0 \tag{5.3}$$

for all solutions v^*. Since x^* and u^* are solutions to the CP, setting $v^* = u^*$ in (5.2) and $v^* = x^*$ in (5.3), and adding the two inequalities, we obtain

$$(x^* - u^*)^T[\mathrm{diag}(a^p)](x^* - u^*) \le 0.$$

Since $a^p \in R^n_{++}$, it follows from the above that $x^* = u^*$. Hence the trajectory is convergent because it has a unique limiting point. It follows from (5.2) that

$$\|N^{1/2}x^*\|^2 \le \|N^{1/2}x^*\|\|N^{1/2}v^*\|,$$

where $N = \mathrm{diag}(u^p)$ and v^* is an arbitrary solution of the CP. Therefore, $\|N^{1/2}x^*\| \le \|N^{1/2}v^*\|$ for any solution v^*, i.e., x^* is a least N-norm solution. In particular, if $a = \alpha e$ for some positive scalar $\alpha > 0$, then (5.2) reduces to $(x^*)^T(x^* - v^*) \le 0$ for all solutions v^*, which implies that x^* is the unique least 2-norm solution.

We now show the result under (C2). It is evident that, under condition (C2), the inequality (4.8) can be written as:

$$\begin{aligned}(v^*)^T\bar{y}(\theta) + f(v^*)^T\bar{x}(\theta) &\\ \leq \theta^q(1+\tau n)e^T a^q - \theta^p \tau n \min_{1\leq i\leq n} a_i^p x_i(\theta)(x_i(\theta) - x_i^*) &\\ - \theta^p[\mathrm{diag}(a^p)x(\theta)]^T(x(\theta) - v^*).&\end{aligned}$$

Dividing both sides of the above inequality by θ^p, and noting that the left-hand side is nonnegative, we have

$$\begin{aligned}0 \leq \theta^{q-p}(1+\tau n)e^T a^q - \tau n \min_{1\leq n} a_i^p x_i(\theta)(x_i(\theta) - v_i^*)&\\ - [\mathrm{diag}(a^p)x(\theta)]^T(x(\theta) - v^*).&\end{aligned}$$

Since $p < q$, the above inequality implies that $\{x(\theta) : \theta \in (0,1]\}$ is bounded, and thus so is $\{y(\theta) : \theta \in (0,1)\}$, by continuity. Let x^* be an arbitrary accumulation point of $\{x(\theta)\}$ as $\theta \to 0$. Letting $\theta \to 0$ in the above inequality, we obtain inequality (5.1) again. It suffices to repeat the proof of (C1). □

The above result states that when $p < q$ the trajectory of the monotone CP converges to an N-norm least solution. The next result studies the case of $p \geq q$.

THEOREM 5.2. *Assume that f is a monotone function.*

(i) *Let $p \geq q$ and $(a,b,c) \in R^n_{++} \times \{0\} \times \{0\}$. If the trajectory $\{x(\theta) : \theta \in (0,1]\}$ generated by (2.3) has an accumulation point as $\theta \to 0$, then any accumulation point of the trajectory, as $\theta \to 0$, is a maximal complementarity solution of the CP.*

(ii) *Let $p > q$ and $(a,b,c) \in R^n_{++} \times \{0\} \times \{0\}$. Assume that the CP has a strict complementarity solution and the trajectory $\{x(\theta) : \theta \in (0,1]\}$ generated by (2.3) has an accumulation point as $\theta \to 0$. Then any accumulation point $(\hat{x}, \hat{y})$ of the trajectory as $\theta \to 0$ is a maximal strict complementarity solution, in the sense that*

$$\sum_{i\in I} a_i^q \log v_i^* + \sum_{j\in J} a_j^q \log f_j(v^*) \leq \sum_{i\in I} a_i^q \log \hat{x}_i + \sum_{j\in J} a_j^q \log f_j(\hat{x}),$$

where v^* is an arbitrary strict complementarity solution. Furthermore, if f is linear, i.e., $f = Mx + u$, where M is an n by n positive semidefinite matrix and $u \in R^n$ is a vector, then the entire trajectory converges, as $\theta \to 0$, to a unique maximal strict complementarity solution.

Proof. Since each accumulation point of the trajectory, as $\theta \to 0$, is a solution to the CP, then under the assumption of the theorem the solution set of the CP is nonempty. Let v^* be an arbitrary solution to the CP. Thus, the inequality (4.8) remains valid. By assumption, we have $\tau = 0$ and $b = c = 0$. Therefore, (4.8) reduces to

$$\begin{aligned}(v^*)^T\bar{y}(\theta) + f(v^*)^T\bar{x}(\theta) &\leq \theta^q e^T a^q - \theta^p \left[\mathrm{diag}(a^p)x(\theta)\right]^T (x(\theta) - v^*)\\ &\leq \theta^q e^T a^q + \theta^p \left[\mathrm{diag}(a^p)x(\theta)\right]^T v^*.\end{aligned}$$

Notice that the solution set of a monotone CP is convex. Let I, J, O be the unique partition of the indexes $\{1, 2, \ldots, n\}$ as defined at the beginning of this section. Then the above inequality further reduces to

$$(v^*)_I^T \bar{y}_I(\theta) + f_J(v^*)^T \bar{x}_J(\theta) \leq \theta^q e^T a^q + \theta^p \left[\mathrm{diag}(a_I^p)x_I(\theta)\right]^T v_I^*.$$

Since $(\bar{x}(\theta), \bar{y}(\theta)) \in R^{2n}_{++}$, we have

$$\begin{aligned}&(v^*)_I^T \bar{X}_I^{-1}(\theta)\bar{X}_I(\theta)\bar{y}_I(\theta) + f_J(v^*)^T \bar{Y}_J^{-1}(\theta)\bar{Y}_J(\theta)\bar{x}_J(\theta)\\ &\le \theta^q e^T a^q + \theta^p \left[\mathrm{diag}(a_I^p)x_I(\theta)\right]^T v_I^*,\end{aligned}$$

where $\bar{X}_I(\theta) = \mathrm{diag}(\bar{x}_I(\theta))$, and $\bar{Y}_J(\theta) = \mathrm{diag}(\bar{y}_J(\theta))$. Since

$$\bar{X}_I(\theta)\bar{y}_I(\theta) = \theta^q a_I^q, \quad \bar{X}_J(\theta)\bar{y}_J(\theta) = \theta^q a_J^q,$$

from the above inequality we have

$$(5.4)\quad (v^*)_I^T \bar{X}_I^{-1}(\theta)a_I^q + f_J(v^*)^T \bar{Y}_J^{-1}(\theta)a_J^q \le e^T a^q + \theta^{p-q}\left[\mathrm{diag}(a_I^p)x_I(\theta)\right]^T v_I^*.$$

The above inequality holds for all solutions v^*. In particular, let v^* be a solution satisfying $v_I^* > 0$ and $f_J(v^*) > 0$, i.e., let v^* be a maximal complementarity solution. Assume that $(\hat{x}, \hat{y})$ is an accumulation point of $(x(\theta), y(\theta))$ as $\theta \to 0$. Taking $\theta \to 0$ in the above inequality we deduce that $\hat{x}_I > 0$ and $\hat{y}_J > 0$ since $p \ge q$. Notice that $(\hat{x}, \hat{y})$ is a solution of the CP; $(\hat{x}, \hat{y})$ is a maximal complementarity solution. Theorem 5.2(i) follows.

We now prove Theorem 5.2(ii). Let v^* in (5.4) be an arbitrary strict complementarity solution. Assume that $(\hat{x}, \hat{y})$ is an arbitrary accumulation point of $\{x(\theta), y(\theta)\}$ as $\theta \to 0$. We now prove that $(\hat{x}, \hat{y})$ is a strict complementarity solution. Taking the limit in (5.4) and noting that $p > q$, we deduce that $\hat{x}_I > 0$, $\hat{y}_J > 0$, and

$$(v^*)_I^T \hat{X}_I^{-1}(\theta)a_I^q + f_J(v^*)^T atY_I^{-1}(\theta)a_J^q \le e^T a^q.$$

Notice that $\hat{y} = f(\hat{x})$. The above inequality can be further written as

$$\sum_{i\in I} a_i^q (v_i^*/\hat{x}_i) + \sum_{j\in J} a_j^q (f_j(v^*)/f_j(\hat{y})) \le e^T a^q.$$

Since $1 + \log t \le t$ for all $t > 0$, from the above we have

$$\sum_{i\in I} a_i^q [1 + \log(v_i^*/\hat{x}_i)] + \sum_{j\in J} a_j^q [1 + \log(f_j(v^*)/f_j(\hat{y}))] \le e^T a^q.$$

Since a strict complementarity solution exists, we have $I \cup J = \{1, \ldots, n\}$, and thus

$$\sum_{i\in I} a_i^q \log(v_i^*/\hat{x}_i) + \sum_{j\in J} a_j^q \log(f_j(v^*)/f_j(\hat{y})) \le 0.$$

Therefore,

$$(5.5)\qquad \sum_{i\in I} a_i^q \log v_i^* + \sum_{j\in J} a_j^q \log f_j(v^*) \le \sum_{i\in I} a_i^q \log \hat{x}_i + \sum_{j\in J} a_j^q \log f_j(\hat{y}).$$

Since v^* is an arbitrary strict complementarity solution, the first part of Theorem 5.2(ii) is proved.

We now consider the linear case, i.e., $f = Mx + u$. Denote by $\mathrm{SSOL}(f)$ the set of strict complementarity solutions of the CP, which is also a convex set by the convexity of $\mathrm{SOL}_{cp}(f)$. Since f is linear, this fact in turn implies that the following set is also convex:

$$\overline{S} = \{(x, y) : y = Mx + u,\ x \in \mathrm{SSOL}(f)\}.$$

To show the second part of result (ii), it is sufficient to prove that the accumulation point $(\hat{x}, \hat{y})$ satisfying (5.5) is unique. In fact, it is easy to see that $(\hat{x}, \hat{y})$ is the solution to the following strict concave program:

$$\begin{aligned} &\text{Maximize } \sum_{i\in I} a_i^q \log x_i + \sum_{i\in J} a_j^q \log y_i \\ &\text{subject to } (x, y) \in \overline{S}. \end{aligned}$$

Since a strict concave program has at most one solution, $(\hat{x}, \hat{y})$ is the unique solution to the above program, which is a maximal strict complementarity solution of the CP. Thus, the entire trajectory is convergent. □

We close this section by proving a general result concerning the characterization of the limiting point of the trajectory proposed in this paper in the case of semimonotone functions.

THEOREM 5.3. *Let f be a continuous semimonotone function from R^n into R^n. Let p, q, and (a, b, c) satisfy one of the following conditions:*

(C1) $0 < p < 1, q \in [1, \infty)$, *and* $(a, b, c) \in R^n_{++} \times R^{2n}$.

(C2) $0 < p < q, q \in [1, \infty)$, *and* $(a, b, c) \in R^n_{++} \times \{0\} \times \{0\}$.

Let $(u(\theta), x(\theta), y(\theta))$ be a solution to system (2.3) for each $\theta \in (0, 1]$. Assume that there exists an accumulation point to the trajectory $(u(\theta), x(\theta), y(\theta))$ as $\theta \to 0$. Then for any accumulation point $(0, x^*, y^*)$ of this trajectory as $\theta \to 0$, x^* is a weak Pareto minimal solution to the CP.

Proof. Let $(0, x^*, y^*)$ be an arbitrary accumulation point of $(u(\theta), x(\theta), y(\theta))$ as $\theta \to 0$. Then there exists a subsequence $\{\theta_k\} \to 0$ such that

$$\{(u(\theta_k), x(\theta_k), y(\theta_k))\} \to (0, x^*, y^*).$$

Assume the contrary: that x^* is not a weak Pareto minimal solution. Then there exists a solution u^* satisfying $u^* < x^*$. Since $x(\theta_k) \to x^*$, we have $x(\theta_k) > u^*$ for all sufficiently large k. By the semimonotone property of f, for each sufficiently large k there is an index i_k such that

$$x_{i_k}(\theta_k) > u^*_{i_k} \quad \text{and} \quad f_{i_k}(x(\theta_k)) \geq f_{i_k}(u^*).$$

Passing through a subsequence, we may assume that there exists an index l such that

$$x_l(\theta_k) > u^*_l \quad \text{and} \quad f_l(x(\theta_k)) \geq f_l(u^*)$$

for all sufficiently large k. Notice that for each θ, the solution $(u(\theta), x(\theta), y(\theta))$ of system (9) satisfies the system (2.8)–(2.11). We still use the symbols $\bar{y}_i(\theta) = y_i(\theta) - \theta b_i/2 > 0$ and $\bar{x}_i(\theta) = x_i(\theta) - \theta b_i/2 > 0$. By (4.6) we have

$$(\bar{y}_l(\theta_k) - f_l(u^*))(\bar{x}_l(\theta_k) - u^*_l) \leq \theta_k^q a_l^q. \tag{5.6}$$

On the other hand, we have

$$\begin{aligned} &(\bar{y}_l(\theta_k) - f_l(u^*))(\bar{x}_l(\theta_k) - u^*_l) \\ &= (\bar{y}_l(\theta_k) - f_l(u^*))(-\theta_k b_l/2) + (\bar{y}_l(\theta_k) - f_l(u^*))(x_l(\theta_k) - u^*_l) \\ &= (\bar{y}_l(\theta_k) - f_l(u^*))(-\theta_k b_l/2) \\ &\quad + (f_l(x(\theta_k)) + \theta_k^p a_l^p x_l(\theta_k) + \theta_k c_l - \theta_k b_l/2 - f_l(u^*))(x_l(\theta_k) - u^*_l) \\ &= (y_l(\theta_k) - \theta_k b_l/2 - f_l(u^*))(-\theta_k b_l/2) \end{aligned}$$

$$
\begin{aligned}
&+ (f_l(x(\theta_k)) - f_l(u^*))(x_l(\theta_k) - u_l^*) \\
&+ (\theta_k^p a_l^p x_l(\theta_k) + \theta_k c_l - \theta_k b_l/2)\,(x_l(\theta_k) - u_l^*) \\
\geq\ & (y_l(\theta_k) - \theta_k b_l/2 - f_l(u^*))(-\theta_k b_l/2) \\
&+ (\theta_k^p a_l^p x_l(\theta_k) + \theta_k c_l - \theta_k b_l/2)\,(x_l(\theta_k) - u_l^*).
\end{aligned}
$$

Combining inequality (5.6) and the inequality above and dividing both sides by θ_k^p, we have

$$
\begin{aligned}
\theta_k^{q-p} a_l^2 \geq\ & (y_l(\theta_k) - \theta_k b_l/2 - f_l(u^*))(-\theta_k^{1-p} b_l/2) \\
&+ (a_l^p x_l(\theta_k) + \theta_k^{1-p} c_l - \theta_k^{1-p} b_l/2)(x_l(\theta_k) - u_l^*).
\end{aligned}
$$

Let $\theta_k \to 0$. It is easy to see that under either condition (C1) or (C2) we have

$$
a_l^p x_l^*(x_l^* - u_l^*) \leq 0,
$$

which contradicts the assumption $0 \leq u^* < x^*$. □

6. Final remarks. We have proved the existence and the boundedness of a new homotopy continuation trajectory for nonlinear P_0-complementarity problems. The assumption imposed in the paper is weaker than most existing conditions widely used in interior-point and non–interior-point methods. Particularly, this assumption is satisfied if the P_0-CP has a nonempty and bounded solution set. Therefore, the method proposed in this paper can tackle all P_0-CPs with bounded solution sets. Since this assumption can be satisfied even when the strict feasibility condition fails to hold, the proposed method can also be used to tackle some P_0-CPs with unbounded solution sets. For P_*-CPs, the existence and the boundedness of the new continuation trajectory can be guaranteed, provided that the solution set of the CP is nonempty (whether the solution set is bounded or not). Moreover, under some choices of p, q, and (a, b, c), the entire trajectory for any continuous monotone CP always converges to a solution of the CP, provided that a solution exists. Based on the results of this paper, we may design a non–interior-point path-following algorithm for CPs that can solve all solvable P_*-CPs, all P_0-CPs with bounded solution sets, and some P_0-CPs with unbounded solution sets.

Since the proposed method can deal with all solvable P_*-CPs, a natural question is whether this method can attack all solvable P_0-CPs. That is, can Condition 3.1 in Theorem 4.2 be replaced by the nonemptiness assumption of the solution set? The answer is no. For some P_0-CPs with unbounded solution sets, the proposed continuation trajectory might be divergent to infinity. This phenomena has also been seen in the canonical Tikhonov regularization trajectory for a P_0-CP with an unbounded solution set (Sznajder and Gowda [29]). In fact, in section 3.4 of [20], Kojima et al. pointed out that a certain knapsack problem can be transformed into a linear P_0-CP with an unbounded solution set. For this P_0-CP, it is easy to verify that both the canonical Tikhonov regularization trajectory and the continuation trajectory proposed in this paper are divergent to infinity. This implies that the example in section 3.4 of [20] does not satisfy Condition 3.1.

REFERENCES

[1] J. V. Burke and S. Xu, *The global linear convergence of a non-interior path-following algorithm for linear complementarity problems*, Math. Oper. Res., 23 (1998), pp. 719–734.

[2] J. V. Burke and S. Xu, *A non-interior predictor-corrector path-following method for LCP*, in Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, Appl. Optim. 22, Kluwer, Dordrecht, The Netherlands, 1999, pp. 25–63.

[3] J. V. Burke and S. Xu, *A non-interior predictor-corrector path following algorithm for the monotone linear complementarity problem*, Math. Program., 87 (2000), pp. 113–130.

[4] J. V. Burke and S. Xu, *The Complexity of A Non-Interior-Path Following Method for the Linear Complementarity Problem*, Technical report, Department of Mathematics, University of Washington, Seattle, WA, 1999.

[5] B. Chen and P. T. Harker, *A non-interior-point continuation method for linear complementarity problems*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1168–1190.

[6] B. Chen and X. Chen, *A global and local superlinear continuation-smoothing method for P_0 and R_0 NCP or monotone NCP*, SIAM J. Optim., 9 (1999), pp. 624–645.

[7] B. Chen, X. Chen, and C. Kanzow, *A penalized Fischer–Burmeister NCP-function: Theoretical investigation and numerical results*, Math. Program., 88 (2000), pp. 211–216.

[8] R. W. Cottle, J. S. Pang, and R. E. Stone, *The Linear Complementarity Problem*, Academic Press, Boston, 1992.

[9] R. W. Cottle, J. S. Pang, and V. Venkateswaran, *Sufficient matrices and the linear complementarity problems*, Linear Algebra Appl., 114/115 (1989), pp. 231–249.

[10] F. Facchinei, *Structural and stability properties of P_0 nonlinear complementarity problems*, Math. Oper. Res., 23 (1998), pp. 735–745.

[11] F. Facchinei and C. Kanzow, *Beyond monotonicity in regularization methods for nonlinear complementarity problems*, SIAM J. Control Optim., 37 (1999), pp. 1150–1161.

[12] F. Facchinei and J. S. Pang, *Total Stability of Variational Inequalities*, Technical report, Dipartimento di Informatica e Sistemistica, Università di Roma, Roma, Italy, 1998.

[13] M. S. Gowda and M. A. Tawhid, *Existence and limiting behavior of trajectories associated with P_0-equations*, Comput. Optim. Appl., 12 (1999), pp. 229–251.

[14] P. T. Harker and J. S. Pang, *Finite-dimensional variational inequality and nonlinear complementarity problems: A survey of theory, algorithms, and applications*, Math. Programming Ser. B, 48 (1990), pp. 161–220.

[15] W. P. M. H. Heemels, J. M. Schumacher, and S. Weiland, *Linear complementarity systems*, SIAM J. Appl. Math., 60 (2000), pp. 1234–1269.

[16] K. Hotta and A. Yoshise, *Global convergence of a class of non-interior-point algorithms using Chen–Harker–Kanzow functions for nonlinear complementarity problems*, Math. Program., 86 (1999), pp. 105–133.

[17] K. Hotta, M. Inaba, and A. Yoshise, *A complexity analysis of a smoothing method using CHKS-function for monotone linear complementarity problems*, Comput. Optim. Appl., 17 (2000), 183–201.

[18] C. Kanzow, *Some noninterior continuation methods for linear complementarity problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 851–868.

[19] M. Kojima, N. Megiddo, and T. Noma, *Homotopy continuation methods for nonlinear complementarity problems*, Math. Oper. Res., 16 (1991), pp. 754–774.

[20] M. Kojima, N. Megiddo, T. Noma, and A. Yoshise, *A Unified Approach to Interior Point Algorithms for Linear Complementarity Problems*, Lecture Notes in Comput. Sci. 538, Springer-Verlag, New York, 1991.

[21] M. Kojima, M. Mizuno, and T. Noma, *Limiting behavior of trajectories generated by a continuation method for monotone complementarity problems*, Math. Oper. Res., 43 (1990), pp. 662–675.

[22] L. Lloyd, *Degree Theory*, Cambridge University Press, Cambridge, UK, 1978.

[23] P. Lötstedt, *Mechanical systems of rigid bodies subject to unilateral constraints*, SIAM J. Appl. Math., 42 (1982), pp. 281–296.

[24] J. S. Pang, *Least Element Complementarity Theory*, Ph.D. thesis, Department of Operations Research, Stanford University, Stanford, CA, 1976.

[25] H. D. Qi, *Tikhonov regularization methods for variational inequality problems*, J. Optim. Theory Appl., 102 (1999), pp. 193–201.

[26] L. Qi and D. Sun, *Improving the convergence of non-interior point algorithm for nonlinear complementarity problems*, Math. Comp., 69 (2000), pp. 283–304.

[27] G. Ravindran and M. S. Gowda, *Regularization of P_0-functions in box variational inequality problems*, SIAM J. Optim., 11 (2000), pp. 748–760.

[28] D. Sun, *A regularization Newton method for solving nonlinear complementarity problems*, Appl. Math. Optim., 40 (1999), pp. 315–339.

[29] R. Sznajder and M. S. Gowda, *On the limiting behavior of the trajectory of regularized solutions of P_0 complementarity problems*, in Reformulation: Nonsmooth, Piecewise Smooth,

Semismooth and Smoothing Methods, Appl. Optim. 22, Kluwer, Dordrecht, The Netherlands, 1999, pp. 371–379.

[30] P. Tseng, *Growth behavior of a class of merit functions for the nonlinear complementarity problems*, J. Optim. Theory Appl., 89 (1996), pp. 17–37.

[31] P. Tseng, *Error bounds for regularized complementarity problems*, in Ill-Posed Variational Problems and Regularization Techniques, Lecture Notes in Econom. and Math. Systems 477, Springer-Verlag, Berlin, 1998, pp. 247–274.

[32] P. Tseng, *Analysis of a non-interior continuation method based on Chen–Mangasarian smoothing functions for complementarity problems*, in Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, Appl. Optim. 22, Kluwer, Dordrecht, The Netherlands, 1999, pp. 381–404.

[33] H. Väliaho, *P_* matrices are just sufficient*, Linear Algebra Appl., 239 (1996), pp. 103–108.

[34] V. Venkateswaran, *An algorithm for the linear complementarity problem with a P_0-matrix*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 967–977.

[35] S. Xu, *The Global Linear Convergence and Complexity of a Non-Interior Path-Following Algorithm for Monotone LCP Based on Chen–Harker–Kanzow–Smale Smoothing Function*, Technical report, Department of Mathematics, University of Washington, Seattle, WA, 1997.

[36] S. Xu and J. V. Burke, *A polynomial time interior-point path-following algorithm for LCP based on Chen-Harker-Kanzow smoothing techniques*, Math. Program., 86 (1999), pp. 91–103.

[37] A. J. van der Schaft and J. M. Schumacher, *Complementarity modeling of hybrid systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 483–490.

[38] Y. B. Zhao and J. Han, *Two interior-point methods for nonlinear $P_*(\tau)$-complementarity problems*, J. Optim. Theory Appl., 102 (1999), pp. 659–679.

[39] Y. B. Zhao and G. Isac, *Quasi-P_*-maps, $P(\tau,\alpha,\beta)$-maps, exceptional family of elements and complementarity problems*, J. Optim. Theory Appl., 105 (2000), pp. 213–231.

[40] Y.-B. Zhao and G. Isac, *Properties of a multivalued mapping associated with some nonmonotone complementarity problems*, SIAM J. Control Optim., 39 (2000), pp. 571–593.

[41] Y. B. Zhao and D. Li, *Strict feasibility conditions in complementarity problems*, J. Optim. Theory Appl., 107 (2000), pp. 641–664.

[42] Y. B. Zhao and D. Li, *On a new homotopy continuation trajectory for nonlinear complementarity problems*, Math. Oper. Res., 26 (2001), pp. 119–146.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 3, pp. 925–946

SECOND ORDER METHODS FOR OPTIMAL CONTROL OF TIME-DEPENDENT FLUID FLOW*

MICHAEL HINZE† AND KARL KUNISCH‡

Abstract. Second order methods for open loop optimal control problems governed by the two-dimensional instationary Navier–Stokes equations are investigated. Optimality systems based on a Lagrangian formulation and adjoint equations are derived. The Newton and quasi-Newton methods as well as various variants of SQP methods are developed for applications to optimal flow control, and their complexity in terms of system solves is discussed. Local convergence and rate of convergence are proved. A numerical example illustrates the feasibility of solving optimal control problems for two-dimensional instationary Navier–Stokes equations by second order numerical methods in a standard workstation environment.

Key words. optimal control Navier–Stokes equations, Newton method, SQP method, second order sufficient optimality

AMS subject classifications. 49J20, 49M37, 90C30, 35Q30

PII. S0363012999361810

1. Introduction.

This research is devoted to the analysis of second methods for solving optimal control problems involving the time-dependent Navier–Stokes equations. Thus we consider

$$\min\ J(y,u) \text{ over } (y,u) \tag{1.1}$$

subject to

$$\begin{cases} \frac{\partial y}{\partial t} + (y\cdot\nabla)y - \nu\Delta y + \nabla p = Bu & \text{in } Q = (0,T)\times\Omega, \\ \operatorname{div} y = 0 & \text{in } Q, \\ y(t,\cdot) = 0 & \text{on } \Sigma = (0,T)\times\partial\Omega, \\ y(0,\cdot) = y_0 & \text{in } \Omega. \end{cases} \tag{1.2}$$

Here Ω is a bounded domain in $\mathbb{R}^2$ with sufficiently smooth boundary $\partial\Omega$. The final time $T > 0$ and the initial condition y_0 are fixed. The vector valued variable y and the scalar valued variable p represent the velocity and the pressure of the fluid. Further, u denotes the control variable and B the control operator. The precise functional analytic setting of problem (1.1), (1.2) will be given in section 2. For the moment it suffices to say that typical cost functionals include tracking-type functionals

$$J(y,u) = \frac{1}{2}\int_Q |y-z|^2 dx\,dt + \frac{\alpha}{2}\,|u|^2 \tag{1.3}$$

*Received by the editors September 24, 1999; accepted for publication (in revised form) May 3, 2001; published electronically October 31, 2001.
http://www.siam.org/journals/sicon/40-3/36181.html

†Fakultät für Mathematik und Naturwissenschaften, Institut für Numerische Mathematik, Technische Universität Dresden, Zellescher Weg 12-14, D-01069 Dresden, Germany (hinze@math.tu-dresden.de). The research of this author was supported by a Marie-Curie Fellowship of the European Union and by the German Research Foundation under Sfb 557 "Beeinflussung komplexer turbulenter Scherströmungen."

‡Institut für Mathematik, Karl-Franzens-Universität Graz A-8010 Graz, Austria (karl.kunisch@kfunigraz.ac.at). The research of this author was supported in part by the Fonds zur Förderung der wissenschaftlichen Forschung under SFB 03 "Optimierung und Kontrolle."

and functionals involving the vorticity of the fluid

$$J(y,u) = \frac{1}{2}\int_Q |\text{curl } y(t,\cdot)|^2_{\mathbb{R}^n}\, dx\, dt + \frac{\alpha}{2}\,|u|^2, \tag{1.4}$$

where $\alpha > 0$ and z are given. For the following discussion it will be convenient to formally represent all the equality constraints involved in (1.2) by $\hat{e}(y,p,u) = 0$ so that (1.1), (1.2) can be expressed in the form

$$(P) \qquad \begin{cases} \min\ J(y,u) \text{ over } (y,u) \\ \text{subject to} \\ \hat{e}(y,p,u) = 0. \end{cases}$$

In this form solving (1.1), (1.2) appears at first to be a standard task; see [AM, DHV, G, GT, IK, KS2, NT] and the references given there. However, the formidable size of (1.1), (1.2) and the goal of analyzing second order methods necessitate an independent analysis.

For second order methods applied to optimal control problems two classes can be distinguished depending on whether (y,p) in (1.1), (1.2) are considered as independent variables or as functions of the control variable u. In the former case $\hat{e}(y,p,u) = 0$ represents an explicit constraint for the optimization problem, whereas in the latter case $\hat{e}(y(u),p(u),u) = 0$ serves the purpose of describing the evaluation of (y,p) as a function of u. In fact, (P) can be expressed as the reduced problem

$$(\hat{P}) \qquad \min\ \hat{J}(u) = J(y(u),u) \text{ over } u, \tag{$\hat{P}$}$$

where $y(u)$ is implicitly defined via $\hat{e}(y(u),p(u),u) = 0$.

To obtain a second order method in the case when (y,p) are considered as independent variables, one can derive the optimality system for (P) and apply the Newton algorithm to the optimality system. This is referred to as the sequential quadratic programming (SQP) method. Alternatively, if (y,p) are considered as functions of u, then Newton's method can be applied to $(\hat{P})$ directly. The relative merits of these two approaches will be discussed in section 4. To anticipate some of this discussion let us point out that the difference in numerical effort between these two methods is rather small. In fact, after proper rearrangements, the difference in computational cost per iteration of the SQP method for (P) and the Newton method for $(\hat{P})$ consists in solving either the linearized equation (1.2) or the full nonlinear equation itself. In view of the time dependence of either of these two equations, an iterative procedure is used for their solution so that the difference between solving the linearized and nonlinear equation per sweep is not so significant. A second consideration that may influence the choice between SQP method or Newton method applied to $(\hat{P})$ concerns initialization. Initial guesses (y_0,p_0) and u_0 for (y,p,u) can clearly be used independently of each other in the SQP method, where the states are decoupled from the controls. It is sometimes hinted at that this decoupling is not only important for the initialization but also during the iteration and that as a consequence the SQP method may require fewer iterations than Newton's method for $(\hat{P})$, [H]. As we shall see below, the variables y and p can be initialized independently from u_0 also in the Newton method. Specifically, if (y_0,p_0) and u_0 are available, it is not necessary to abandon (y_0,p_0) and compute $(y(u_0),p(u_0))$ from u_0. As for the choice of the initial guess (y_0,p_0,u_0), one possibility is to rely on one of the suboptimal strategies that were developed in the recent past to obtain approximate solutions to (1.1), (1.2). We mention re-

duced order techniques [IR], proper orthogonal decomposition (POD)-based methods [HK, KV, LT], and the instantaneous control method [CTMK, BMT, CHK, HKK]. As another possibility, one can carry out some gradient steps before one switches to the Newton iteration. Let us stress that the numerical realization of the discussed algorithms does not require the availability of the matrix representation of an approximation to the Hessian of the optimal control problem. In fact, storage of the Hessian would be unfeasible. Rather, only the action of the Hessian to a vector is required. In the context of shape optimization this point was elaborated upon in [LA], for example.

Let us briefly comment on some related contributions. In [AT] optimality systems are derived for problems of the type (1.1), (1.2). A gradient technique is proposed in [GM] for the solution of (1.1), (1.2). Similarly, in [B] gradient techniques are analyzed for a boundary control problem related to (1.1), (1.2). In [FGH] the authors analyze optimality systems for exterior boundary control problems. One of the few contributions focusing on second order methods for optimal control of fluids is given in [GB, H]. These works are restricted to stationary problems, however.

This paper, on the other hand, focuses on second order methods for time-dependent problems. We show that despite the difficulties due to the size of (1.1), (1.2) and the fact that the optimality systems contain a two point boundary value problem, forward in time for the primal and backwards in time for the adjoint variables, second order methods are computationally feasible. We establish that the initial approximation to the reduced Hessian is only a compact perturbation of the Hessian at the minimizer. In addition, we give conditions for second order sufficient optimality conditions of tracking-type problems. These results imply superlinear convergence of quasi-Newton as well as SQP methods. While the present paper focuses on distributed control problems, in a future paper we plan to address the case of velocity control along the boundary.

The paper is organized as follows. Section 2 contains the necessary analytic prerequisites. First and second order derivatives of the cost functional with respect to the control are computed in section 3. Section 4 contains a comparison of second order methods to solve (1.1), (1.2). In section 5, the convergence of the quasi-Newton method and SQP methods applied to $(\hat{P})$ is analyzed. Numerical results for the Newton method and comparisons to a gradient method are contained in section 6.

2. The optimal control problem. In this section we consider the optimal control problem (1.1), (1.2) in the abstract form

$$\begin{cases} \min\ J(y,u) \text{ over } (y,u) \in W \times U \\ \text{subject to } e(y,u) = 0. \end{cases} \tag{2.1}$$

To define the spaces and operators arising in (2.1), we assume Ω to be a bounded domain in $\mathbb{R}^2$ with Lipschitz boundary and introduce the solenoidal spaces

$$H = \{v \in C_0^\infty(\Omega)^2 \colon \text{ div } v = 0\}^{-|\cdot|_{L^2}}, V = \{v \in C_0^\infty(\Omega)^2 \colon \text{ div } v = 0\}^{-|\cdot|_{H^1}},$$

with the superscripts denoting closures in the respective norms. Further, we define

$$W = \{v \in L^2(V) \colon v_t \in L^2(V^*)\} \quad \text{and} \quad Z := L^2(V) \times H,$$

W endowed with the norm

$$|v|_W = (|v|^2_{L^2(V)} + |v_t|^2_{L^2(V^*)})^{1/2},$$

$$H^{2,1}(Q) := \{v \in L^2(V \cap H^2(\Omega)^2); v_t \in L^2(H)\}$$

equipped with the norm

$$|v|^2_{H^{2,1}(Q)} := |v|^2_{L^2(V\cap H^2(\Omega)^2)} + |v_t|^2_{L^2(H)},$$

and we set $\langle\cdot,\cdot\rangle := \langle\cdot,\cdot\rangle_{L^2(V^*),L^2(V)}$ with V^* denoting the dual space of V. Here $L^2(V)$ is an abbreviation for $L^2(0,T;V)$, and similarly $L^2(V^*) = L^2(0,T;V^*)$. Recall that up to a set of measure zero in $(0,T)$ elements $v \in W$ can be identified with elements in $C([0,T];H)$, and elements $w \in H^{2,1}(Q)$ can be identified with elements in $C([0,T];V)$. In (2.1), further, U denotes the Hilbert space of controls which is identified with its dual U^*. For the cost functional $J\colon L^2(V)\times U \to \mathbb{R}$ we make the following assumptions:

(H0)
1. J is bounded from below, i.e., $J(y,u) \geq C > -\infty$ for all $(y,u) \in L^2(V)\times U$.
2. J is weakly lower semicontinuous.
3. J is twice Fréchet differentiable with locally Lipschitzian second derivative.
4. J can be decomposed as $J(y,u) = J_1(y) + J_2(u)$.
5. J is radially unbounded in u, i.e., $J(y,u) \to \infty$ as $|u|_U \to \infty$ for every $y \in W$.

The nonlinear mapping

$$e\colon W \times U \to Z^* = L^2(V^*)\times H$$

is defined by

$$e(y,u) = \left(\frac{\partial y}{\partial t} + (y\cdot\nabla)y - \nu\Delta y - Bu, y(0) - y_0\right),$$

where $B \in \mathcal{L}(U, L^2(V^*))$ and $y_0 \in H$. Comparing (1.1), (1.2) to (2.1), we note that the conservation of mass as well as the boundary condition are realized in the choice of the space W, while the dynamics are described by the condition $e(y,u) = 0$. In variational form the constraints in (2.1) can be equivalently expressed as the following: given $u \in U$ find $y \in W$ such that $y(0) = y_0$ and

$$\langle y_t, v\rangle + \langle (y\cdot\nabla)y, v\rangle + \nu(\nabla y, \nabla v)_{L^2(L^2)} = \langle Bu, v\rangle \text{ for all } v \in L^2(V). \tag{2.2}$$

The following existence result for the Navier–Stokes equations in dimension two is well known [CF, L], [T, Chapter III].

PROPOSITION 2.1. *There exists a constant C such that for every $u \in U$ there exists a unique element $y = y(u) \in W$ satisfying*

$$e(y(u),u) = 0$$

and

$$|y|_{C(0,T;H)} + |y|_W \leq C(|y_0|_H + |u|_U + |y_0|^2_H + |u|^2_U).$$

From Proposition (2.1) we conclude that with respect to existence (2.1) is equivalent to

$$\min \hat{J}(u) = J(y(u),u) \text{ subject to } u \in U, \tag{2.3}$$

where $y(u) \in W$ satisfies $e(y(u), u) = 0$.

THEOREM 2.2. *Problem* (2.1) *admits a solution* $(y^*, u^*) = (y(u^*), u^*) \in W \times U$.

Proof. With the above formalism the proof is quite standard, and we give only a brief outline. Since by assumption (H0) 1 J is bounded from below, there exists a minimizing sequence $\{(y_n, u_n)\} = \{y(u_n), u_n\}$ in $W \times U$. Due to (H0) 4, the radial unboundedness property (H0) 5 of J, and Proposition 2.1, the sequence $\{(y_n, u_n)\}$ is bounded in $W \times U$, and hence there exists a subsequence with a weak limit $(y^*, u^*) \in W \times U$. The weakly lower semicontinuity assumption (H0) 2 of $(y, u) \mapsto J(y, u)$ implies that

$$J(y^*, u^*) = \inf\{J(y, u) \colon (y, u) \in W \times U,\ e(y, u) = 0\},$$

and it remains to show that $y^* = y(u^*)$. This can be achieved by passing to the limit in (2.2) with (y, u) replaced by $(y(u_n), u_n)$. ∎

We shall also require the following result concerning strong solutions to the Navier–Stokes equation [T, Theorem III. 3.10], [T1].

PROPOSITION 2.3. *If* $y_0 \in V$ *and* $B \in \mathcal{L}(U, L^2(H))$, *then for every* $u \in U$ *the solution* $y = y(u) \in W$ *to* $e(y, u) = 0$ *satisfies* $y \in H^{2,1}(Q)$. *Moreover, for every bounded set* $\mathcal{U}$ *in* U

$$\{y(u) \colon u \in \mathcal{U}\} \text{ is bounded in } H^{2,1}(Q).$$

We shall frequently refer to the linearized Navier–Stokes system and the adjoint equations given next:

$$\left\{ \begin{array}{l} v_t + (v \cdot \nabla)y + (y \cdot \nabla)v - \nu \Delta v = f \quad \text{in } \Omega \text{ a.e. on } (0, T], \\ v(0) = v_0 \end{array} \right. \tag{2.4}$$

and

$$\left\{ \begin{array}{l} -w_t + (\nabla y)^t w - (y \cdot \nabla)w - \nu \Delta w = g \quad \text{in } \Omega \text{ a.e. on } [0, T), \\ w(T) = 0. \end{array} \right. \tag{2.5}$$

PROPOSITION 2.4. *Let* $y \in W$, $v_0 \in H$, $f \in L^2(V^*)$, *and* $g \in L^\alpha(V^*) \cap W^*$, *with* $\alpha \in [1, \frac{4}{3}]$. *Then* (2.4) *admits a unique variational solution* $v \in W$, *and* (2.5) *has a unique variational solution* $w \in L^2(V)$ *with* $w_t \in L^\alpha(V^*) \cap W^*$, $w \in C(H)$, *and the first equation in* (2.5) *holding in* $L^\alpha(V^*) \cap W^*$. *Moreover, the following estimates hold.*

(i) $|v|_{L^\infty(H)} + |v|_{L^2(V)} \le C(|y|_{L^2(V)}) \left\{ |f|_{L^2(V^*)} + |v_0|_H \right\}$,

(ii) $|v_t|_{L^2(V^*)} \le C(|y|_{L^2(V)}, |y|_{L^\infty(H)}) \left\{ |f|_{L^2(V^*)} + |v_0|_H \right\}$,

(iii) $|w|_{L^2(V)} + |w_t|_{L^\alpha(V^*)} \le C(|y|_{L^2(V)}, |y|_{L^\infty(H)}) \left\{ |g|_{L^\alpha(V^*)} + |g|_{W^*} \right\}$.

If, in addition, $y \in L^\infty(V)$ *and* $g \in L^2(V^*)$, *then* $w \in W$ *and*

(iv) $|w|_{L^2(V)} + |w_t|_{L^2(V^*)} \le C(|y|_{L^\infty(V)})\, |g|_{L^2(V^*)}$.

For $\partial\Omega \in C^2$, $y \in W \cap L^\infty(V) \cap L^2(H^2(\Omega)^2)$, $v_0 \in V$, *and* $f, g \in L^2(H)$ *the unique solutions* v *of* (2.4) *and* w *of* (2.5) *are elements of* $H^{2,1}(Q)$ *and satisfy the a priori estimates*

(v) $|v|_{H^{2,1}(Q)} \le C(|y|_{L^\infty(V)}, |y|_{L^2(H^2(\Omega)^2)}) \left\{ |f|_{L^2(H)} + |v_0|_V \right\}$

and

(vi) $|w|_{H^{2,1}(Q)} \le C(|y|_{L^\infty(V)}, |y|_{L^2(H^2(\Omega)^2)})\, |g|_{L^2(H)}$.

Proof. The proof can be found in [HH, Appendix]. ∎

3. Derivatives. In this section representations for the first and second derivatives of $\hat{J}$ appropriate for the treatment of (2.3) by the Newton and quasi-Newton method are derived. We shall utilize the notation

$$X = W \times U \text{ and } x = (y, u) \text{ for } (y, u) \in W \times U.$$

PROPOSITION 3.1. *The operator $e = (e^1, e^2)\colon X \to Z^*$ is twice continuously differentiable with Lipschitz continuous second derivative. The action of the first two derivatives of e^1 are given by*

$$\begin{aligned}\langle e^1_x(x)(w,s),\phi\rangle &= \langle w_t,\phi\rangle + \langle (w\cdot\nabla)y,\phi\rangle + \langle (y\cdot\nabla)w,\phi\rangle \\ &\quad + \nu(\nabla w,\nabla\phi)_{L^2(L^2)} - \langle Bs,\phi\rangle,\end{aligned}$$

where $x = (y,u) \in X, (w,s) \in X$, and $\phi \in L^2(V)$, and

$$\begin{aligned}\langle e^1_{xx}(x)(w,s)(v,r),\phi\rangle &= \langle e^1_{yy}(x)(w,v),\phi\rangle \\ &= \langle (w\cdot\nabla)v,\phi\rangle + \langle (v\cdot\nabla)w,\phi\rangle =: \langle v, H(\phi)w\rangle_{W,W^*},\end{aligned} \tag{3.1}$$

where $(v,r) \in X$.

Proof. Since e^2 is linear, we restrict our attention to e^1. Let $b\colon V \times V \times V \to \mathbb{R}$ be defined by

$$b(u,v,\phi) = \langle (u\cdot\nabla)v,\phi\rangle_{V^*,V},$$

and recall that, due to the assumption that $\Omega \subset \mathbb{R}^2$,

$$|b(u,v,\phi)|^2 \le 2|u|_H\,|u|_V\,|v|_H\,|v|_V\,|\phi|_V^2 \tag{3.2}$$

for all $(u,v,\phi) \in V \times V \times V$ [T, p. 293]. To argue local Lipschitz continuity of e, let $x, \tilde{x} \in X$ and $\phi \in L^2(V)$. We find

$$\begin{aligned}\langle e^1(x) - e^1(\tilde{x}),\phi\rangle &= \langle (y-\tilde{y})_t,\phi\rangle + \langle ((y-\tilde{y})\cdot\nabla)\tilde{y},\phi\rangle \\ &\quad + \langle (y\cdot\nabla)(y-\tilde{y}),\phi\rangle + \nu(\nabla(y-\tilde{y}),\nabla\phi)_{L^2(L^2)} + \langle B(\tilde{u}-u),\phi\rangle \\ &\le \sqrt{2}\int_0^T |y-\tilde{y}|_H^{1/2}|y-\tilde{y}|_V^{1/2}(|\tilde{y}|_H^{1/2}|\tilde{y}|_V^{1/2} + |y|_H^{1/2}|y|_V^{1/2})|\phi|_V dt \\ &\quad + C|x-\tilde{x}|_X\,|\phi|_{L^2(V)}.\end{aligned}$$

Here and below, C denotes a constant independent of $x, \tilde{x}$, and ϕ. Due to the continuous embedding of W into $L^\infty(H)$, we have

$$\begin{aligned}\langle e^1(x) - e^1(\tilde{x}),\phi\rangle &\le C\Bigg[|x-\tilde{x}|_X\,|\phi|_{L^2(V)} \\ &\quad + |y-\tilde{y}|_{L^\infty(H)}^{1/2}\left(|\tilde{y}|_{L^\infty(H)}^{1/2} + |y|_{L^\infty(H)}^{1/2}\right)\int_0^T |y-\tilde{y}|_V^{1/2}\left(|\tilde{y}|_V^{1/2} + |y|_V^{1/2}\right)|\phi|_V dt\Bigg].\end{aligned}$$

Using Hölder's inequality this further implies the estimate

$$\begin{aligned}\langle e^1(x) - e^1(\tilde{x}),\phi\rangle &\le C\Bigg[|x-\tilde{x}|_X + |y-\tilde{y}|_{L^\infty(H)}^{1/2} \\ &\quad \times \left(|\tilde{y}|_{L^\infty(H)}^{1/2} + |y|_{L^\infty(H)}^{1/2}\right)\left(\int_0^T |y-\tilde{y}|_V\}\,(|\tilde{y}|_V + |y|_V)\,dt\right)^{1/2}\Bigg]|\phi|_{L^2(V)},\end{aligned}$$

and, consequently, after redefining C one last time,

$$\langle e^1(x) - e^1(\tilde{x}), \phi\rangle \le C\,|x-\tilde{x}|_X(|y|_W + |\tilde{y}|_W)|\phi|_{L^2(V)}. \tag{3.3}$$

This estimate establishes the local Lipschitz continuity of e. To verify that the formula for e_x given above represents the Fréchet derivative of e, we estimate

$$\begin{aligned}
&|e^1(\tilde{x}) - e^1(x) - e^1_x(x)(\tilde{x}-x)|_{L^2(V^*)} = \sup_{|\phi|_{L^2(V)}=1} \int_0^T |b(y-\tilde{y}, y-\tilde{y}, \phi)|\,dt \\
&\le \sup_{|\phi|_{L^2(V)}=1} \int_0^T |y-\tilde{y}|_H\,|y-\tilde{y}|_V\,|\phi|_V dt \\
&\le C|y-\tilde{y}|_W \sup_{|\phi|_{L^2(V)}=1} \int_0^T |y-\tilde{y}|_V|\phi|_V\,dt \;\le\; C|y-\tilde{y}|^2_W,
\end{aligned}$$

and the Fréchet differentiability of e follows. To show Lipschitz continuity of the first derivative, let $x, \tilde{x}$, and (v,r) be in X, and estimate

$$\begin{aligned}
&|(e^1_x(\tilde{x}) - e^1_x(x))(v,r)|_{L^2(V^*)} = \sup_{|\phi|_{L^2(V)}=1} \int_0^T |b(y-\tilde{y}, v, \phi) + b(v, y-\tilde{y}, \phi)|\,dt \\
&\le 2\sqrt{2} \sup_{|\phi|_{L^2(V)}=1} \int_0^T |y-\tilde{y}|_H\,|y-\tilde{y}|_V\,|v|_H\,|v|_V\,|\phi|_V dt \\
&\le C|y-\tilde{y}|_W\,|v|_W.
\end{aligned}$$

The expression for the second derivative can be verified by an estimate analogous to the one for the first derivative. The second derivative is independent of the point at which it is taken, and thus it is necessarily Lipschitz continuous. ∎

From (3.2) it follows that for $\phi \in L^2(V)$ and $w \in W$ the mapping

$$\sigma : v \mapsto \langle v, H(\phi)w\rangle_{W,W^*}$$

is an element of W^*. In section 4 we shall use the fact that σ can also be identified with an element of $L^4(V)^* = L^{4/3}(V^*)$.

LEMMA 3.2. *For $\phi \in L^2(V)$ and $w \in W$ the functional σ can be identified with an element in $W^* \cap L^{4/3}(V^*)$.*

Proof. To argue that $\sigma \in L^{4/3}(V^*)$, let $v \in L^4(V)$ and estimate using (3.2)

$$\begin{aligned}
\sigma(v) = \int_0^T b(w,v,\phi) + b(v,w,\phi)\,dt &\le 2\sqrt{2k}|w|^{\frac{1}{2}}_{L^\infty(H)} \int_0^T |w|^{\frac{1}{2}}_V |v|_V |\phi|_V\,dt \\
&\le 2\sqrt{2k}|w|^{\frac{1}{2}}_{L^\infty(H)}|\phi|_{L^2(V)}|w|^{\frac{1}{2}}_{L^2(V)}|v|_{L^4(V)},
\end{aligned}$$

where k is the embedding constant of V into H. This gives the claim. ∎

PROPOSITION 3.3. *Let $x = (y,u) \in W \times U$. Then $e_y(x)\colon W \to Z^*$ is a homeomorphism. Moreover, if the inverse of its adjoint $e_y^{-*}(x)\colon W^* \to Z$ is applied to an element $g \in W^* \cap L^\alpha(V^*)$, $\alpha \in [1, 4/3]$, then, setting $(w, w_0) := e_y^{-*}(x)g \in L^2(V) \times H$, we have $w_t \in L^\alpha(V^*)$, $w(0) = w_0$, and w is the variational solution to (2.5).*

Proof. Due to Proposition 3.1, $e_y(x)$ is a bounded linear operator. By the closed range theorem the claim follows once it is argued that (2.4) has a unique solution $v \in W$ for every $(f, v_0) \in Z^*$. This is a direct consequence of Proposition 2.4(i) and

(ii). The assertion concerning the adjoint follows from (iii) of the same proposition and its proof. □

As a consequence of Propositions 3.1 and 3.3 and the implicit function theorem, the first derivative of the mapping $u \to y(u)$ at u in direction δu is given by

$$y'(u)\delta u = -e_y^{-1}(x)e_u(x)\delta u, \tag{3.4}$$

where $x = (y(u), u)$. By the chain rule we thus obtain

$$\langle \hat{J}'(u), \delta u\rangle_U = \langle J_u(x) - e_u^*(x)e_y^{-*}(x)J_y(x), \delta u\rangle_U.$$

Introducing the variable

$$\lambda = -e_y^{-*}(x)J_y(x) \in Z, \tag{3.5}$$

we obtain, utilizing Proposition 2.4(iii) with $g = -J_y(x) \in L^2(V^*)$, the Riesz representation for the first derivative of $u \to \hat{J}(u)$:

$$\hat{J}'(u) = J_u(x) + e_u^*\lambda. \tag{3.6}$$

Here $\lambda = (\lambda^1, \lambda^0) \in Z$, $\lambda_t^1 \in L^{4/3}(V^*)$, $\lambda^1 \in C(H)$, and λ^1 is the variational solution of

$$\begin{cases} -\lambda_t^1 + (\nabla y)^t\lambda^1 - (y\cdot\nabla)\lambda^1 - \nu\Delta\lambda^1 = -J_y(x), \\ \lambda^1(T) = 0, \end{cases} \tag{3.7}$$

where the first equation holds in $L^{4/3}(V^*) \cap W^*$.

The computation of the second derivative of $\hat{J}''(u) \in \mathcal{L}(U)$ of $\hat{J}$ is more involved. Let $(\delta u, \delta v) \in U \times U$, and note that the second derivative of $u \to y(u)$ from U to W can be expressed as

$$y''(u)(\delta u, \delta v) = -e_y^{-1}(x)e_{yy}(x)(y'(u)\delta u, y'(u)\delta v). \tag{3.8}$$

By the chain rule and since $W \subset L^2(V)$ and hence $L^2(V^*) \subset W^*$, we have

$$\begin{aligned} \langle \hat{J}''(u)\delta u, \delta v\rangle_U &= \langle J_{yy}(x)y'(u)\delta u, y'(u)\delta v\rangle \\ &\quad + \langle J_y(x), y''(u)(\delta u, \delta v)\rangle + \langle J_{uu}(x)\delta u, \delta v\rangle_U \\ &= \langle J_{yy}(x)y'(u)\delta u, y'(u)\delta v\rangle - \langle J_y(x), e_y^{-1}e_{yy}(x)(y'(u)\delta u, y'(u)\delta v)\rangle \\ &\quad + \langle J_{uu}(x)\delta u, \delta v\rangle_U \\ &= \langle J_{yy}(x)y'(u)\delta u, y'(u)\delta v\rangle + \langle \lambda^1, e_{yy}^1(x)(y'(u)\delta u, y'(u)\delta v)\rangle \\ &\quad + \langle J_{uu}(u)\delta u, \delta v\rangle_U. \end{aligned}$$

We introduce the Lagrangian $L\colon X \times Z \to \mathbb{R}$

$$L(x, \lambda) = J(x) + \langle e(x), \lambda\rangle_{Z^*,Z} \tag{3.9}$$

and the matrix operator

$$T(x) = \begin{pmatrix} -e_y^{-1}(x)e_u(x) \\ Id_U \end{pmatrix} \in \mathcal{L}(U, X). \tag{3.10}$$

We observe that the second derivative of L with respect to x can be expressed as

$$L_{xx}(x, \lambda) = \begin{pmatrix} J_{yy}(x) + & \langle e_{yy}^1(x)(\cdot,\cdot),\, \lambda^1\rangle & 0 \\ & 0 & J_{uu}(x) \end{pmatrix} \in \mathcal{L}(X, X^*).$$

The above computation for $\hat{J}''(u)$ together with (3.4) imply that

$$\hat{J}''(u) = T^*(x)L_{xx}(x, \lambda)T(x), \tag{3.11}$$

where $x = (y(u), u)$.

4. Second order methods. This section contains a description and a comparison of second order methods to solve (2.1). Throughout, u^* denotes a (local) solution to (2.1).

4.1. Newton and quasi-Newton algorithm. For the sake of reference let us specify the Newton algorithm.

ALGORITHM 4.1 (Newton algorithm).

1. *Choose* $u^0 \in N(u^*)$*; set* $k = 0$.
2. *Do until convergence:*
   (i) *solve* $\hat{J}''(u^k)\delta u^k = -\hat{J}'(u^k)$,
   (ii) *update* $u^{k+1} = u^k + \delta u^k$,
   (iii) *set* $k = k + 1$.

Let us consider the linear system in 2(i). Its dimension is that of the control space U. From the characterization of the Hessian $\hat{J}''(u^k)$ we conclude that the number of solutions to the linearized Navier–Stokes equation (3.4) with appropriate right-hand sides that are required for its evaluation equals the dimension of U. If U is infinite dimensional, then an appropriate discretization must be carried out. Let us assume now that the dimension of U is large so that direct evaluation of $\hat{J}''(u^k)$ is not feasible. In this case 2(i) must be solved iteratively, e.g., by a conjugate gradient technique. We shall then refer to 2(i) as the "inner" loop as opposed to the do loop in 2, which is the "outer" loop of the Newton algorithm. The inner loop at iteration level k of the outer loop requires us to

(i) evaluate $\hat{J}'(u^k)$, i.e., given u^k, compute $y(u^k)$ from (1.2) and λ^1 from (3.7) with $x = (y(u^k), u^k)$;
(ii) iteratively evaluate the action of $\hat{J}''(u^k)$ on δ_j^k, the jth iterate of the inner loop on the kth level of the outer loop.

The iterate $q = \hat{J}''(u^k)\delta_j^k$ can be evaluated by successively applying the following steps:

(a) solve in $L^2(V^*)$ for $v \in W$

$$\begin{aligned} v_t + (v \cdot \nabla)y + (y \cdot \nabla)v - \nu\Delta v &= B\delta_j^k, \\ v(0) &= 0, \end{aligned}$$

where $y = y(u^k)$;
(b) evaluate $J_{yy}(x)v + \langle e^1_{yy}(x)(v, \cdot), \lambda^1\rangle$;
(c) solve in W^* for $w \in L^2(V)$

$$e_y(x)^* w \;=\; J_{yy}(x)v + \langle e^1_{yy}(x)(v, \cdot), \lambda^1\rangle;$$

(d) and, finally, set $q := J_{uu}\delta u + B^* w$.

We recall that $\lambda^1 \in L^2(V)$ and that for $s \in W$

$$\langle e^1_{yy}(x)(v, s), \lambda^1\rangle = \int_0^T \int_\Omega ((v \cdot \nabla)s\lambda^1 + (s \cdot \nabla)v\lambda^1)dx\, dt.$$

Moreover, by Lemma 3.2 the functional appearing in (b) is an element of $W^* \cap L^{4/3}(V^*)$. Hence by Proposition 2.4 the adjoint equation in (c) can equivalently be rewritten as

$$\begin{aligned} -w_t + (\nabla y)^t w - (y \cdot \nabla)w - \nu\Delta w &= J_{yy}(x)v + \langle e^1_{yy}(x)(v, \cdot), \lambda^1\rangle, \\ w(T) &= 0, \end{aligned}$$

where the first equation holds in $W^* \cap L^{4/3}(V^*)$. Summarizing, for the outer iteration of the Newton method one Navier–Stokes solve for $y(u^k)$ and one linearized Navier–Stokes solve for $\lambda(u^k)$ are required. For the inner loop one forward (in time) as well as one backward linearized Navier–Stokes solve per iteration are necessary.

Concerning initialization, we observe that if initial guesses $(y_0, u_0) \in W \times U$ are available (with y_0 not necessarily $y(u_0)$), then, alternatively to the initialization in Algorithm 4.1, this information can be used advantageously to compute the adjoint variable λ^1 required for the initial guess for the right-hand side of the linear system as well as to carry out steps (a)–(c) for the evaluation of the Hessian. There is no necessity to recompute $y(u_0)$ from u_0.

To avoid the difficulties of evaluating the action of the exact Hessian in Algorithm 4.1, one can resort to quasi-Newton algorithms. Here we specify one of the most prominent candidates, the BFGS method. For w and z in U we define the rank-one operator $w \otimes z \in \mathcal{L}(U)$, the action of which is given by

$$(w \otimes z)(v) = \langle z, v \rangle_U \, w.$$

In the BFGS method the Hessian $\hat{J}''$ at u^* is approximated by a sequence of operators H^k.

Algorithm 4.2 (BFGS algorithm).

1. *Choose* $u^0 \in N(u^*)$, $H^0 \in \mathcal{L}(U)$ *symmetric, set* $k = 0$.
2. *Do until convergence:*
   - (i) *solve* $H^k \delta u^k = -\hat{J}'(u^k)$,
   - (ii) *update* $u^{k+1} = u^k + \delta u^k$,
   - (iii) *compute* $\hat{J}'(u^{k+1})$,
   - (iv) *set* $s^k = u^{k+1} - u^k$, $d^k = \hat{J}'(u^{k+1}) - \hat{J}'(u^k)$,
   - (v) *update* $H^{k+1} = H^k + \frac{d^k \otimes d^k}{\langle d^k, s^k \rangle_U} - \frac{H^k s^k \otimes H^k s^k}{\langle H^k s^k, s^k \rangle_U}$,
   - (vi) *set* $k = k + 1$.

Note that the BFGS algorithm requires no more system solves than the gradient algorithm applied to (2.1), which is one forward solution of the nonlinear equation to obtain $y(u^k)$ and one backward solve of the linearized equation (3.7) to obtain the adjoint variable $\lambda(u^k)$.

In order to compare Newton's method to the SQP method derived in the next section, we rewrite the update step 2(i) in Algorithm 4.1. To begin with, we observe that the right-hand side in the update step can be written with the help of the adjoint variable λ from (3.5) and the operator $T(x)$ defined in (3.10) as

$$-\hat{J}_u(u) = -J_u(x) - e_u^*(x)\lambda = -T^*(x) \begin{bmatrix} 0 \\ J_u(x) + e_u^*(x)\lambda \end{bmatrix}, \tag{4.1}$$

where we dropped the iteration indices. As a consequence, with $\delta y = y'(u)\delta u$ from (3.3), the update can be written as

$$T^*(x) L_{xx}(x, \lambda) \begin{bmatrix} \delta y \\ \delta u \end{bmatrix} = -T^*(x) \begin{bmatrix} 0 \\ J_u(x) + e_u^*(x)\lambda \end{bmatrix} \tag{4.2}$$

so that

$$L_{xx}(x, \lambda) \begin{bmatrix} \delta y \\ \delta u \end{bmatrix} + \begin{bmatrix} 0 \\ J_u(x) + e_u^*(x)\lambda \end{bmatrix} \in \mathcal{N}(T^*(x))$$

holds. Since $e_x(x) \in \mathcal{L}(X, Z^*)$ and $\mathcal{N}(e_x(x)) = \mathcal{R}(T) \subseteq X$, it follows that $\mathcal{R}(T)$ is closed, and we have the sequence of identities

$$\mathcal{N}(T^*(x)) = \mathcal{R}(T(x))^\perp = \mathcal{N}(e_x(x))^\perp = \mathcal{R}(e_x^*(x)).$$

Thus there exists $\delta\lambda \in Z$ such that

$$-e_x^*(x)\delta\lambda = L_{xx}(x,\lambda) \begin{bmatrix} \delta y \\ \delta u \end{bmatrix} + \begin{bmatrix} 0 \\ J_u(x) + e_u^*(x)\lambda \end{bmatrix}.$$

Using this equation together with the definition of δy, we may rewrite Newton's update as

$$\begin{bmatrix} L_{xx}(x^k,\lambda^k) & e_{x^*}(x^k) \\ e_x(x^k) & 0 \end{bmatrix} \begin{bmatrix} \delta y \\ \delta u \\ \delta\lambda \end{bmatrix} = - \begin{bmatrix} 0 \\ J_u(x) + e_u^*(x)\lambda \\ 0 \end{bmatrix}. \tag{4.3}$$

4.2. Basic SQP method. Here we regard (2.1) as a minimization problem of the functional J over the space X subject to the explicit constraint $e(x) = 0$. The basic SQP algorithm consists in applying Newton's method to the first order optimality system

$$\begin{aligned} L_x(x,\lambda) &= 0 \quad \text{in } X^*, \\ L_\lambda(x,\lambda) &= 0 \quad \text{in } Z^*, \end{aligned} \tag{4.4}$$

where the Lagrangian L is defined in (3.9).

With x^* denoting a solution to problem (P), $e_x(x^*)$ is surjective by Proposition 3.3, and hence there exists a Lagrange multiplier $\lambda^* \in Z$, which is even unique such that (4.4) holds. The SQP method will be well defined and locally second order convergent if in addition to the surjectivity of $e_x(x^*)$ the following second order optimality condition holds.

(H1) There exists $\alpha > 0$ such that
$\langle L_{xx}(x^*,\lambda^*)x, x\rangle_{X^*,X} \geq \alpha\,|x|_X^2$ for all $x \in \ker(e_x(x^*))$.

If (H1) holds, then, due to the regularity properties of e, there exists a neighborhood $B((x^*,\lambda^*))$ such that $L_{xx}(x,\lambda)$ is uniformly positive definite on $\ker(e_x(x))$ for every $x \in B((x^*,\lambda^*))$.

ALGORITHM 4.3 (SQP algorithm).

1. *Choose* $(x^0,\lambda^0) \in B((x^*,\lambda^*))$, *set* $k = 0$.
2. *Do until convergence:*
   (i) *solve*

$$\begin{pmatrix} L_{xx}(x^k,\lambda^k) & e_x^*(x^k) \\ e_x(x^k) & 0 \end{pmatrix} \begin{pmatrix} \delta x^k \\ \delta\lambda^k \end{pmatrix} = - \begin{pmatrix} J_x(x^k) & + e_x^*(x^k)\lambda^k \\ & e(x^k) \end{pmatrix}, \tag{4.5}$$

   (ii) *update* $(x^{k+1},\lambda^{k+1}) = (x^k,\lambda^k) + (\delta x^k, \delta\lambda^k)$,
   (iii) *set* $k = k+1$.

Just as for Newton's method, step 2(i) is the difficult one. While in contrast to Newton's method, neither the Navier–Stokes equation nor its linearization needs to be solved; the dimension of the system matrix which is twice the dimension of the state plus the dimension of the control space is formidable for applications in fluid

mechanics. In addition, from experience with Algorithm 4.3 for other optimal control problems (see [KA, V], for example) it is well known that preconditioning techniques must be applied to solve (4.5) efficiently. As a preconditioner one might consider the (action of the) operator $P\colon X^* \times Z^* \to X \times Z$ given by

$$P = \begin{bmatrix} 0 & 0 & R \\ 0 & J_{uu}(x^k)^{-1} & 0 \\ R^* & 0 & 0 \end{bmatrix},$$

where $R\colon Z^* \to H$ is the inverse to the (discretized) instationary Stokes operator or the (discretized) linearization of the Navier–Stokes equation at the state y^k, either one with homogenous boundary conditions.

One iteration of the preconditioned version of Algorithm 4.3 therefore requires two linear parabolic solves, one forward and one backward in time. As a consequence, even with the application of preconditioning techniques, the numerical expense counted in number of parabolic system solves is less for the SQP method than for Newton's method. However, the number of iterations of iterative methods applied to solve the system equations in Algorithms 4.1 and 4.3 strongly depends on the system dimension, which is much larger for Algorithm 4.3 than for Algorithm 4.1.

To further compare the structure of the Newton and the SQP methods, let us assume, for an instance, that x^k is feasible for the primal equation, i.e., $e(x^k) = 0$ and (x^k, λ^k) is feasible for the adjoint equation (3.5), i.e., $e_y^*(x^k)\lambda^k = -J_y(x^k)$. Then the right-hand side of (4.5) has the form $-[0, J_u(x^k) + e_u^*\lambda^k, 0]^t$, and comparing this to the computation at the end of section 4.1, we observe that the linear systems describing the Newton and the SQP methods coincide. In general, the nonlinear primal and the linearized adjoint equation will not be satisfied by the iterates of the SQP method, and we therefore refer to the SQP method as an outer or unfeasible method, while the Newton method is a feasible one.

4.3. Reduced SQP method. The idea of the reduced SQP method is to replace (4.5) with an equation in $\ker e_x(x)$ so that the reduced system is of smaller dimension than the original one. To develop the reduced system, we follow the lines of [KS]. Recall the definition of $T(x)\colon U \to X$, and define $A(x)\colon Z^* \to X$ by

$$A(x) = \begin{pmatrix} e_y^{-1}(x) \\ 0 \end{pmatrix}. \tag{4.6}$$

Note that A is a right-inverse to $e_x(x)$. In fact, we have

(i) $\ker e_x(x) = \mathcal{R}(T(x)) = \{\begin{pmatrix} -e_y^{-1}(x)e_u(x)v \\ v \end{pmatrix}\colon v \in U\}$,
(ii) $e_x(x)T(x) = 0$ in Z^*,
(iii) $e_x(x)A(x) = I_{Z^*}$.

By Proposition 3.3 and due to $B \in \mathcal{L}(U, L^2(V^*))$, the operator $T(x)$ is an isomorphism from U to $\ker e_x(x)$, and hence the second equality in (4.5) given by

$$e_x(x)\delta x = -e(x)$$

can be expressed as

$$\delta x = T(x)\delta u - A(x)e(x). \tag{4.7}$$

Using this in the first equality of (4.5), we find

$$L_{xx}(x,\lambda)T(x)\delta u - L_{xx}(x,\lambda)A(x)e(x) + e_x^*(x)\delta\lambda = -(J_x(x) + e_x^*(x)\lambda).$$

Applying $T^*(x)$ to this last equation and (ii) from above implies that if δu is a solution coordinate of (4.5), then it also satisfies

$$T^*(x)L_{xx}(x,\lambda)T(x)\delta u = T^*(x)L_{xx}(x,\lambda)A(x)e(x) - T^*(x)J_x(x). \tag{4.8}$$

Once δu is computed from (4.8), then δy and $\delta\lambda$ can be obtained from (4.7) (which requires one forward linear parabolic solve) and the first equation in (4.5) (another backwards linear parabolic solve).

Let us note that if x is feasible, then the first term on the right-hand side of (4.8) is zero and (4.8) is identical to step 2(i) in Newton's algorithm (Algorithm 4.1).

This again reflects the fact that Newton's method can be viewed as an SQP method that obeys the feasibility constraint $e(x) = 0$. It also points at the fact that the amount of work (measured in equation solves) for the inner loop coincides for both the Newton and the reduced SQP methods. The significant difference between the two methods lies in the outer iteration. To make this evident we next specify the reduced SQP algorithm.

ALGORITHM 4.4 (reduced SQP algorithm).

1. *Choose* $x^0 \in B(x^*)$, *set* $k = 0$.
2. *Do until convergence:*
   (i) *Lagrange multiplier update: solve*
   $$e_y^*(x^k)\lambda^k = -J_y(x^k),$$
   (ii) *solve*
   $$\alpha)\ T^*(x^k)L_{xx}(x^k,\lambda^k)T(x^k)\delta u^k = T^*(x^k)L_{xx}(x^k,\lambda^k)A(x^k)e(x^k) - T^*(x^k)J_x(x^k)$$
   $$\beta)\ e_y(x^k)\delta y^k = -e(x^k) - e_u(x^k)\delta u^k,$$
   (iii) *update*
   $$x^{k+1} = x^k + (\delta y^k, \delta u^k),$$
   (iv) *set* $k = k + 1$.

Note that in the algorithm that we specified we did not follow the procedure outlined above for the update of the Lagrange variable. In fact, for reduced SQP methods there is no "optimal" update strategy for λ. The two choices described above are natural and frequently used. To implement Algorithm 4.4 two linear parabolic systems have to be solved in steps 2(i) and 2(ii)β) and, in addition, two linear parabolic systems are necessary to evaluate the term involving the operator A on the right-hand side of 2(ii)α). In applications this term is often neglected since it vanishes at x^*.

The reduced SQP method and Newton's method turn out to be very similar. Let us discuss the points in which they differ:

(i) Most significantly, the velocity field is updated by means of the nonlinear equation in Newton's method and via the linearized equation in the reduced SQP method.

(ii) The right-hand sides of the linear systems differ due to the appearance of the term involving the operator A. As mentioned above, this term is frequently not implemented.

(iii) Formally there is a difference in the initialization procedure in that y^0 is chosen independently from u^0 in the reduced SQP method and $y^0 = y(u^0)$ in Newton's method. However, as explained in section 4.1 above, if a good initial guess y^0 independent from $y(u^0)$ is available, it can be utilized in Newton's method as well.

5. Convergence analysis. We present local convergence results for the algorithms introduced in section 4 for cost functionals J which are of separable type; i.e., (H0) 4 is satisfied. For this purpose it will be essential to derive conditions that ensure positive definiteness of $\hat{J}''(u^*)$ and (H1). The key to these conditions are the a priori estimates of Proposition 2.4. We shall also prove that the difference $\hat{J}''(u^*) - J_{uu}(x^*)$ is compact. This property is required for the rate of convergence analysis of quasi-Newton methods. In our first result we assert positive definiteness of the Hessian provided that $J_y(x)$ is sufficiently small, a condition which is applicable to tracking-type problems.

LEMMA 5.1 (positive definiteness of the Hessian). *Let $u \in U$, and assume that $J_{yy}(x) \in \mathcal{L}(L^2(V), L^2(V^*))$ is positive semidefinite and $J_{uu}(x) \in \mathcal{L}(U)$ is positive definite, where $x = (y(u), u)$. Then the Hessian $\hat{J}''(u)$ is positive definite provided that $|J_y(x)|_{L^2(V^*)}$ is sufficiently small.*

Proof. We recall from (3.11) that

$$\hat{J}''(u) = T^*(x) L_{xx}(x, \lambda) T(x),$$

where $x = (y(u), u)$ and $\lambda = \lambda(x)$ is the solution to (3.7). It follows that

$$\begin{aligned} \hat{J}''(u) = {} & e_u^*(x) e_y^{-*}(x) J_{yy}(x) e_y^{-1}(x) e_u(x) \\ & + e_u^*(x) e_y^{-*}(x) \langle e_{yy}^1(x)(e_y^{-1}(x) e_u(x), \cdot), \lambda^1(x) \rangle + J_{uu}(x). \end{aligned} \tag{5.1}$$

Here we note that for $\delta u \in U$ the functional

$$w \mapsto \langle e_{yy}(x)(e_y^{-1}(x) e_u(x) \delta u, w), \lambda^1 \rangle$$

is an element of W^*. Since $J_{yy}(x)$ is assumed to be positive definite and $J_{uu}(x)$ is positive definite, the result will follow provided the operator norm of

$$\mathcal{R} := e_u^*(x) e_y^{-*}(x) \langle e_{yy}^1(x)(e_y^{-1}(x) e_u(x), \cdot), \lambda^1(x) \rangle \ \in \mathcal{L}(U) \tag{5.2}$$

can be bounded by $|J_y(x)|_{L^2(V^*)}$. Straightforward estimation gives

$$\begin{aligned} \|\mathcal{R}\|_{\mathcal{L}(U)} \ &\leq \ \|e_u^*(x) e_y^{-*}(x)\|_{\mathcal{L}(W^*, U)} \\ &\quad \|\langle e_{yy}^1(x)(\cdot, \cdot), \lambda^1(x) \rangle\|_{\mathcal{L}(W, W^*)} \, \|e_y^{-1}(x) e_u(x)\|_{\mathcal{L}(U, W)} \\ &= \ \|e_y^{-1}(x) e_u(x)\|_{\mathcal{L}(U,W)}^2 \, \|\langle e_{yy}^1(x)(\cdot, \cdot), \lambda^1(x) \rangle\|_{\mathcal{L}(W, W^*)}. \end{aligned} \tag{5.3}$$

From Proposition 2.4 we conclude that

$$\|e_y^{-1}(x) e_u(x)\|_{\mathcal{L}(U,W)} \ \leq \ C(|y|_{L^2(V)}, |y|_{L^\infty(H)}, \|B\|_{\mathcal{L}(U, L^2(V^*))}).$$

To estimate $\|\langle e_{yy}^1(x)(\cdot, \cdot), \lambda^1(x) \rangle\|_{\mathcal{L}(W, W^*)}$ we recall that for $g, h \in W$

$$\langle e_{yy}^1(x)(g, h), \lambda^1(x) \rangle \ = \ \int_0^T \int_\Omega (g \cdot \nabla) h \lambda^1 \ + \ (h \cdot \nabla) g \lambda^1 \, dx dt.$$

Using (3.2) and the continuity of the embedding $W \hookrightarrow L^\infty(H)$, we may estimate

$$|\langle e_{yy}^1(x)(g, h), \lambda^1(x) \rangle| \ \leq \ C \, |g|_W |h|_W \, |\lambda^1|_{L^2(V)},$$

with a constant C independent of g and h. Therefore,

$$\|\mathcal{R}\|_{\mathcal{L}(U)} \;\le\; C(|y|_{L^2(V)}, |y|_{L^\infty(H)}, \|B\|_{\mathcal{L}(U,L^2(V^*))})\,|\lambda^1|_{L^2(V)} \\ \le C(|y|_{L^2(V)}, |y|_{L^\infty(H)}, \|B\|_{\mathcal{L}(U,L^2(V^*))})\,|J_y(x)|_{L^2(V^*)},$$

where we applied (iii) in Proposition 2.4 to (3.7). □

LEMMA 5.2. *Let $x \in X$, and denote by $\lambda = \lambda(x) \in Z$ the function defined in* (3.5). *Then, under the assumptions of Lemma* 5.1 *on J condition,* (H1) *is satisfied with (x^*, λ^*) replaced by (x, λ).*

Proof. Let $(v, u) \in \mathcal{N}(e_x(x))$. Then v solves (2.4) with $v_0 = 0$ and $f = Bu$. Due to Proposition 2.3, $v \in W$ and satisfies

$$|v|_W \;\le\; C(|y|_{L^2(V)}), |y|_{L^\infty(H)}, \|B\|_{\mathcal{L}(U,L^2(V^*))})\,|u|_U. \tag{5.4}$$

Let $\delta > 0$ be chosen such that $J_{uu}(x)(u,u) \ge \delta|u|_U^2$ for all $u \in U$. We find

$$\langle L_{xx}(x,\lambda)(v,u),(v,u)\rangle_{X^*,X} = J_{yy}(x)(v,v) + \langle e^1_{yy}(x)(v,v),\lambda^1\rangle + J_{uu}(x)(u,u) \\ \ge\; \delta\,|u|_U^2 - 2\sqrt{2}\int_0^T |v|_H|v|_V|\lambda^1|_V\,dt \;\ge\; \delta\,|u|_U^2 - C|u|_U^2\,|\lambda^1|_{L^2(V)}.$$

Here and below, C denotes a generic constant independent of (v,u), and $\lambda = \lambda(x)$. Due to (3.5) and Proposition 2.4

$$|\lambda|_{L^2(V)} \;\le\; C\,|J_y(x)|_{L^2(V^*)}.$$

These estimates imply

$$\langle L_{xx}(x,\lambda)(v,u),(v,u)\rangle_{X^*,X} \;\ge\; \big(\delta - C|J_y(x)|_{L^2(V^*)}\big)\,|u|_U^2,$$

and combined with (5.4) the claim follows. □

LEMMA 5.3. *If $B \in \mathcal{L}(U, L^2(H))$, then the difference*

$$\hat{J}''(u) - J_{uu}(x)$$

is compact for every $u \in U$.

Proof. Utilizing (5.2), we may rewrite

$$\hat{J}''(u) - J_{uu}(x) = e_u^*(x)e_y^{-*}(x)J_{yy}(x)e_y^{-1}(x)e_u(x) \,+\, \mathcal{R}, \tag{5.5}$$

where $x = (y(u), u)$. It will be shown that both summands define compact operators on U. For this purpose let $\mathcal{U}$ be a bounded subset of U. Utilizing $B \in \mathcal{L}(U, L^2(H)) \subset \mathcal{L}(U, L^2(V^*))$ and Proposition 2.4, we conclude that

$$S = \{e_y^{-1}(x)e_u(x)\delta u : \delta u \in \mathcal{U}\}$$

is a bounded subset of W and hence of $L^2(V)$. Since by assumption J is twice continuously Fréchet differentiable with respect to y from $L^2(V)$ to $\mathbb{R}$, it follows that $J_{yy}(S)$ is a bounded subset of $L^2(V^*)$. Proposition 2.4(iii) implies that, consequently, $e_y^{-*}(J_{yy}(S))$ is bounded in $W^2_{4/3} \times H$, where $W^2_{4/3} := \{v \in L^2(V) : v_t \in L^{4/3}(V^*)\}$. Since $W^2_{4/3}$ is compactly embedded in $L^2(H)$ [CF] and $B \in \mathcal{L}(U, L^2(H))$, it follows from the fact that $e_u^*(x)(z^1, z^0) = -B^*z^1$ for $z = (z^1, z^0) \in L^2(V) \times H$ that

$$\{e_u^*(x)e_y^{-*}(x)J_{yy}(x)(z) : z \in S\} \tag{5.6}$$

is precompact in U.

Let us turn to the second addend in (5.5). Due to Lemma 3.2 and its proof, the set

$$\{\langle e^1_{yy}(x)(z,\cdot),\lambda^1\rangle : z \in S\}$$

is a bounded subset of $W^* \cap L^{4/3}(V^*)$. It follows, utilizing Proposition 2.4, that

$$\{e_y^{-*}(x)\langle e^1_{yy}(x)(z,\cdot),\lambda^1\rangle : z \in S\}$$

is a bounded subset of $W^2_{4/3} \times H$. As above, the assumption that $B \in \mathcal{L}(U, L^2(H))$ implies that

$$\{e_u^*(x)e_y^{-*}(x)\langle e^1_{yy}(x)(z,\cdot),\lambda^1\rangle; z \in S\}$$

is precompact in U, and the lemma is verified. □

The following lemma concerning the operators $T(x)$ and $A(x)$ defined in (3.10) and (4.6) will be required for the analysis of the reduced SQP method.

LEMMA 5.4. *The mappings $x \mapsto A(x)$ from X to $\mathcal{L}(Z^*, X)$ and $x \mapsto T(x)$ from X to $\mathcal{L}(U, X)$ are Fréchet differentiable with Lipschitz continuous derivatives.*

Proof. The proof is an immediate consequence of (i), (ii) in Proposition 2.4 and the identities (ii) and (iii) in section (4.3) together with the differentiability properties of the mapping $x \mapsto e_x(x)$. □

We are now in the position to prove local convergence for the algorithms discussed in section 4. Throughout we assume that (y^*, u^*) is a local solution to (2.1) and set $y^* = y(u^*)$, $x^* = (y^*, u^*)$. In addition to the general conditions on J, B, and e, we require

(H2) $J_{yy}(x^*) \in \mathcal{L}(L^2(V), L^2(V^*))$ is positive semidefinite, $J_{uu}(x^*) \in \mathcal{L}(U)$ is positive definite, and $|J_y(x^*)|_{L^2(V^*)}$ is sufficiently small.

With (H2) holding, (H1) is satisfied due to Lemma 5.1. In particular, a second order sufficient optimality condition holds, and (y^*, u^*) is a strict local solution to (2.1). The following theorem follows from well-known results on Newton's algorithm.

THEOREM 5.5. *If* (H2) *holds, then there exists a neighborhood $\mathcal{U}(u^*)$ such that for every $u^0 \in \mathcal{U}(u^*)$ the iterates $\{u^n\}_{n\in\mathbb{N}}$ of Newton's algorithm (Algorithm* 4.1*) converge quadratically to u^*.*

THEOREM 5.6. *If* (H2) *holds, then there exists a neighborhood $\mathcal{U}(u^*)$ and $\epsilon > 0$ such that for all $u^0 \in \mathcal{U}(u^*)$ and all positive definite operators $H^0 \in L(U)$ with*

$$|H^0 - \hat{J}''(u^*)|_{\mathcal{L}(U)} < \epsilon,$$

the BFGS method of Algorithm 4.2 *converges linearly to u^*. If in addition $B \in \mathcal{L}(U, L^2(H))$ and $H^0 := J_{uu}(x^*)$, then the convergence is superlinear.*

Proof. For the first part of the theorem we refer to [GR, section 4], for example. For the second claim we observe that the difference $\hat{J}''(u^*) - J_{uu}(x^*)$ is compact by Lemma 5.3 so that the claim follows from [GR, Theorem 5.1]; see also [KS1]. □

THEOREM 5.7. *Assume that* (H2) *holds, and let λ^* be the Lagrange multiplier associated to x^*. Then there exists a neighborhood $\mathcal{U}(x^*, \lambda^*) \subset X \times Z$ such that for*

all $(x^0, \lambda^0) \in \mathcal{U}(x^*, \lambda^*)$ *the SQP algorithm (Algorithm* 4.3*) is well defined, and the iterates* $\{(x^n, \lambda^n)\}_{n\in\mathbb{N}}$ *converge quadratically to* (x^*, λ^*).

Proof. Since J and e are twice continuously differentiable with Lipschitz continuous second derivative, $e_x(x^*)$ is surjective by Proposition 3.3, and (H1) is satisfied, second order convergence of the SQP method follows from standard results; see, for instance, [IK]. ∎

We now turn to the reduced SQP method.

THEOREM 5.8. *Assume that* (H1) *holds, and let* λ^* *denote the Lagrange multiplier associated to* x^*. *Then there exists a neighborhood* $\mathcal{U}(x^*) \subset X$ *such that for all* $x^0 \in \mathcal{U}(x^*)$ *the reduced SQP algorithm (Algorithm* 4.4*) is well defined, and its iterates* $\{x^k\}_{k\in\mathbb{N}}$ *converge two-step quadratically to* x^*, *i.e.,*

$$|x^{k+1} - x^*|_X \;\leq\; C\,|x^{k-1} - x^*|_X^2$$

for some positive constant C *independent of* $k \in \mathbb{N}$.

Proof. First note that (H1) implies positive definiteness of $T(x^*)^* L_{xx}(x^*, \lambda^*) T(x^*)$ in a neighborhood $\tilde{\mathcal{U}}(x^*)$ of x^*. By Lemma 5.4 the mappings $x \mapsto T(x)$ and $x \mapsto A(x)$ are Fréchet differentiable with Lipschitz continuous derivatives. Furthermore, it can be shown that the mapping $x \mapsto \lambda(x)$ is locally Lipschitz continuous, where λ is defined through (3.5) [HH, Lemma 4.5.2]. This, in particular, implies for the Lagrange multiplier updates λ^k the estimate

$$|\lambda^k - \lambda|_Z \;\leq\; C\,|x^k - x^*|_X, \quad x^k \in \tilde{\mathcal{U}}(x^*),$$

where the constant C is positive and depends on x^* and on $\sup\{|J_{yy}(x)|_{\mathcal{L}(L^2(V),L^2(V^*)};\ x \in \tilde{\mathcal{U}}(x^*)\}$. Altogether, the assumptions for Corollary 3.6 in [K] are met, and there exists a neighborhood $\hat{\mathcal{U}}(x^*)$ such that for all $x^0 \in \mathcal{U}(x^*) := \hat{\mathcal{U}}(x^*) \cap \tilde{\mathcal{U}}(x^*)$ the claim follows. ∎

6. Numerical results. Here we present a numerical example that should first demonstrate the feasibility of utilizing Newton's method for optimal control of the two-dimensional instationary Navier–Stokes equations in a workstation environment despite the formidable size of the optimization problem. The total number of unknowns (primal, adjoint, and control variables) in example 1 below, for instance, is of order $2.2{*}10^6$. The time horizon could still be increased or the mesh size decreased by utilizing reduced storage techniques at the expense of additional cpu time, but we shall not pursue this aspect here. The control problem is given by (1.1), (1.2) with cost function J defined by

$$J(y,u) := \frac{1}{2}\int_{Q_o} |y - z|^2\,dxdt + \frac{\alpha}{2}\int_{Q_c} |u|^2\,dxdt, \tag{6.1}$$

where $Q_c := \Omega_c \times (0,T)$ and $Q_o := \Omega_o \times (0,T)$, with Ω_c and Ω_o subsets of $\Omega = (0,1)^2$ denoting the control and observation volumes, respectively. In our example, $T = 1$, $U := L^2(Q_c)$, $\nu = \frac{1}{\mathrm{Re}} = 400$, and B is the indicator function of Q_c. The results for Newton's method will be compared to those of the gradient algorithm, which we recall here for the sake of convenience.

ALGORITHM 6.1 (gradient algorithm).

1. *Set* $k = 0$ *and choose* u^0,
2. *set* $d := -\hat{J}'(u^k)$ *and compute*

$$\rho^* = \arg\min_{\rho>0} I(\rho) := \hat{J}(u^k + \rho d),$$

3. *set*

$$u^{k+1} = u^k + \rho^* d,$$

4. *set* $k = k + 1$ *and goto* 2.

Given a control u, the evaluation of the gradient of J at a point u amounts to solving (1.2) for the state y and (3.7) for the adjoint variable λ. Implementing a stepsize rule to determine an approximation of ρ^* is numerically expensive as every evaluation of the functional J at a control u requires solving the instationary Navier–Stokes equations with right-hand side Bu.

In the numerical example presented below, we use the following procedure to compute an approximation to the step size ρ^*. For a given search direction $d \in U$ we insert the linearization of the mapping $\rho \mapsto y(u + \rho d)$ at $\rho = 0$,

$$y(u + \rho d) \doteq y(u) + \rho y'(u)d,$$

into the cost functional J. This results in the quadratic approximation

$$I_1(\rho) := J(y(u) + \rho y'(u)d, u + \rho d)$$

of the functional $I(\rho)$. Now we use the unique root

$$\rho_1^* = \frac{-\langle \hat{J}'(u), d\rangle_U}{\alpha |d|_U^2 + |v|^2_{L^2(Q_o)}} \tag{6.2}$$

of the equation $I_1'(\rho) = 0$ as approximation of ρ^*, with v given by (2.4) with $f = Bd$ and $v_0 = 0$.

Altogether, every iteration of the gradient algorithm amounts to solving the nonlinear Navier–Stokes equations forward in time and the associated adjoint equations backward in time for the computation of the gradient, and to solving linearized Navier–Stokes equations forward in time for the step size proposal. For a detailed discussion of step size rules for the gradient algorithm, see [HH, Algorithm 4.7.1].

The inner iteration of Newton's method is performed by the conjugate gradient method, the choice of which is justified in a neighborhood of a local solution u^* of the optimal control problem by the positive definiteness of $\hat{J}''(u^*)$, provided the desired state z is sufficiently close to the optimal state $y(u^*)$.

For the numerical tests the target flow z is given by the Stokes flow with boundary condition $z_1 = 1$ in tangential direction (Figure 6.1), and the uncontrolled flow is the Navier–Stokes flow at Re=400 with the same boundary values as z; see Figure 6.3, top left. The termination criterion for the jth iterate u_j^k in the conjugate gradient method is chosen as

$$\frac{|\hat{J}''(u^k)\delta u_j^k + \hat{J}'(u^k)|}{|\hat{J}'(u^0)|} \le \min\left\{ \left(\frac{|\hat{J}'(u^k)|}{|\hat{J}'(u^0)|} \right)^{\frac{3}{2}}, 10^{-2} \frac{|\hat{J}'(u^k)|}{|\hat{J}'(u^0)|} \right\} \text{ or } j \ge 50.$$

The initialization for Newton's method was $u^0 := 0$.

The discretization of the Navier–Stokes equations, its linearization and adjoint, was carried out by using parts of the code developed by Bänsch in [BA], which is based on Taylor–Hood finite elements for spatial discretization. As time step size we took $\delta t = .00625$, which resulted in 160 grid points for the time grid and 545 pressure and

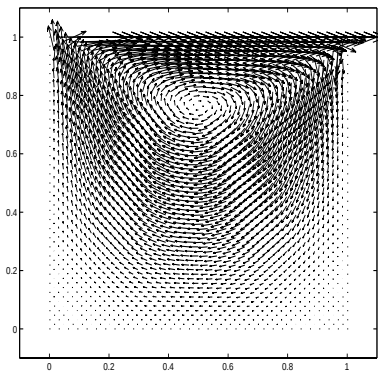


FIG. 6.1. *Control target: Stokes flow in the cavity.*

TABLE 6.1
Performance of Newton's method.

Iteration	CG-steps	$\frac{\lvert\hat{J}'(u)\rvert}{\lvert\hat{J}'(u^0)\rvert}$	$\frac{\lvert\delta u^k\rvert_U}{\lvert\delta u^{k-1}\rvert_U}$	$\hat{J}(u^k)$
1	-	1.e0	-	1.196202e-2
2	13	3.358825e-1	1.	3.226486e-3
3	11	5.058497e-2	0.492	1.617913e-3
4	18	8.249029e-3	0.422	1.482032e-3
5	17	1.409278e-4	0.079	1.480533e-3
6	19	4.686819e-6	0.032	1.480534e-3

2113 velocity nodes for the spatial discretization. All computations were performed on a DEC-ALPHA^{TM}station 500.

We now present the results for $\Omega_c = \Omega_o = (0,1)^2$ and $\alpha = 10^{-2}$. Table 6.1 confirms superlinear convergence of the inexact Newton method. To achieve the same accuracy as Newton's method the gradient algorithm requires 96 iterations. The computing time with Newton's method is approximately 45 minutes, whereas the gradient method requires 110 minutes. This demonstrates the superiority of Newton's method over the gradient algorithm for this example. For larger values of α and coarser time and space grids the difference in computing time is less drastic. In fact, this difference increases with decreasing α and increasing mesh refinement. As expected, a significant amount of computing time is spent for read-write actions of the variables to the hard disc in the subproblems.

In Figure 6.2 the evolution of the cost functional is documented. It can be observed that Newton's method (left) tends to overestimate the control in the first iteration step, whereas the gradient algorithm (right) approximates the optimal control from below. Graphically there is no significant change after the second iteration for Newton's method. These comments hold for quite a wide range of values for α.

In Figure 6.3 a snapshot of the uncontrolled flow at $t = 1$ together with a snapshot of the control action at t=0.75 and two zooms are presented. As one can see, two major vortices develop; the one in the upper left corner (Figure 6.3, bottom left) is responsible for tearing back the vortex of the uncontrolled flow (Figure 6.3, top left), while the second control vortex (Figure 6.3, bottom right) pushes back this vortex towards the vortex of the target flow (Figure 6.1). The controlled flow at t=1 is optically nearly indistinguishable from the target flow.

In this example the observation volume Ω_o and the control volume Ω_c each cover

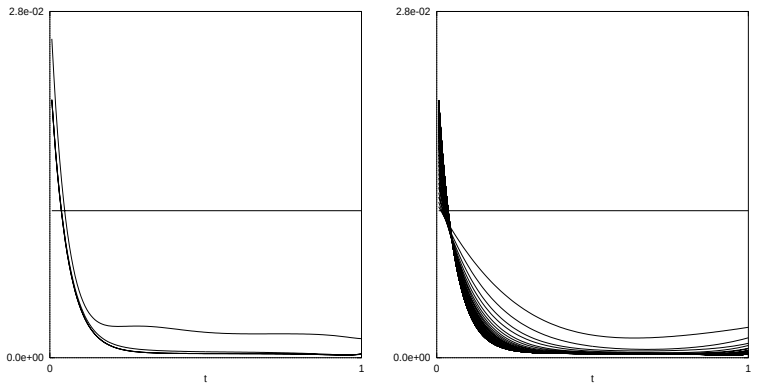


FIG. 6.2. *Newton's method (left,* 6 *iterations) versus gradient algorithm (right),* Re = 400, $\alpha = 10^{-2}$*: Evolution of cost functional for relative accuracy* $= 1.d - 5$.

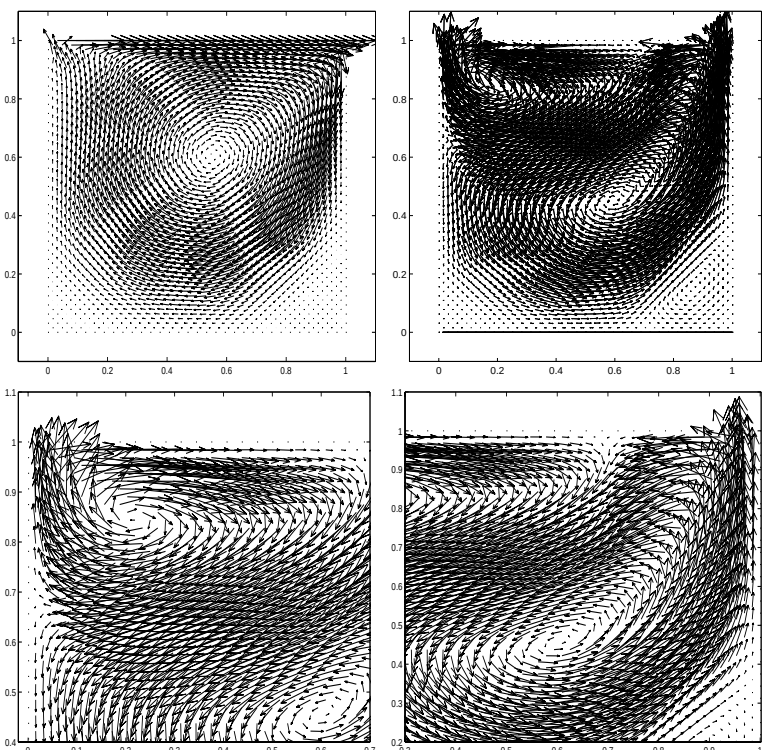


FIG. 6.3. *Top left: Uncontrolled flow at* Re = 400. *Top right: Control force at* t=0.75 *for* $\alpha = 10^{-2}$. *Bottom left: Control force, zoom on* $[0, 0.7] \times [0.4, 1]$. *Bottom right: Zoom on* $[0.3, 1] \times [0.2, 1]$.

the whole spatial domain. From the practical point of view this is not feasible. However, from the numerical standpoint this is a complicated situation, since the inhomogeneities in the primal and adjoint equations are large. Further numerical examples with different observation and control domains can be found in [HH, Chapter 4.7].

Acknowledgment. The authors would like to thank an anonymous referee for a careful reading of the first version of the paper and for many helpful comments.

REFERENCES

[AM] W. Alt and K. Malanowski, *The Lagrange–Newton method for state constrained optimal control problems*, Comput. Optim. Appl., 4 (1995), pp. 217–239.

[AT] F. Abergel and R. Temam, *On some control problems in fluid mechanics*, Theoret. Comput. Fluid Dynamics, 1 (1990), pp. 303–325.

[BA] E. Bänsch, *An adaptive finite element strategy for the three-dimensional time-dependent Navier–Stokes equations*, J. Comput. Appl. Math., 36 (1991), pp. 3–28.

[B] M. Berggren, *Numerical solution of a flow-control problem: Vorticity reduction by dynamic boundary action*, SIAM J. Sci. Comput., 19 (1998), pp. 829–860.

[BMT] T. R. Bewley, P. Moin, and R. Temam, *DNS–based predictive control of turbulence: An optimal benchmark for feedback algorithms*, J. Fluid Mech., to appear.

[CF] P. Constantin and C. Foias, *Navier–Stokes Equations*, The University of Chicago Press, Chicago, 1989.

[CHK] H. Choi, M. Hinze, and K. Kunisch, *Instantaneous control of backward-facing-step flows*, Appl. Numer. Math., 31 (1999), pp. 133–158.

[CTMK] H. Choi, R. Temam, P. Moin, and J. Kim, *Feedback control for unsteady flow and its application to the stochastic Burgers equation*, J. Fluid Mech., 253 (1993), pp. 509–543.

[DL5] R. Dautray and J. L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol. 5, Springer-Verlag, Berlin, 1992.

[DHV] J. E. Dennis, M. Heinkenschloss, and L. N. Vicente, *Trust-region interior-point SQP algorithms for a class of nonlinear programming problems*, SIAM J. Control Optim., 36 (1998), pp. 1750–1794.

[FGH] A. V. Fursikov, M. D. Gunzburger, and L. S. Hou, *Boundary value problems and optimal boundary control for the Navier–Stokes system: The two-dimensional case*, SIAM J. Control Optim., 36 (1998), pp. 852–894.

[G] R. Glowinski, *Numerical Methods for Nonlinear Variational Problems*, Springer-Verlag, Berlin, 1984.

[GB] O. Ghattas and J. J. Bark, *Optimal control of two- and three-dimensional incompressible Navier–Stokes flows*, J. Comput. Phys., 136 (1997), pp. 231–244.

[GR] A. Griewank, *The local convergence of Broyden-like methods in Lipschitzian problems in Hilbert spaces*, SIAM J. Numer. Anal., 24 (1987), pp. 684–705.

[GM] M. D. Gunzburger and S. Manservisi, *The velocity tracking problem for Navier–Stokes flows with bounded distributed controls*, SIAM J. Control Optim., 37 (1999), pp. 1913–1945.

[GT] H. Goldberg and F. Tröltzsch, *Second-order sufficient optimality conditions for a class of nonlinear parabolic control problems*, SIAM J. Control Optim., 31 (1993), pp. 1007–1025.

[H] M. Heinkenschloss, *Formulation and analysis of a sequential quadratic programming method for the optimal Dirichlet boundary control of Navier–Stokes flow*, in Optimal Control: Theory, Algorithms, and Applications, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 178–203.

[HH] M. Hinze, *Optimal and Instantaneous Control of the Instationary Navier–Stokes Equations*, Habilitation Thesis, Technische Universität Berlin, Berlin, Germany, 2000.

[HK] M. Hinze and A. Kauffmann, *Reduced Order Modelling and Suboptimal Control of a Solid Fuel Ignition Model*, preprint 636/99, Technische Universität Berlin, Berlin, Germany, 1999.

[HKK] M. Hinze and K. Kunisch, *Control strategies for fluid flows—optimal versus suboptimal control*, in ENUMATH 97, H. G. Bock et al., eds., World Scientific, Singapore, pp. 351–358.

[IK] K. Ito and K. Kunisch, *Augmented Lagrangian–SQP methods for nonlinear optimal control problems of tracking type*, SIAM J. Control Optim., 34 (1996), pp. 874–891.

[IR] K. Ito and S. S. Ravindran, *Optimal control of thermally convected fluid flows*, SIAM J. Sci. Comput., 19 (1998), pp. 1847–1869.

[KA] A. Kauffmann, *Optimal Control of the Solid Fuel Ignition Model*, Ph.D. thesis, Fachbereich Mathematik, Technische Universität Berlin, Berlin, Germany, 1998.

[KS] K. Kunisch and E. W. Sachs, *Reduced SQP methods for parameter identification problems*, SIAM J. Numer. Anal. 29 (1992), pp. 1793–1820.

[K] F.-S. Kupfer, *An infinite-dimensional convergence theory for reduced SQP methods in Hilbert space*, SIAM J. Optim., 6 (1996), pp. 126–163.

[KS1] C. T. Kelley and E. W. Sachs, *Quasi-Newton methods and unconstrained optimal control problems*, SIAM J. Control Optim., 25 (1987), pp. 1503–1516.

[KS2] C. T. Kelley and E. W. Sachs, *Mesh independence of the gradient projection method for optimal control problems*, SIAM J. Control Optim., 30 (1992), pp. 477–493.

[KV] K. Kunisch and S. Volkwein, *Control of Burgers equation by a reduced order approach using proper orthogonal decomposition*, J. Optim. Theory Appl., 102 (1999), pp. 345–371.

[LA] M. Laumen, *Newton's method for a class of optimal shape design problems*, SIAM J. Optim., 10 (2000), pp. 503–533.

[L] P. L. Lions, *Mathematical Topics in Fluid Mechanics* I, Clarendon, Oxford University Press, New York, 1996.

[LT] H. V. Ly and H. T. Tran, *Modelling and control of physical processes using proper orthogonal decomposition*, Math. Comput. Modelling, 33 (2001), pp. 223–236.

[NT] P. Neittaanmäki and D. Tiba, *Optimal Control of Nonlinear Parabolic Systems*, Marcel Dekker, New York, 1994.

[T] R. Temam, *Navier–Stokes Equations*, North-Holland, Amsterdam, 1979.

[T1] R. Temam, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, 2nd ed., Springer-Verlag, Berlin, 1997.

[V] S. Volkwein, *Mesh-Independence of an Augmented Lagrangian-SQP Method in Hilbert Spaces and Control Problems for the Burgers Equation*, Ph.D. thesis, Fachbereich Mathematik, Technische Universität Berlin, Berlin, Germany, 1997.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 3, pp. 947–968

GENERAL ORTHONORMAL BASES FOR ROBUST IDENTIFICATION IN H_∞*

HÜSEYIN AKÇAY†

Abstract. In this paper, the problem of system identification in H_∞ with general orthonormal basis functions is investigated. A two-stage algorithm is shown to be robust, provided that the number of basis elements as a function of the amount of data does not increase faster than $O(N^{\frac{2}{5}})$. Worst-case identification error bounds in the H_∞ norm are derived. The algorithm also works on nonuniformly spaced frequency response measurements. An example is provided to illustrate the application of the algorithm.

Key words. robust identification, orthonormal basis functions, H_∞

AMS subject classifications. 93A30, 42C15, 42A24

PII. S0363012999360397

1. Introduction.

Consider a linear-time-invariant, single-input/single-output, discrete-time system with impulse response $g(k)$. It is assumed that this system is ℓ_2 bounded-input/bounded-output stable so that the associated power-series representation

$$G(z) = \sum_{k=0}^{\infty} g(k) z^k \tag{1.1}$$

lies in $H_\infty(\mathbf{D})$. *Note that the definition of the z-transform is such that the stability corresponds to having no poles in the closed unit disk.* Here $H_\infty(\mathbf{D})$ denotes the space of bounded analytic functions in the open unit disk $\mathbf{D}$. Let $C(\mathbf{T})$ denote the set of continuous functions on $\mathbf{T}$, the unit circle, and let $A(\mathbf{D}) = H_\infty(\mathbf{D}) \cap C(\mathbf{T})$. We further require $G \in A(\mathbf{D})$. In the series expansion, we have employed a set of orthonormal functions $1, z, z^2, \ldots$, where orthogonality is with respect to the inner product

$$\langle f, g \rangle \triangleq \frac{1}{2\pi} \int_0^{2\pi} f(e^{i\omega})\ \overline{g(e^{i\omega})}\, d\omega.$$

Although power-series representation or *finite-impulse response* modelling is suitable for most applications, it fails to be successful when the number of basis coefficients to be estimated from the data becomes very large, especially in the modelling of large flexible structures. This increase in parameter dimension to maintain a certain level of accuracy over a broad range of transfer functions is a drawback of finite-impulse response models in robust controller design. Alternatively, we can write (1.1) in a generalized form as

$$G(z) = \sum_{k=0}^{\infty} a_k B_k(z), \tag{1.2}$$

*Received by the editors August 19, 1999; accepted for publication (in revised form) April 10, 2001; published electronically November 15, 2001. This work was supported by the Alexander von Humboldt Foundation.

http://www.siam.org/journals/sicon/40-3/36039.html

†Department of Electrical and Electronics Engineering, Anadolu University, 26470 Eskişehir, Turkey (huakcay@anadolu.edu.tr).

where $\{B_0, B_1, B_2, \ldots\}$ is an orthonormal set of rational basis functions in $H_2(\mathbf{D})$. Here $L_2(\mathbf{T})$ is the space of square integrable functions on $\mathbf{T}$, and $H_2(\mathbf{D})$ denotes its intersection with the set of analytic functions on $\mathbf{D}$. The choice of basis poles reflects the desire to adapt basis functions to the specific properties of the system. With appropriately chosen basis poles, the convergence rate of the series expansion in (1.2) can be very fast, and hence the number of basis coefficients to be estimated can become very small.

The undermodelling-induced component of the estimation error can be reduced significantly in comparison to the use of a finite-impulse response model structure in (1.1) by employing the well-known *Laguerre* and *Kautz* bases [22, 21]. The use of the Laguerre and Kautz bases for approximation and identification of linear time-invariant dynamics has been investigated in [17, 30, 24, 32, 40, 14, 41, 42, 12]. More recently, in [43, 44] the *rational wavelet* bases have been suggested to robustly estimate linear-time-invariant infinite-dimensional dynamics.

Recently, in [19, 38, 39], the general orthonormal basis functions, which generalize the Laguerre and Kautz bases and to some extent incorporate the dynamics of the underlying system, were introduced and used in a Hilbert-space setting, mostly for time-domain identification. These basis functions are obtained from the rational orthonormal basis functions considered in detail in [28, 29, 2, 4], which are defined by a choice of numbers $z_k \in \mathbf{D}$, $k = 0, 1, \ldots$, as

$$B_k(z) \triangleq \frac{\sqrt{1-|z_k|^2}}{1-\overline{z_k}\,z}\,\psi_k(z), \qquad \psi_k(z) \triangleq \prod_{j=0}^{k-1} \frac{z-z_j}{1-\overline{z_j}\,z}, \quad \psi_0(z) \triangleq 1, \tag{1.3}$$

with the restriction

$$z_{j+nm} = z_j, \qquad j = 1, 2, \ldots, n, \quad m = 1, 2, \ldots, \tag{1.4}$$

where n is a fixed number, $z_0 = 0$, and $1/\overline{z_1}, \ldots, 1/\overline{z_n}$ denote the chosen basis poles in complex conjugate pairs. The general orthonormal basis functions are constructed from an initial set of orthonormal functions by repeated multiplication of an all-pass function that acts as a generalized shift. The rational orthonormal basis functions in (1.3) are *complete* in the spaces $H_p(\mathbf{D})$ $(1 \le p < \infty)$ and $A(\mathbf{D})$ if and only if $\sum_{k=1}^{\infty}(1-|z_k|) = \infty$ [2].

In a series of papers [2, 3, 4, 7, 10, 9], completeness, approximation, and identification properties of the basis functions defined by (1.3) and their continuous-time versions have been investigated. In particular, it was established in [2] that by using a min-max criterion, provided that the basis functions are complete and the model order increases at most linearly with the amount of data, these bases lead to robust estimators for which error bounds in the H_∞ norm can be explicitly quantified. The purpose of the current paper is to establish a similar robust estimation result for the general orthonormal basis functions defined by (1.3)–(1.4), using a two-stage scheme which first appeared in [18].

In this paper, we will consider the problem of system identification in H_∞ initiated by [18] in the general orthonormal basis set-up. The system identification with H_∞ criterion has received a growing interest since the appearance of the H_∞ formulation of robust control [46, 18, 16, 34, 25, 27, 15, 6]. Given N noise-corrupted samples of the frequency response function

$$e_k = G(e^{i\omega_k}) + \nu_k, \qquad k = 1, \ldots, N, \tag{1.5}$$

at the frequencies $\omega_k \in [0, 2\pi)$, $k = 1, \ldots, N$, where ν is the frequency response measurement noise bounded in amplitude by ε, the objective is to find an identification algorithm which maps the data e_k, $k = 1, \ldots, N$, to an identified model $\widehat{G}_N \in A(\mathbf{D})$ such that for all $G \in A(\mathbf{D})$

$$\lim_{\substack{\varepsilon \to 0 \\ N \to \infty}} \sup_{\|\nu\|_\infty \le \varepsilon} \|\widehat{G}_N - G\|_\infty = 0. \tag{1.6}$$

An algorithm that satisfies (1.6) and does not use the prior knowledge of ε and the uncertainty set in which the system lies is called *robustly convergent*. It should be noted that the latter requirement restricts the nature of identification algorithms to be used. If this restriction is removed, a large class of nonlinear *tuned* algorithms that satisfy (1.6) exists. This problem formulation is, in a technical sense, somewhat different from the usual formulation adopted in [18]. The main difference is that in this definition we have not taken a supremum of the identification error over the set of unknown systems. This formulation is similar to the point of view taken in [25].

Several robustly convergent nonlinear algorithms have been proposed for solving this problem of system identification in H_∞ [18, 16, 25, 6]. These algorithms share a common two-stage structure. In the first stage, a (stable and unstable) finite-impulse response model is linearly estimated from the frequency response data. In the second stage, the identified model is obtained by solving a Nehari problem. We will also adapt the same two-stage strategy. In a very recent paper [37], Szabó, Bokor, and Schipp have obtained some general results for the same identification problem. However, our results and algorithm are *quite different from and independent of* those in [37].

In this paper, we obtain explicit worst-case identification error bounds, provided that the number of general orthonormal basis functions used for the estimation does not increase too fast with the amount of data. In comparison to the min-max algorithms in [2], worst-case error bounds are large for small amounts of data and, for a given amount of data, the number of general orthonormal basis functions that could be used for the estimation is rather restricted.

In spite of these weaknesses, a couple of distinct features of our approach deserve to be mentioned. First, due to the linearity of the first stage, the algorithm of this paper is easier to implement and analyze than the min-max algorithms. In this stage, our algorithm is also computationally more efficient. When model complexity is restricted, the second stage could be omitted since the linear estimate diverges very slowly in the worst-case. The second and most compelling reason is that frequently low complexity models are desired for the purpose of subsequent controller design, and by iteratively updating basis poles it may be possible to significantly reduce the undermodelling-induced component of the estimation error.

Now we will briefly describe the contents of this paper. In section 2, uniform approximation of complex-valued continuous functions, by the general orthonormal basis functions defined by (1.3)–(1.4) and the complementary general orthonormal functions, is studied. The basis functions spanning the orthogonal complement of $H_2(\mathbf{D})$ are defined by a choice of numbers $x_k \in \mathbf{D}$, $k = 1, 2, \ldots$, as

$$B_{-k}(z) \triangleq \frac{\sqrt{1 - |x_k|^2}}{z - x_k}\, \phi_k(z), \qquad \phi_{k-1}(z) \triangleq \prod_{j=1}^{k} \frac{1 - \overline{x_j}\, z}{z - x_j}, \quad \phi_0(z) \triangleq 1. \tag{1.7}$$

Then the complementary general orthonormal functions are obtained from (1.7) by setting

$$x_k = z_k \qquad \text{for all } k. \tag{1.8}$$

In section 3, an overdetermined system of linear equations is studied. The purpose of this section is to link the approximation results derived in section 2 to the identification problem. In section 4, the main result of this paper is presented. In section 5, an example is given to illustrate the use of the general orthonormal basis functions defined by (1.3)–(1.4) in an iterative-identification scheme. Section 6 concludes the paper.

We consider only single-input/single-output systems. This is not at all a restriction. The results extend with no modifications to multi-input/multi-output systems. They are also applicable to continuous-time systems, since the bilinear mapping

$$s \triangleq \lambda \frac{1-z}{1+z}, \qquad \lambda > 0,$$

preserves the supremum norms between the space of functions which are analytic and bounded on the upper half plane and $H_\infty(\mathbf{D})$. The details can be found in [1].

2. Approximation of continuous functions. In this section, we derive error bounds in the L_∞ norm for the approximation of continuous functions on $\mathbf{T}$ by a particular weighted Fourier series. These bounds will be used in the analysis of a proposed algorithm that solves the worst-case identification problem posed in section 1.

Let $\mathcal{S}_k f$ denote the partial sums of the Fourier series of $f \in C(\mathbf{T})$ with respect to the orthonormal system (1.3) and (1.7) defined by

$$\mathcal{S}_k f(e^{i\theta}) \triangleq \sum_{j=-k}^{k} \langle f, B_j \rangle \, B_j(e^{i\theta}). \tag{2.1}$$

The (block) Cesàro means of f is defined as

$$\mathcal{F}_m f \triangleq \frac{1}{m+1} \sum_{j=0}^{m} \mathcal{S}_{nj} f. \tag{2.2}$$

We take the (block) de la Vallée Poussin estimate of f defined by

$$\mathcal{V}_m f \triangleq 2\mathcal{F}_{2m+1} f - \mathcal{F}_m f \tag{2.3}$$

as the L_∞-approximant of f. This estimate can be written as

$$\mathcal{V}_m f = \sum_{k=-n(2m+1)}^{n(2m+1)} \chi_m(k) \, \langle f, B_k \rangle \, B_k$$

for a nonnegative symmetric window function χ_m:

$$\begin{aligned} \chi_m(k) &\triangleq 1, \quad k = 0, \ldots, n(m+1), \\ \chi_m\left(n(m+j)+k\right) &\triangleq 1 - \frac{j}{m+1}, \quad k = 1, \ldots, n, \ j = 1, \ldots, m, \\ \chi_m(k) &\triangleq 0, \quad k > n(2m+1). \end{aligned} \tag{2.4}$$

The de la Vallée Poussin estimator was introduced to the field of worst-case identification by Partington [33]. It is possible to use windows other than χ_m provided that the chosen window function satisfies certain regularity conditions. The choice of window functions is discussed in [37].

We want to show that $\mathcal{V}_m f \to f$ uniformly on $\mathbf{T}$ for all continuous functions f. To this end, we first derive a simple expression for the so-called Dirichlet kernel

$$D_{k,k'}(s,\theta) \triangleq \sum_{j=-k'}^{k} \overline{B_j(e^{is})}\, B_j(e^{i\theta}). \tag{2.5}$$

We undertake this as follows.

LEMMA 2.1. *Consider the rational basis functions defined by* (1.3) *and* (1.7). *Let* $D_{k,k'}(s,\theta)$ *be as in* (2.5). *Then*

$$D_{k,k'}(s,\theta) = e^{i\mu_{k,k'}(s,\theta)}\ \frac{\sin\left(\lambda_{k,k'}(s,\theta)\right)}{\sin\left(\frac{\theta-s}{2}\right)}, \tag{2.6}$$

where

$$\begin{aligned} \mu_{k,k'}(s,\theta) &\triangleq \frac{1}{2}\int_s^\theta \left(\sum_{j=0}^{k} |B_j(e^{iy})|^2 - 1 - \sum_{j=-k'}^{-1} |B_j(e^{iy})|^2 \right) dy, \\ \lambda_{k,k'}(s,\theta) &\triangleq \frac{1}{2}\int_s^\theta \sum_{j=-k'}^{k} |B_j(e^{iy})|^2\, dy. \end{aligned} \tag{2.7}$$

Proof. See [10]. This lemma can also be proved using the results in [36] after applying some algebra. □

The partial sums in (2.1) can be written as

$$\mathcal{S}_k f(e^{i\theta}) = \frac{1}{2\pi}\int_{\theta-\pi}^{\theta+\pi} f(e^{is})\, D_{k,k}(s,\theta)\, ds. \tag{2.8}$$

Hence

$$\|\mathcal{S}_k\| \triangleq \sup_{\|f\|_\infty \le 1} \|\mathcal{S}_k f\|_\infty = \sup_\theta \frac{1}{2\pi}\int_0^{2\pi} |D_{k,k}(s,\theta)|\, ds.$$

The right-hand side is called the Lebesgue constant for the basis $\{B_k\}$. Thus the Fourier series (2.1) converges uniformly for every $f \in C(\mathbf{T})$ if and only if the Lebesgue constants are uniformly bounded, and in this case $\{B_k\}$ is said to form a basis for $C(\mathbf{T})$.

It is known [31] that no uniformly bounded orthonormal system can form a basis for the space $C(\mathbf{T})$. This result applies to the disk algebra with the basis functions defined by (1.3) as well. Notice that the uniform boundedness of the basis defined by (1.3) and (1.7) is equivalent to

$$\sup_n \{|z_n|, |x_n|\} \triangleq r < 1. \tag{2.9}$$

Although unbounded, the rational wavelets considered in [43, 44] cannot form a basis for $A(\mathbf{D})$ as well [8]. It is unknown whether there exists an (unbounded) rational basis defined by (1.3) such that every function in $A(\mathbf{D})$ has a convergent Fourier series with respect to this basis. However, Fourier series of every Dini–Lipschitz continuous function with respect to the uniformly bounded orthonormal bases defined by (1.3) and (1.7) always converge uniformly on $\mathbf{T}$ [10].

The situation is quite different if one considers orthonormal systems other than the rational system defined by (1.3). There are certainly orthonormal bases for $H_2(\mathbf{D})$ which consists of rational functions (even polynomials) and also form bases in the disk algebra. See, for example, the construction in [45].

Having seen that the Fourier series of a continuous function may not necessarily converge uniformly with respect to a given (uniformly) bounded basis, we now study the convergence properties of the windowed Fourier series defined in (2.2) with respect to the general orthonormal basis functions. Recall that the general orthonormal basis functions are obtained from the rational functions in (1.3) and (1.7) by enforcing (1.8) and the periodicity condition (1.4).

The Cesàro means of f defined in (2.2) can be written as

$$\mathcal{F}_m f(e^{i\theta}) = \frac{1}{2\pi} \int_{\theta-\pi}^{\theta+\pi} f(e^{is})\, \sigma_m(s,\theta)\, ds, \tag{2.10}$$

where $\sigma_m(s,\theta)$ is the Fejér kernel defined by

$$\sigma_m(s,\theta) \triangleq \frac{1}{m+1} \sum_{k=0}^{m} D_{nk,nk}(s,\theta). \tag{2.11}$$

Assuming that the basis poles are cyclically repeated according to (1.4), we derive a formula for the Fejér kernel in the following.

LEMMA 2.2. *Consider the general orthonormal basis functions defined by* (1.3), (1.7), (1.4), *and* (1.8). *Let* $\sigma_m(s,\theta)$ *be as in* (2.11). *Then*

$$\sigma_m(s,\theta) = \alpha_m(s,\theta) + \beta(s,\theta)\, \frac{1}{m+1}\, \frac{\sin^2\left((m+1)\,\dfrac{\zeta}{2}\right)}{\sin^2\left(\dfrac{\zeta}{2}\right)}, \tag{2.12}$$

where

$$\alpha_m(s,\theta) \triangleq \left\{ \cos\left(m\,\frac{\zeta}{2}\right) - \cos\left(\frac{\eta}{2}\right) \cos\left((m+1)\,\frac{\zeta}{2}\right) \frac{\sin\left(\dfrac{\zeta}{2}\right)}{\sin\left(\dfrac{\eta}{2}\right)} \right\} \frac{\sin\left((m+1)\,\dfrac{\zeta}{2}\right)}{(m+1)\,\sin\left(\dfrac{\zeta}{2}\right)},$$

$$\beta(s,\theta) \triangleq \cos\left(\frac{\zeta}{2}\right) \cos\left(\frac{\eta}{2}\right) \frac{\sin\left(\frac{\zeta}{2}\right)}{\sin\left(\frac{\eta}{2}\right)}, \tag{2.13}$$

$$\begin{aligned} \zeta &\triangleq \sum_{k=1}^{n} \int_{s}^{\theta} \left|B_k(e^{iy})\right|^2 dy, \\ \eta &\triangleq \theta - s. \end{aligned} \tag{2.14}$$

Proof. Since $z_j = x_j$ and $|\psi_j(e^{i\theta})| = |\phi_j(e^{i\theta})| = 1$, note that $|B_j(e^{i\theta})| = |B_{-j}(e^{i\theta})|$ for all j and θ. Thus in (2.7), $\mu_{k,k}(s,\theta) = 0$ for all s, θ, and k. Due to (1.4), we also

have

$$\lambda_{nk,nk}(s,\theta) = \frac{\theta - s}{2} + \int_s^\theta \sum_{j=1}^{nk} \left|B_j(e^{iy})\right|^2 \, dy$$
$$= \frac{\eta}{2} + k\,\zeta.$$

Hence from Lemma 2.1,

$$D_{nk,nk}(s,\theta) = \frac{\sin\left(\frac{\eta}{2} + k\,\zeta\right)}{\sin\left(\frac{\eta}{2}\right)}, \qquad k \geq 0. \tag{2.15}$$

Plugging (2.15) into (2.11), we get (2.12) from several applications of the identities:

$$\sin\theta = \frac{e^{i\theta} - e^{-i\theta}}{2i},$$

$$\sum_{k=0}^{m-1} e^{ik\theta} = \frac{e^{im\theta} - 1}{e^{i\theta} - 1},$$

$$\sin(x+y) = \sin x \cos y + \cos x \sin y. \qquad \square$$

The Cesàro means of f defined in (2.2) and the de la Vallée Poussin estimate of f converge uniformly to f. This will follow from the following lemma.

LEMMA 2.3. *Consider the general orthonormal basis functions defined by* (1.3), (1.7), (1.4), *and* (1.8). *Let* $\sigma_m(s,\theta)$ *be as in* (2.12). *Then for all* $m \geq 1$,

$$\sup_\theta \frac{1}{2\pi} \int_{-\pi}^{\pi} |\sigma_m(s,\theta)| \, ds \leq C_1 + c_m, \tag{2.16}$$

where

$$C_1 \triangleq \frac{4 + 6\pi + 4\ln n}{(1-r)^2},$$
$$c_m \triangleq \frac{28n}{(1-r)^2} \, \frac{\ln(m+1)}{m+1}. \tag{2.17}$$

Proof. From Lemma 2.2, by a change of variables $s = \theta + t$ we have

$$\int_{-\pi}^{\pi} |\sigma_m(s,\theta)| \, ds \leq \int_{-\pi}^{\pi} |\alpha_m(\theta + t, \theta)| \, dt$$
$$+ \int_{-\pi}^{\pi} \frac{|\beta(\theta + t, \theta)|}{m+1} \underbrace{\frac{\sin^2\left((m+1)\,\frac{\zeta(t)}{2}\right)}{\sin^2\left(\frac{\zeta(t)}{2}\right)}}_{[\Upsilon(\zeta(t))]^2} \, dt. \tag{2.18}$$

To bound the first integral on the right-hand side, note from (2.14) that

$$\frac{\zeta}{\eta} = \frac{\displaystyle\sum_{k=1}^{n} \int_{\theta+t}^{\theta} \left|B_k(e^{iy})\right|^2 \, dy}{\theta - (\theta + t)} \leq \sum_{k=1}^{n} \|B_k\|_\infty^2,$$

where the terms $\|B_k\|_\infty^2$ are bounded as

$$\|B_k\|_\infty^2 = \max_t \frac{1 - |z_k|^2}{|1 - \overline{z_k}\, e^{it}|^2} \leq \frac{1+r}{1-r}. \tag{2.19}$$

Hence

$$\frac{\zeta}{\eta} \leq \frac{2n}{1-r}. \tag{2.20}$$

We also have

$$\frac{\zeta}{\eta} \geq \sum_{k=1}^{n} \min_t |B_k(e^{it})| \geq n\, \frac{1-r}{1+r}. \tag{2.21}$$

Thus from (2.20) and (2.21),

$$\frac{n(1-r)}{2} \leq \frac{d\zeta}{d\eta} \leq \frac{2n}{1-r}. \tag{2.22}$$

Next from (2.20), the inequalities $|\sin x| \leq |x|$ for all x, and

$$\frac{\sin x}{x} \geq \frac{2}{\pi}, \qquad |x| \leq \frac{\pi}{2}, \tag{2.23}$$

we get

$$\frac{\left|\sin\left(\dfrac{\zeta}{2}\right)\right|}{\left|\sin\left(\dfrac{\eta}{2}\right)\right|} \leq \frac{\pi\,\zeta}{2\eta} \leq \frac{\pi\, n}{1-r}, \qquad |\eta| \leq \pi. \tag{2.24}$$

Hence

$$\int_{-\pi}^{\pi} |\alpha_m(\theta + t, \theta)|\, dt \leq \frac{1}{m+1} \left(1 + \frac{\pi n}{1-r}\right) \int_{-\pi}^{\pi} |\Upsilon(\zeta(t))|\, dt. \tag{2.25}$$

By a change of the variables $\zeta = \int_{\theta+t}^{\theta} \sum_{k=1}^{n} |B_k(e^{is})|^2\, ds$ and $\eta = -t$, we have from (2.22) the following:

$$\begin{aligned} \int_{-\pi}^{\pi} |\Upsilon(\zeta(t))|\, dt &= \int_{\zeta(\pi)}^{2\pi n + \zeta(\pi)} |\Upsilon(\zeta)| \left(\frac{\mathrm{d}\zeta}{\mathrm{d}\eta}\right)^{-1} d\zeta \\ &\leq \frac{2}{n(1-r)} \int_{\zeta(\pi)}^{2\pi n + \zeta(\pi)} |\Upsilon(\zeta)|\, d\zeta \\ &= \frac{2}{1-r} \int_{-\pi}^{\pi} |\Upsilon(\zeta)|\, d\zeta, \end{aligned} \tag{2.26}$$

where the first two (in)equalities follow from (2.22) and

$$\int_{\theta}^{\theta+2\pi} |B_k(e^{is})|^2\, ds = 2\pi \qquad \text{for all } \theta \text{ and } k, \tag{2.27}$$

and the last equality from the fact that $|\Upsilon(\zeta)|$ is a periodic function with period 2π. The integral in (2.26) can be bounded from (2.23) as follows:

$$\begin{aligned}\int_{-\pi}^{\pi} |\Upsilon(\zeta)|\, d\zeta &= 2\left\{ \int_0^{\frac{\pi}{m+1}} \frac{\left|\sin\left((m+1)\,\frac{\zeta}{2}\right)\right|}{\sin\left(\frac{\zeta}{2}\right)}\, d\zeta + \int_{\frac{\pi}{m+1}}^{\pi} \frac{\left|\sin\left((m+1)\,\frac{\zeta}{2}\right)\right|}{\sin\left(\frac{\zeta}{2}\right)}\, d\zeta \right\} \\ &\le 2\left\{ \int_0^{\frac{\pi}{m+1}} \frac{(m+1)\,\pi}{2}\, d\zeta + \int_{\frac{\pi}{m+1}}^{\pi} \frac{\pi}{\zeta}\, d\zeta \right\} \\ &= \pi^2 + 2\pi\, \ln(m+1).\end{aligned} \tag{2.28}$$

Hence from (2.25), (2.26), and (2.28), we get

$$\begin{aligned}\int_{-\pi}^{\pi} |\alpha_m(\theta+t,\theta)|\, dt &\le \frac{2(\pi n+1-r)}{(m+1)(1-r)^2}\left(\pi^2 + 2\pi\, \ln(m+1)\right) \\ &\le \frac{55\pi n}{(1-r)^2}\, \frac{\ln(m+1)}{m+1}, \qquad m \ge 1.\end{aligned} \tag{2.29}$$

To bound the second integral on the right-hand side of (2.18), change the variables again to $\zeta = \int_{\theta+t}^{\theta} \sum_{k=1}^{n} |B_k(e^{is})|^2\, ds$. Note that ζ is decreasing and (2.27) implies that it maps $[-\pi,\pi]$ onto $[\zeta(\pi), 2\pi n + \zeta(\pi)]$. Thus we may define an inverse map $\Xi : \zeta \mapsto t$. Let $t_k = \Xi(\zeta(\pi) + 2\pi k)$, $k = 0, 1, \ldots, n$, and suppose that for some j, $0 \in [t_{j+1}, t_j)$. Then

$$\begin{aligned}\int_{-\pi}^{\pi} |\beta(\theta+t,\theta)|\, \frac{[\Upsilon(\zeta(t))]^2}{m+1}\, dt &= \int_{\zeta(\pi)}^{2\pi n+\zeta(\pi)} |\beta(\theta+\Xi(\zeta),\theta)| \left(\frac{d\zeta}{d\eta}\right)^{-1} \frac{[\Upsilon(\zeta)]^2}{m+1}\, d\zeta \\ &\le \frac{2}{n(1-r)} \int_{\zeta(\pi)}^{2\pi n+\zeta(\pi)} \underbrace{\frac{\left|\sin\left(\frac{\zeta}{2}\right)\right|}{\left|\sin\left(-\frac{\Xi(\zeta)}{2}\right)\right|}\, \frac{[\Upsilon(\zeta)]^2}{m+1}}_{\Lambda(\zeta)}\, d\zeta,\end{aligned} \tag{2.30}$$

where the inequality follows from (2.22). Split the integral above as follows:

$$\int_{\zeta(\pi)}^{\zeta(\pi)+2\pi n} \Lambda(\zeta)\, d\zeta = \sum_{k=1}^{j-1} \int_{\zeta(t_{k-1})}^{\zeta(t_k)} \Lambda(\zeta)\, d\zeta + \sum_{k=j+3}^{n} \int_{\zeta(t_{k-1})}^{\zeta(t_k)} \Lambda(\zeta)\, d\zeta + \int_{\zeta(t_{j-1})}^{\zeta(t_{j+2})} \Lambda(\zeta)\, d\zeta.$$

Note from (2.22) the following inequalities:

$$\frac{n(1-r)}{2}\,(t_k - t_{k+1}) \le 2\pi = \zeta(t_{k+1}) - \zeta(t_k) \le \frac{2n}{1-r}\,(t_k - t_{k+1}),$$

which imply

$$\frac{\pi(1-r)}{n} \le (t_k - t_{k+1}) \le \frac{4\pi}{n(1-r)}.$$

Thus for $k \ge j+2$,

$$t_k = t_{j+1} + \sum_{l=j+2}^{k} (t_l - t_{l-1}) \le -\frac{\pi(1-r)}{n}(k-j-1), \tag{2.31}$$

and for $k \le j-1$,

$$t_k = t_j + \sum_{l=k+1}^{j} (t_{l-1} - t_l) \ge \frac{\pi(1-r)}{n}(j-k). \tag{2.32}$$

Hence from (2.24), the fact that $\Upsilon^2(\zeta)$ is a periodic function with period 2π, and the identity (see [20])

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{[\Upsilon(t)]^2}{m+1}\,dt = 1, \tag{2.33}$$

we get

$$\begin{aligned}
\int_{\zeta(t_{j-1})}^{\zeta(t_{j+2})} \Lambda(\zeta)\,d\zeta &\le \frac{\pi n}{1-r}\int_{\zeta(t_{j-1})}^{\zeta(t_{j+2})} \frac{[\Upsilon(\zeta)]^2}{m+1}\,d\zeta \\
&= \frac{\pi n}{1-r}\int_{\zeta(\pi)+2\pi(j-1)}^{\zeta(\pi)+2\pi(j+2)} \frac{[\Upsilon(\zeta)]^2}{m+1}\,d\zeta \\
&\le \frac{6\pi^2 n}{1-r}.
\end{aligned}$$

From (2.23), (2.33), (2.31), and (2.32),

$$\begin{aligned}
\sum_{k\ne j,j+1,j+2}\int_{\zeta(t_{k-1})}^{\zeta(t_k)} \Lambda(\zeta)\,d\zeta &\le 2\pi\sum_{k=1}^{j-1}\frac{\pi}{t_k} - 2\pi\sum_{k=j+3}^{n}\frac{\pi}{t_{k-1}} \\
&\le \frac{2\pi n}{1-r}\left(\sum_{k=1}^{j-1}\frac{1}{j-k} + \sum_{k=j+3}^{n}\frac{1}{k-j-2}\right) \\
&\le \frac{4\pi n}{1-r}\sum_{k=1}^{n}\frac{1}{k} \\
&\le \frac{4\pi n}{1-r}(1+\ln n).
\end{aligned}$$

Hence

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} |\beta(\theta+t,\theta)|\,\frac{|\Upsilon(\zeta(t))|^2}{m+1}\,dt \le \frac{4+6\pi+4\ln n}{(1-r)^2}. \tag{2.34}$$

This bound takes care of the cases $j<2$ and $j>n-3$ for all possible values of n. The combination of the inequalities (2.29), (2.34), and (2.18) yields $\|\sigma_m(\cdot,\theta)\|_1 \le C_1 + c_m$. Then taking the supremum with respect to θ completes the proof. □

Let X_k be the linear span of the general orthonormal basis functions B_j, $|j| \le k$, and define

$$\delta_k(f, C(\mathbf{T})) \triangleq \inf_{g \in X_k} \|g - f\|_\infty, \qquad f \in C(\mathbf{T}). \tag{2.35}$$

Thus $\delta_k(f, C(\mathbf{T}))$ is the best approximation error of f by functions in X_k. A best approximation always exists since X_k is finite-dimensional. Furthermore, $\delta_k(f, C(\mathbf{T})) \to 0$ $(k \to \infty)$ because the basis functions defined by (1.3)–(1.4) and (1.7)–(1.8) are complete in $C(\mathbf{T})$.

Let f be a given function in $C(\mathbf{T})$, and for $k = n(m+1)$ let g be a minimizing solution in (2.35). Let $h = f - g$ denote the approximation error. Observe that $\mathcal{V}_m g = g$ since $g \in X_{n(m+1)}$ and $\chi_m(j) = 1$ for all $|j| \le n(m+1)$. Due to the linearity of $\mathcal{V}_m$, notice also that $\mathcal{V}_m h = \mathcal{V}_m f - \mathcal{V}_m g$. Thus from (2.3), Lemma 2.3, and (2.35) we find

$$\begin{aligned} \|\mathcal{V}_m f - f\|_\infty &= \|\mathcal{V}_m h - h\|_\infty \\ &\le (2\|\mathcal{F}_{2m+1}\| + \|\mathcal{F}_m\|)\, \|h\|_\infty + \|h\|_\infty \\ &\le (3\, C_1 + 3\, c_m + 1)\ \delta_{n(m+1)}(f, C(\mathbf{T})). \end{aligned} \tag{2.36}$$

Hence $\|\mathcal{V}_m f - f\|_\infty \to 0$ $(m \to \infty)$ as claimed.

The inequality (2.36) shows that the approximation error of the estimate defined by (2.3) is $O(\delta_{n(m+1)}(f, C(\mathbf{T})))$, which compares well with the best possible error $\delta_{n(2m+1)}(f, C(\mathbf{T}))$. For example, if $f(z)$ is analytic on a region that contains $\mathbf{T}$, then for some $\gamma \in (0, 1)$, $\delta_k(f, C(\mathbf{T})) = O(\gamma^k)$ and thus $\|\mathcal{V}_m f - f\|_\infty = O(\gamma^{n(m+1)})$. This result is in sharp contrast with the Cesàro means defined in (2.2), where $O(\frac{1}{n(m+1)})$ is the best possible convergence rate for the same f using the trigonometric basis functions $e^{\pm ik\theta}$.

In [7, 28], real-valued impulse response versions of the basis functions defined by (1.3) and (1.7) have been formulated. (The details and examples can be found in [7, section 5].) It was established that the new basis functions denoted by $\widetilde{B}_k$ have the same closure and approximation properties as the original basis functions (1.3) and (1.7). In particular,

$$\mathcal{S}_k f = \sum_{j=-k}^{k} \langle f, \widetilde{B}_j \rangle \widetilde{B}_j \triangleq \widetilde{\mathcal{S}}_k f$$

whenever $\{z_0, \ldots, z_k\}$ contains complex conjugates as well. In this case, if f has a real-valued impulse response, then both $\mathcal{S}_k f$ and $\widetilde{\mathcal{S}}_k f$ will have real-valued impulse responses. Thus the de la Vallée Poussin estimate of f defined by (2.3) always has a real-valued impulse response, since the basis functions are generated by a complex-conjugate closed pole-parameter set $\{z_1, \ldots, z_n\}$ through (1.3), (1.7), (1.4), and (1.8).

3. Overdetermined system of linear equations. In this section, the approximation results derived in section 2 will be linked to the identification problem formulated in section 1.

Let $p(N)$ be a nonnegative integer-valued function. Let

$$\Theta_N \triangleq \begin{bmatrix} B_{-p}(e^{i\omega_1}) & \cdots & B_0(e^{i\omega_1}) & \cdots & B_p(e^{i\omega_1}) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ B_{-p}(e^{i\omega_N}) & \cdots & B_0(e^{i\omega_N}) & \cdots & B_p(e^{i\omega_N}) \end{bmatrix}, \quad E_N \triangleq \begin{bmatrix} e_1 \\ \vdots \\ e_N \end{bmatrix}. \tag{3.1}$$

For convenience of notation, the columns of Θ_N are indexed from $-p$ to p. Since the frequencies $\{\omega_k\}$ are distinct, Θ_N has full column rank, provided that $N \geq 2p+1$. Then we may compute the Moore–Penrose pseudo-inverse of Θ_N defined by

$$\Theta_N^\dagger \triangleq (\Theta_N^* \Theta_N)^{-1} \Theta_N^*,$$

where Θ_N^* is the complex-conjugated transpose of Θ_N. It is well known that $\Theta_N^\dagger E_N$ is the unique solution of the minimum distance problem:

$$\min_{x \in \mathbf{C}^{2p+1}} \|\Theta_N x - E_N\|_2. \tag{3.2}$$

For each N, we define a map $\mathcal{T}_N : \mathbf{C}^{2p+1} \to C(\mathbf{T})$ by

$$\mathcal{T}_N X(z) \triangleq \sum_{k=-p}^{p} w_p(k)\, [\Theta_N^\dagger X](k)\, B_k(z), \tag{3.3}$$

where w_p is a symmetric nonnegative window function.

Assuming that the frequencies in (1.5) satisfy the condition

$$0 \leq \omega_{k+1} - \frac{2\pi}{N}\, k < \frac{2\pi}{N}, \qquad k = 0, \ldots, N-1, \tag{3.4}$$

we extend the discrete data $\{e_1, \ldots, e_N\}$ into L_∞ as follows:

$$\mathcal{P}_N E_N(e^{i\omega}) \triangleq e_k \ \text{ if } \ \omega \in \left[\frac{2\pi}{N}\,(k-1), \frac{2\pi}{N}\, k\right), \ \ k = 1, 2, \ldots, N. \tag{3.5}$$

Next we define a sequence of operators $\mathcal{W}_N : L_\infty \to C(\mathbf{T})$ as follows:

$$\mathcal{W}_N f(z) \triangleq \sum_{k=-p}^{p} w_p(k)\, \langle f, B_k \rangle\, B_k(z). \tag{3.6}$$

Thus the composite operator $\mathcal{W}_N \mathcal{P}_N$ maps $E_N \in \mathbf{C}^{2p+1}$ into $C(\mathbf{T})$.

In the following result, we derive an upper bound on the difference between the operators $\mathcal{T}_N$ and $\mathcal{W}_N \mathcal{P}_N$, assuming that the frequencies in (1.5) are in one-to-one correspondence with the uniformly spaced frequencies: $\frac{2\pi k}{N}$, $k = 0, \ldots, N-1$.

LEMMA 3.1. *Consider the basis functions defined by* (1.3) *and* (1.7). *Suppose that they are uniformly bounded and let* r *be as in* (2.9). *Let* $\mathcal{T}_N$, $\mathcal{P}_N$, $\mathcal{W}_N$ *be as in* (3.3), (3.5), (3.6), *respectively. Assume that the frequencies in* (1.5) *satisfy* (3.4). *Suppose also that*

$$N \geq 26 p^2 \left(\frac{1+r}{1-r}\right)^2. \tag{3.7}$$

Then for all $p \geq 3$,

$$\|\mathcal{T}_N - \mathcal{W}_N \mathcal{P}_N\| \leq \|w_p\|_\infty\, K_p, \tag{3.8}$$

where

$$K_p \triangleq 76 \left(\frac{1+r}{1-r}\right)^{\frac{5}{2}} \frac{p^{\frac{5}{2}}}{N}. \tag{3.9}$$

Proof. Let

$$\widetilde{e}_N(z) \triangleq \sum_{k=-p}^{p} w_p(k)\,\frac{1}{N}\left[\Theta_N^* E_N\right](k)\,B_k(z). \tag{3.10}$$

Then the sums

$$\frac{1}{N}[\Theta_N^* E_N](k) = \frac{1}{N}\sum_{j=1}^{N} e_j\,\overline{B_k(e^{i\omega_j})}, \quad k = 0, \pm 1, \ldots,$$

are approximations to the Fourier coefficients of $\mathcal{P}_N E_N$,

$$\langle \mathcal{P}_N E_N, B_k\rangle = \frac{1}{2\pi}\int_0^{2\pi} \mathcal{P}_N E_N(e^{is})\,\overline{B_k(e^{is})}\,ds, \quad k = 0, \pm 1, \ldots,$$

since from (3.5) and the mean value theorem for integrals we have

$$\begin{aligned}\frac{1}{2\pi}\int_0^{2\pi} \mathcal{P}_N E_N(e^{is})\,\overline{B_k(e^{is})}\,ds &= \frac{1}{2\pi}\sum_{j=1}^{N} e_j \int_{\frac{2\pi}{N}(j-1)}^{\frac{2\pi}{N}j} \overline{B_k(e^{is})}\,ds \\ &= \frac{1}{N}\sum_{j=1}^{N} e_j\,\overline{B_k(e^{is_j})},\end{aligned}$$

where for each j, s_j lies in the interval $(\frac{2\pi}{N}(j-1), \frac{2\pi}{N}j)$. Then the approximation errors can be bounded from (3.4) as follows:

$$\begin{aligned}\left|\frac{1}{N}[\Theta_N^* E_N](k) - \langle \mathcal{P}_N E_N, B_k\rangle\right| &\le \Omega_{\overline{B_k}}\left(\frac{2\pi}{N}\right)\|E_N\|_\infty \\ &\le \frac{2\pi}{N}\max_s\left|\frac{d}{ds}B_k(e^{is})\right|\|E_N\|_\infty, \quad k = 0, \pm 1, \ldots,\end{aligned} \tag{3.11}$$

where Ω_f is the modulus of continuity of f defined by

$$\Omega_f(\varrho) \triangleq \sup_{|x-y|\le\varrho} |f(x) - f(y)|.$$

It is known from [13] (see also [2, Lem. 13]) that if $g(z)$ is a rational function with poles outside the disk $\{z : |z| < R\}$ $(R > 1)$ and with numerator and denumerator degrees p and q, then

$$\left\|\frac{dg}{dz}\right\|_\infty \le \max\{p, q\}\,\frac{R+1}{R-1}\,\|g\|_\infty.$$

Applying this result to the basis functions B_k, $k \ge 0$, with $R = 1/r$, where r is as defined in (2.9), we get from (2.19)

$$\left\|\frac{\mathrm{d}B_k}{\mathrm{d}z}\right\|_\infty \le \left(\frac{1+r}{1-r}\right)^{\frac{3}{2}} k. \tag{3.12}$$

By the substitution $z \mapsto z^{-1}$, note that the same inequality holds for $k < 0$ with $|k|$ replacing k on the right-hand side. Hence from (3.11) and (3.12) we have

$$\begin{aligned}\sum_{k=-p}^{p} \left| \frac{1}{N}[\Theta_N^* E_N](k) - \langle \mathcal{P}_N E_N, B_k \rangle \right| &\leq \frac{2\pi}{N} \left(\frac{1+r}{1-r} \right)^{\frac{3}{2}} \|E_N\|_\infty \sum_{k=-p}^{p} |k| \\ &\leq \frac{2\pi}{N}\, p(p+1) \left(\frac{1+r}{1-r} \right)^{\frac{3}{2}} \|E_N\|_\infty.\end{aligned}$$

Thus

$$\begin{aligned}\|\mathcal{W}_N \mathcal{P}_N E_N - \widetilde{e}_N\|_\infty &\leq \|w_p\|_\infty \max_{-p \leq k \leq p} \|B_k\|_\infty \sum_{k=-p}^{p} \left| \frac{1}{N}[\Theta_N^* E_N](k) - \langle \mathcal{P}_N E_N, B_k \rangle \right| \\ &\leq \underbrace{\frac{2\pi}{N}\, p(p+1) \left(\frac{1+r}{1-r} \right)^{2}}_{\triangleq \gamma} \|w_p\|_\infty \|E_N\|_\infty. \end{aligned} \tag{3.13}$$

Next, from

$$\mathcal{T}_N E_N - \widetilde{e}_N = \sum_{k=-p}^{p} w_p(k) \left(I - \frac{1}{N}\, \Theta_N^* \Theta_N \right) [\Theta_N^\dagger E_N](k)\, B_k, \tag{3.14}$$

where I denotes the $2p+1$ by $2p+1$ identity matrix, we have

$$\|\mathcal{T}_N E_N - \widetilde{e}_N\|_\infty \leq \|w_p\|_2 \left\| I - \frac{1}{N} \Theta_N^* \Theta_N \right\|_2 \|\Theta_N^\dagger E_N\|_2 \max_{-p \leq k \leq p} \|B_k\|_\infty, \tag{3.15}$$

where $\|A\|_2$ denotes the spectral norm of A. For the entries of the second term in (3.14), upper bounds are derived as follows:

$$\begin{aligned}\left| I_{k,j} - \frac{1}{N}\, [\Theta_N^* \Theta_N]_{k,j} \right| &= \left| \frac{1}{2\pi} \int_0^{2\pi} \overline{B_k(e^{is})}\, B_j(e^{is})\, ds - \frac{1}{N}\, [\Theta_N^* \Theta_N]_{k,j} \right| \\ &= \frac{1}{N} \sum_{l=1}^{N} \left| \frac{N}{2\pi} \int_{\frac{2\pi}{N}(l-1)}^{\frac{2\pi}{N} l} \overline{B_k(e^{is})}\, B_j(e^{is})\, ds - \overline{B_k(e^{i\omega_l})}\, B_j(e^{i\omega_l}) \right| \\ &= \frac{1}{N} \sum_{l=1}^{N} \left| \overline{B_k(e^{is_l})}\, B_j(e^{is_l})\, ds - \overline{B_k(e^{i\omega_l})}\, B_j(e^{i\omega_l}) \right| \\ &\leq \Omega_{\overline{B_k} B_j} \left(\frac{2\pi}{N} \right) \\ &\leq \frac{2\pi}{N} \left(\|B_j\|_\infty \left\| \frac{dB_k}{dz} \right\|_\infty + \|B_k\|_\infty \left\| \frac{dB_j}{dz} \right\|_\infty \right) \\ &\leq \frac{2\pi}{N} \left(\frac{1+r}{1-r} \right)^{2} (|k| + |j|)\,, \end{aligned}$$

where the last equality has followed from the mean-value theorem for integrals and, for each l, s_l, lies in the interval $(\frac{2\pi}{N}(l-1), \frac{2\pi}{N} l)$. Thus from $\|A\|_2^2 \leq \sum_{k,l} |A_{k,l}|^2$, $p \geq 3$, and the last inequality above, we get

$$\left\| I - \frac{1}{N} \Theta_N^* \Theta \right\|_2 \leq \frac{4\pi}{N} \left(\frac{1+r}{1-r} \right)^{2} p^2 \leq 2\gamma. \tag{3.16}$$

Hence using $\|A\|_2^2 = \|A^*A\|_2$ and $\sqrt{1+x} \le 1+(x/2),\ |x| \le 1$, we obtain the following inequalities:

$$\left\| \frac{\Theta_N}{N^{\frac{1}{2}}} \right\|_2 \le 1+\gamma, \qquad \left\| \frac{\Theta_N}{N^{\frac{1}{2}}} \right\|_{\min} \ge 1-\gamma, \tag{3.17}$$

where $\|A\|_{\min}$ denotes the smallest nonzero singular value of A. Therefore

$$\|\Theta_N^\dagger E_N\|_2 \le \frac{\|\Theta_N\|_2\ \|E_N\|_2}{N^{\frac{1}{2}}\ \|\Theta_N\|_{\min}^2} \le \frac{1+\gamma}{1-\gamma}\,\|E_N\|_\infty. \tag{3.18}$$

Thus from (3.15), (3.16), and (3.18) we get

$$\begin{aligned} \|\mathcal{T}_N E_N - \widetilde{e}_N\|_\infty &\le 2\gamma\,\|w_p\|_2\,\frac{1+\gamma}{1-\gamma}\,\left(\frac{1+r}{1-r}\right)^{\frac{1}{2}}\,\|E_N\|_\infty \\ &\le 4\gamma\,\frac{1+\gamma}{1-\gamma}\,\left(\frac{1+r}{1-r}\right)^{\frac{1}{2}}\,p^{\frac{1}{2}}\,\|w_p\|_\infty\,\|E_N\|_\infty, \end{aligned} \tag{3.19}$$

where the second inequality follows from $\|w_p\|_2 \le (2p+1)^{\frac{1}{2}}\|w_p\|_\infty$.

Assuming $p \ge 3$ and $\gamma \le \frac{1}{3}$, an application of the triangle inequality to (3.13) and (3.19) yields

$$\begin{aligned} \sup_{\|E_N\|_\infty \le \varepsilon} \|\mathcal{T}_N E_N - \mathcal{W}_N \mathcal{P}_N E_N\|_\infty &\le \left\{ p^{-\frac{1}{2}} + 4\,\frac{1+\gamma}{1-\gamma}\,\left(\frac{1+r}{1-r}\right)^{\frac{1}{2}} \right\} p^{\frac{1}{2}}\,\gamma\,\|w_p\|_\infty\,\varepsilon \\ &\le 9\left(\frac{1+r}{1-r}\right)^{\frac{1}{2}}\,p^{\frac{1}{2}}\,\gamma\,\|w_p\|_\infty\,\varepsilon. \end{aligned}$$

Hence, from the above inequality and $p \ge 3$, we have

$$\|\mathcal{T}_N - \mathcal{W}_N \mathcal{P}_N\| \le 76\left(\frac{1+r}{1-r}\right)^{\frac{5}{2}}\,\|w_p\|_\infty\,\frac{p^{\frac{5}{2}}}{N},$$

which we set out to prove. Finally, $\gamma \le \frac{1}{3}$ is implied by (3.7). □

Since the system to be identified has a real-valued impulse response, its transfer function $G(z)$ must satisfy the complex-conjugate symmetry:

$$G(e^{-i\omega_k}) = \overline{G(e^{i\omega_k})}, \qquad k = 1, \ldots, N. \tag{3.20}$$

Therefore it suffices to measure the frequency response at frequencies up to π. Then the frequency response on $(\pi, 2\pi)$ is obtained from (3.20). Thus without loss of generality we assume that E_N defined in (3.1) obeys (3.20), which forces E_N to be real-valued at the frequencies 0 and π. That can simply be satisfied by taking the real parts of E_N at those frequencies. Assuming that the frequencies are ordered as $\omega_k < \omega_{k+1}$, $k = 1, \ldots, N$, a subset satisfying (3.4) can be extracted from $\{\omega_k,\ k = 1, \ldots, N\}$ by the following process.

Regularization. Let κ_N be the *maximum frequency gap* defined by

$$\kappa_N \triangleq \max\left\{ \max_{1 \le k < N} (\omega_{k+1} - \omega_k),\, 2\pi + \omega_1 - \omega_N \right\}. \tag{3.21}$$

Let N' be the largest integer rounding $\frac{2\pi}{\kappa_N}$ down. For $j = 1, \ldots, N'$, pick one frequency ω_{k_j} from each interval $[\frac{2\pi(j-1)}{N'}, \frac{2\pi j}{N'})$ and let $e'_j \triangleq e_{k_j}$ and $\omega'_j \triangleq \omega_{k_j}$.

Thus $\{\omega'_j,\ j = 1, \ldots, N'\}$ satisfies (3.7), and Lemma 3.1 applies to the subset of the data $\{e'_j,\ j = 1, \ldots, N'\}$. Henceforth, without loss of generality, we set $N = N'$ and $k_j = j$ for all j. The regularization process does not require two adjacent points to be well separated. Recall the robustness definition that excludes prior information beyond $G(z) \in A(\mathbf{D})$. By removing closely spaced frequencies, we don't expect to lose much information.

Since $\|\mathcal{P}_N\| = 1$, from Lemma 3.1 we have

$$\|\mathcal{T}_N\| \le \|\mathcal{W}_N\| + \|w_p\|_\infty K_p. \tag{3.22}$$

Assuming $\sup_p \|w_p\|_\infty < \infty$ and $p = o(N^{\frac{2}{5}})$, (3.22) then implies $\sup_N \|\mathcal{T}_N\| < \infty$ if $\sup_N \|\mathcal{W}_N\| < \infty$. In addition to $\sup_N \|\mathcal{W}_N\| < \infty$, if the following holds

$$\lim_{p\to\infty} w_p(k) = 1 \qquad \text{for all } \ k, \tag{3.23}$$

then $\|\mathcal{W}_N f - f\|_\infty \to 0$ for all $f \in C(\mathbf{T})$ [11]. This implies by Lemma 3.1 that the sequence of operators $\mathcal{T}_N$ defined in (3.3) asymptotically recovers continuous functions from noise-free samples.

Given a set of arbitrary uniformly bounded basis functions, the problem of finding a window function that satisfies

$$\sup_N \|\mathcal{W}_N\| = \sup_p \max_s \int_{-\pi}^{\pi} \left| \sum_{k=-p}^{p} w_p(k)\, \overline{B_k(e^{is})}\, B_k(e^{i\theta})\, d\theta \right| < \infty \tag{3.24}$$

is nontrivial. In section 2, we demonstrated that (3.24) holds for the general orthonormal basis functions and the de la Vallée Poussin window function. We combine Lemma 2.3 and Lemma 3.1 to obtain the following important result.

LEMMA 3.2. *Consider the general orthonormal basis functions defined by* (1.3), (1.7), (1.4), *and* (1.8). *Let $m(N)$ be a nonnegative integer-valued function. Let $p = n(2m+1)$ and*

$$\widehat{\delta}_m(G) \triangleq \inf \left\{ \|f - G\|_\infty :\ f \in \mathrm{sp}\{B_j\}_{j=0}^{n(m+1)} \right\}. \tag{3.25}$$

Let χ_m be as in (2.4). *Define the* preidentified model*:*

$$\widetilde{G}_N(z) \triangleq \sum_{k=-p}^{p} \chi_m(k)\, [\Theta_N^\dagger E_N](k)\, B_k(z). \tag{3.26}$$

Let C_1, c_m, and K_p be as in (2.17) *and* (3.9). *Then for all $G(z) \in A(\mathbf{D})$,*

$$\sup_{\|\nu\|_\infty \le \varepsilon} \|\widetilde{G}_N - G\|_\infty \le \widehat{\delta}_m(G) + (3C_1 + 3c_m + K_p) \left(\widehat{\delta}_m(G) + \varepsilon \right). \tag{3.27}$$

Proof. Decompose G as $G = g + h$, where $g \in \mathrm{sp}\{B_j\}_{j=0}^{n(m+1)}$ and h is a minimizing solution in (3.25). Let $U_N \triangleq [g(e^{i\omega_1}) \cdots g(e^{i\omega_N})]^T$ and $Y_N \triangleq [h(e^{i\omega_1}) + \nu_1 \cdots h(e^{i\omega_N}) + \nu_N]^T$. Consider $\mathcal{T}_N$ in (3.3) with $w_p = \chi_m$. Thus $E_N = U_N + Y_N$ and $\mathcal{W}_N = \mathcal{V}_m$. Since $\mathcal{T}_N$ is linear,

$$\widetilde{G}_N = \mathcal{T}_N E_N = \mathcal{T}_N U_N + \mathcal{T}_N Y_N.$$

Note that $\mathcal{T}_N U_N = g(z)$ since $g \in \mathrm{sp}\{B_j\}_{j=0}^{n(m+1)}$ and $\chi_m(k) = 1, \;\; k = 0, \ldots, n(m+1)$. To bound $\mathcal{T}_N Y_N$, we first derive an upper bound for $\|\mathcal{T}_N\|$ as follows:

$$\begin{aligned}\|\mathcal{T}_N\| &\le \|\mathcal{W}_N\| + \|w_p\|_\infty K_p \\ &= \|\mathcal{V}_m\| + K_p \\ &\le (2\|\mathcal{F}_{2m+1}\| + \|\mathcal{F}_m\|) + K_p \\ &\le 3(C_1 + c_m) + K_p,\end{aligned}$$

where the second inequality follows from (2.3), and the third from (2.10) and Lemma 2.3. Hence

$$\begin{aligned}\|\mathcal{T}_N Y_N\|_\infty &\le \|\mathcal{T}_N\| \|Y_N\| \\ &\le (3C_1 + 3c_m + K_p)\,(\|h\|_\infty + \|\nu\|_\infty) \\ &\le (3C_1 + 3c_m + K_p)\left(\widehat{\delta}_m(G) + \varepsilon\right),\end{aligned}$$

where the last inequality follows from $\|h\|_\infty = \widehat{\delta}_m(G)$. Thus

$$\begin{aligned}\|G - \widetilde{G}_N\|_\infty &= \|h - \mathcal{T}_N Y_N\|_\infty \\ &\le \|h\|_\infty + \|\mathcal{T}_N Y_N\|_\infty \\ &\le \widehat{\delta}_m(G) + (3C_1 + 3c_m + K_p)\left(\widehat{\delta}_m(G) + \varepsilon\right).\end{aligned}$$

Taking the supremum of the left-hand side with respect to ν completes the proof. □

Now we are in a position to state the main result of this paper in the next section.

4. Two-stage nonlinear algorithm. In the previous section, we computed an approximant to the system from the noisy data. Although this estimate is close enough to the system so that it satisfies the criterion in (1.6), it can not be taken as the identified model since it contains unstable dynamics. The stable identified model is then obtained from $\widetilde{G}_N$ by solving the following nonlinear optimization problem known as the Nehari distance problem:

$$\widehat{G}_N \triangleq \arg \min_{f \in H_\infty} \|\widetilde{G}_N - f\|_\infty. \tag{4.1}$$

The Nehari problem was first used by Helmicki, Jacobson, and Nett [18] to solve the identification problem in H_∞ formulated in section 1. Notice from (4.1) the following inequality:

$$\|\widehat{G}_N - G\|_\infty \le 2\,\|\widetilde{G}_N - G\|_\infty \tag{4.2}$$

from which the main result of this paper follows.

THEOREM 4.1. *Let $\widehat{G}_N$ and κ_N be as in* (4.1) *and* (3.21). *Suppose $\kappa_N \to 0$. Then*

$$\lim_{\substack{\varepsilon \to 0 \\ N \to \infty}} \sup_{\|\nu\|_\infty \le \varepsilon} \|\widehat{G}_N - G\|_\infty = 0 \quad \text{for all} \quad G \in A(\mathbf{D}), \tag{4.3}$$

and the convergence is robust. Furthermore,

$$\sup_{\|\nu\|_\infty \le \varepsilon} \|\widehat{G}_N - G\|_\infty \le 2\widehat{\delta}_m(G) + 2\,(3C_1 + 3c_m + K_p)\left(\widehat{\delta}_m(G) + \varepsilon\right), \tag{4.4}$$

where $p = n(2m+1)$, *and* C_1, c_m, K_p, $\widehat{\delta}_m(G)$ *are defined by* (2.17), (3.9), (3.25).

Assuming $m = o(N^{\frac{2}{5}})$, then from (4.4) asymptotically in N

$$\sup_{\|\nu\|_\infty \le \varepsilon} \|\widehat{G}_N - G\|_\infty \le 2(1+3C_1)\,\widehat{\delta}_m(G) + 6C_1\varepsilon.$$

The right-hand side converges to zero at a geometric rate of m when the data are noise-free and the identified system is exponentially stable. This bound is similar to the asymptotic error bound computed in [6] for the trigonometric basis $\{e^{ik\theta}\}$. This is not unexpected for the general orthonormal basis functions, since $H_2(\mathbf{D})$ can be decomposed into at most n orthogonal subspaces by means of an inner function constructed from the chosen n basis poles. Then, intuitively, a system lying in $A(\mathbf{D})$ can be approximated on the decomposing subspaces, although how this could be achieved is not clear.

A drawback of the algorithm is the amount of required data. The requirement $m = o(N^{\frac{2}{5}})$ is severe. The minimax algorithm studied in [2] principally due to Mäkilä and Partington, on the other hand, has a modest requirement $m = O(N)$. The minimax algorithm is, however, computationally expensive, in particular for large values of m.

5. Example. In this section, we use a simulation example to illustrate the use of the generalized basis functions defined by (1.3)–(1.4) in an iterative-identification scheme. In the example, the Nehari step is omitted and each iteration step involves the estimation of a large-order model followed by a model reduction step leading to a new set of basis poles.

We consider the identification of a fifth-order system with poles (in the usual stability notion) $0.95\pm0.20i$, $0.85\pm0.10i$, 0.55, and zeros $0.96\pm0.28i$, $0.96\pm0.17i$. The transfer function of the system is normalized so that its H_∞ norm satisfies $\|G\|_\infty = 1$. This system was used in [39, 7] to illustrate the use of the basis functions in a one-step identification algorithm.

As in [7], we assume $N = 500$ frequency response measurements

$$e_k = G(e^{i\omega_k}) + \nu_k, \qquad k = 1, \dots, N, \tag{5.1}$$

are available, where ω_k are equally spaced on the interval $[0, 3]$ and the disturbances ν_k are bounded random variables

$$\nu_k = 0.1\, e^{i\alpha_k},$$

where α_k are independent and uniformly distributed random variables in the interval $[0, 2\pi]$. Note that, by this choice of frequencies, the frequency response is not on a uniform grid of frequencies. For a comparison with approximation results, we also consider the possibility $\nu = 0$ in (5.1).

We will estimate G from the data (5.1) by the following iterative algorithm. First, using the basis functions in (1.3) with

$$z_k = \begin{cases} 0.2, & k \text{ odd}, \\ 0.9, & k \text{ even}, \end{cases}$$

a high-order model is computed from (5.1) by the simple least-squares method as

$$H^{(0)}(z) = \sum_{k=0}^{2m} [\Phi^\dagger E_N]_k B_k(z), \tag{5.2}$$

where E_N is as defined in (3.1), and

$$\Phi = \begin{bmatrix} 1 & \cdots & B_{2m}(e^{i\omega_1}) \\ \vdots & \ddots & \vdots \\ 1 & \cdots & B_{2m}(e^{i\omega_N}) \end{bmatrix}. \tag{5.3}$$

This simple choice of basis functions represents both slow and fast dynamics in the model structure by the Laguerre functions. In the simulation, we fixed the number of basis functions as 21.

We reduced $H^{(0)}$ to a fifth order model denoted by $\widehat{G}_N^{(0)}$, by using the subspace-based identification algorithm in [26] for model reduction purposes. The input to the algorithm in [26] was 4096 equally spaced frequency response data on $[0, 2\pi]$. Note that this amounts to evaluating Φ on a uniform grid of 4096 frequencies for which fast algorithms are known to exist. The size of the Hankel matrix in the subspace algorithm was chosen to be 256 by 256. These values will be used throughout the iterations. The step prior to forming a Hankel matrix was a 4096-point inverse fast Fourier transform.

The reduced model starts the iterations. Let $z_1^{(1)}, \ldots, z_5^{(1)}$ denote the inverses of the poles of $\widehat{G}_N^{(0)}(z)$. The updated basis functions $B_0^{(1)}, \ldots, B_{20}^{(1)}$ are computed from (1.3) with $z_0^{(1)} = 0$ and

$$z_{j+5m}^{(1)} = z_j^{(1)}, \qquad j = 1, \ldots, 5, \quad m = 1, 2, 3.$$

Then $H^{(0)}$ in (5.2) is updated to

$$H^{(1)}(z) = \sum_{k=0}^{20} [\Phi^\dagger E_N]_k B_k^{(1)}(z),$$

where Φ in (5.3) is computed with the updated basis functions. Next $H^{(1)}$ is reduced to a fifth order model $\widehat{G}_N^{(1)}$, using the same model reduction technique. This completes the first iteration. The next iteration starts with the inverses of the poles of $\widehat{G}_N^{(1)}(z)$ in place of $z_1^{(2)}, \ldots, z_5^{(2)}$, and so on.

The quality of the estimated models will be assessed by a measure based on the fit between the data and the model. For this purpose, the maximum error defined as

$$\|\widehat{G}_N^{(k)} - E_N\|_{m,\infty} \stackrel{\Delta}{=} \max_{1 \le j \le N} \left| \widehat{G}_N^{(k)}(e^{i\omega_j}) - e_j \right|$$

will be used. The iterations are then terminated when the sequence $\{\|\widehat{G}_N^{(k)} - E_N\|_{m,\infty}\}$ seems to have reached a steady state.

5.1. Results. The results of the simulation are given in Tables 5.1 and 5.2. The iterations converged at $k = 2$ for the noise-free data and at $k = 1$ for the noisy data. For the noisy data, the poles of $\widehat{G}_N^{(1)}$ are $0.94 \pm 0.20i$, $0.86 \pm 0.10i$, 0.53, and the four significant zeros are $0.96 \pm 0.16i$, $0.96 \pm 0.28i$. They all agree well with the system poles and zeros. Thus the major source of the identification errors in Table 5.2 appears to be the measurement error ν. This observation is justified by computing the poles of $\widehat{G}_N^{(0)}$ as $0.91 \pm 0.12i$, $0.90 \pm 0.17i$, 0.52. In Figure 5.1, the magnitudes of $G(e^{i\omega_k})$ and $\widehat{G}_N^{(1)}(e^{i\omega_k})$, and the identification error $\widehat{G}_N^{(1)}(e^{i\omega_k}) - G(e^{i\omega_k})$, are plotted for the noisy data. In the reduction of $H^{(1)}(z)$ for the noisy data, the first seven singular values of the Hankel matrix were computed as 0.5471, 0.2407, 0.1926, 0.0973, 0.0963, 0.0369, 0.0367. The choice of (final) model order was based on this distribution.

TABLE 5.1
Results for the noise-free data, where k is the number of iterations.

k	$\|H^{(k)} - E_N\|_{m,\infty}$	$\|\widehat{G}_N^{(k)} - E_N\|_{m,\infty}$
0	1.07×10^{-1}	1.00×10^{-1}
1	3.48×10^{-3}	5.21×10^{-4}
2	3.58×10^{-12}	5.54×10^{-6}
3	8.31×10^{-13}	5.54×10^{-6}
4	8.24×10^{-13}	5.54×10^{-6}

TABLE 5.2
Results for the noisy data, where k is the number of iterations.

k	$\|H^{(k)} - E_N\|_{m,\infty}$	$\|\widehat{G}_N^{(k)} - E_N\|_{m,\infty}$
0	0.1890	0.1859
1	0.1383	0.1237
2	0.1387	0.1228
3	0.1390	0.1229
4	0.1390	0.1229

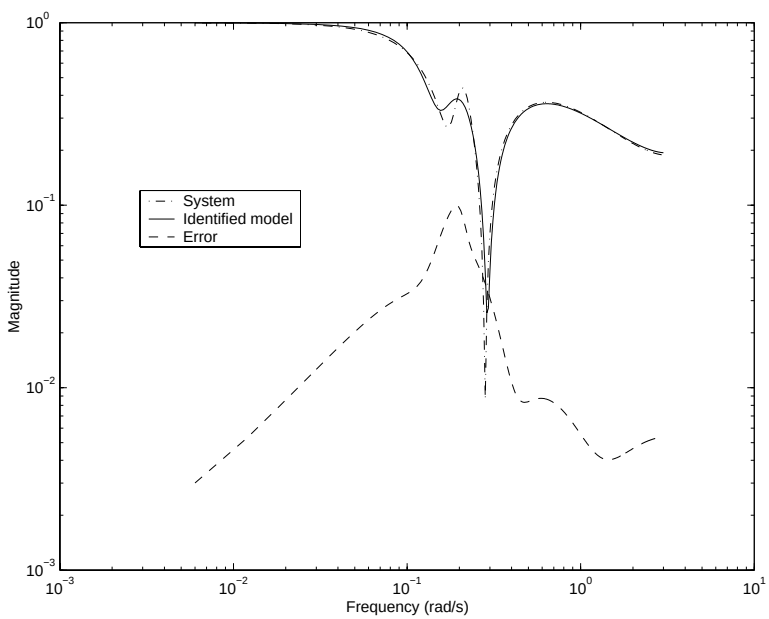


FIG. 5.1. *The magnitude plots of* $G(e^{i\omega_k})$, $\widehat{G}_N^{(1)}(e^{i\omega_k})$, *computed from the noisy data, and the identification error* $\widehat{G}_N^{(1)}(e^{i\omega_k}) - G(e^{i\omega_k})$.

5.2. Discussion. This example has demonstrated that the iterative-identification scheme works. The convergence analysis of this scheme and its relation to other iterative-identification schemes [35, 23] need to be studied. The convergence properties of this scheme with respect to unknown-but-bounded noise are studied in [5].

6. Conclusions. In this paper, we presented a robust two-stage algorithm that uses general orthonormal basis functions to identify linear-time invariant systems from nonuniformly spaced frequency response measurements. We also derived worst-case identification error bounds in the H_∞ norm, assuming that the number of basis

functions does not increase faster than a certain rate relative to the amount of data. The error convergence rate of the algorithm as a function of the number of basis functions, for every choice of basis functions, was shown to be on the order of the best possible.

The work initiated in this paper could be continued in several directions. First, the possibility of extending these results to arbitrary complete bases in $A(\mathbf{D})$ is worth investigating. The second, and practically more relevant, problem is the development of iterative pole updating schemes to improve the quality of transfer function estimates in applications where model complexity is restricted.

REFERENCES

[1] H. Akçay, G. Gu, and P. P. Khargonekar, *A class of algorithms for identification in $\mathcal{H}_\infty$: Continuous-time case*, IEEE Trans. Automat. Control, 38 (1993), pp. 289–294.

[2] H. Akçay and B. Ninness, *Rational basis functions for robust identification from frequency and time-domain measurements*, Automatica J. IFAC, 34 (1998), pp. 1101–1117.

[3] H. Akçay and B. Ninness, *Orthonormal basis functions for modelling continuous-time systems*, Signal Process., 77 (1999), pp. 261–274.

[4] H. Akçay and B. Ninness, *Orthonormal basis functions for continuous-time systems and L_p convergence*, Math. Control Signals Systems, 12 (1999), pp. 295–305.

[5] H. Akçay and P. Heuberger, *A frequency-domain iterative identification algorithm using general orthonormal basis functions*, Automatica J. IFAC, 37 (2001), pp. 663–674.

[6] H. Akçay, *Algorithms for robust identification in $\mathcal{H}_\infty$ with nonuniformly spaced frequency response data*, Math. Control Signals Systems, 11 (1998), pp. 161–181.

[7] H. Akçay, *Discrete-time system modelling in L_p with orthonormal basis functions*, Systems Control Lett., 39 (2000), pp. 365–376.

[8] H. Akçay, *On the existence of a disk algebra basis*, Signal Process., 80 (2000), pp. 903–907.

[9] H. Akçay, *Continuous-time stable and unstable system modelling with orthonormal basis functions*, Internat. J. Robust Nonlinear Control, 10 (2000), pp. 513–531.

[10] H. Akçay, *On the uniform approximation of discrete–time systems by generalized Fourier series*, IEEE Trans. Signal Process., 49 (2001), pp. 1461–1467.

[11] H. Akçay, *A stochastic analysis of robust estimation algorithms in H_∞ with rational basis functions*, Internat. J. Robust Nonlinear Control, to appear.

[12] J. Bokor and F. Schipp, *Approximate identification in Laguerre and Kautz bases*, Automatica J. IFAC, 34 (1998), pp. 463–468.

[13] P. Borwein and T. Erdelyi, *Sharp extensions of Berstein's inequality to rational spaces*, Mathematika, 43 (1996), pp. 413–423.

[14] A. C. Brinker, *Laguerre-domain adaptive filters*, IEEE Trans. Signal Process., 42 (1994), pp. 953–956.

[15] J. Chen, C. N. Nett, and M. K. H. Fan, *Worst-case system identification in $\mathcal{H}_\infty$: Validation of* a priori *information, essentially optimal algorithms, and error bounds*, IEEE Trans. Automat. Control, 40 (1995), pp. 1260–1265.

[16] G. Gu and P. P. Khargonekar, *A class of algorithms for identification in $\mathcal{H}_\infty$*, Automatica J. IFAC, 28 (1992), pp. 229–312.

[17] J. Head, *Approximation to transient by means of Laguerre series*, Proc. Cambridge Philos. Soc., 52 (1956), pp. 640–651.

[18] A. J. Helmicki, C. A. Jacobson, and C. N. Nett, *Control-oriented system identification: A worst-case/deterministic approach in $\mathcal{H}_\infty$*, IEEE Trans. Automat. Control, 36 (1991), pp. 1163–1176.

[19] P. S. C. Heuberger, P. M. J. Van den Hof, and O. Bosgra, *A generalized orthonormal basis for linear dynamical systems*, IEEE Trans. Automat. Control, 40 (1995), pp. 451–465.

[20] Y. Katznelson, *An Introduction to Harmonic Analysis*, Wiley, New York, 1968.

[21] W. H. Kautz, *Transient synthesis in the time domain*, IRE Trans. Circuit Theory, 1 (1954), pp. 29–39.

[22] Y. Lee, *Synthesis of electric networks by means of the Fourier transforms of Laguerre's functions*, J. Math. Phys., 11 (1933), pp. 83–113.

[23] L. Ljung, *System Identification: Theory for the User*, Prentice-Hall, Englewood Cliffs, NJ, 1987.

[24] P. M. Mäkilä, *Laguerre series approximation of infinite dimensional systems*, Automatica J. IFAC, 26 (1990), pp. 985–995.
[25] P. M. Mäkilä and J. R. Partington, *Robust identification of strongly stabilizable systems*, IEEE Trans. Automat. Control, 37 (1992), pp. 1709–1716.
[26] T. McKelvey, H. Akçay, and L. Ljung, *Subspace-based multivariable system identification from frequency response data*, IEEE Trans. Automat. Control, 41 (1996), pp. 960–979.
[27] M. Milanese and A. Vicino, *Information based complexity and nonparametric worst-case identification*, J. Complexity, 9 (1993), pp. 427–446.
[28] B. Ninness and F. Gustafsson, *A unifying construction of orthonormal bases for system identification*, IEEE Trans. Automat. Control, 42 (1997), pp. 515–521.
[29] B. Ninness, H. Hjalmarsson, and F. Gustafsson, *Generalized Fourier and Toeplitz results for rational orthonormal bases*, SIAM J. Control Optim., 37 (1998), pp. 429–460.
[30] Ü. Nurges, *Laguerre models in problems of approximation and identification*, in Adaptive Systems, Plenum Press, New York, 1987, pp. 346–352.
[31] A. M. Olevskii, *Fourier Series with Respect to General Orthogonal Systems*, translated by B. P. Marshall and H. J. Christoffers, Springer-Verlag, Berlin, 1975.
[32] J. R. Partington, *Approximation of delay systems by Fourier–Laguerre series*, Automatica J. IFAC, 27 (1991), pp. 569–572.
[33] J. R. Partington, *Robust identification and interpolation in $\mathcal{H}_\infty$*, Internat. J. Control, 54 (1991), pp. 1281–1290.
[34] J. R. Partington, *Robust identification in $\mathcal{H}_\infty$*, J. Math. Anal. Appl., 166 (1992), pp. 428–441.
[35] C. K. Sanathanan and J. Koerner, *Transfer function synthesis as a ratio of two complex polynomials*, IEEE Trans. Automat. Control, 8 (1963), pp. 56–58.
[36] F. Schipp, L. Gianone, and J. Bokor, *Identification in generalized orthogonal basis—a frequency domain approach*, in Proceedings of the 13th IFAC World Congress, San Francisco, CA, 1996, pp. 387–392.
[37] Z. Szabó, J. Bokor, and F. Schipp, *Identification of rational approximate models in H^∞ using generalized orthonormal basis*, IEEE Trans. Automat. Control, 44 (1999), pp. 153–158.
[38] P. M. J. Van den Hof, P. S. C. Heuberger, and J. Bokor, *System identification with generalized orthonormal basis functions*, Automatica J. IFAC, 31 (1995), pp. 1821–1834.
[39] D. K. de Vries and P. M. J. Van den Hof, *Frequency domain identification with generalized orthonormal basis functions*, IEEE Trans. Automat. Control, 43 (1998), pp. 656–669.
[40] B. Wahlberg, *System identification using Laguerre models*, IEEE Trans. Automat. Control, 36 (1991), pp. 551–562.
[41] B. Wahlberg, *System identification using Kautz models*, IEEE Trans. Automat. Control, 39 (1994), pp. 1276–1282.
[42] B. Wahlberg and P. M. Mäkilä, *On approximation of stable linear dynamical systems using Laguerre and Kautz functions*, Automatica J. IFAC, 32 (1996), pp. 693–708.
[43] N. F. D. Ward and J. R. Partington, *Rational wavelet decomposition of transfer functions in Hardy-Sobolev classes*, Math. Control Signals Systems, 8 (1995), pp. 257–278.
[44] N. F. D. Ward and J. R. Partington, *Robust identification in the disk algebra using rational wavelets and orthonormal bases functions*, Internat. J. Control, 64 (1996), pp. 409–423.
[45] P. Wojtaszczyk and K. Woźniakowski, *Orthonormal polynomial bases in function spaces*, Israel J. Math., 75 (1991), pp. 167–191.
[46] G. Zames, *On the metric complexity of causal linear systems: ϵ-entropy and ϵ-dimension for continuous time*, IEEE Trans. Automat. Control, 24 (1979), pp. 222–230.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 3, pp. 969–991

PICK MATRIX CONDITIONS FOR SIGN-DEFINITE SOLUTIONS OF THE ALGEBRAIC RICCATI EQUATION*

H. L. TRENTELMAN† AND P. RAPISARDA‡

Abstract. We study the existence of positive and negative semidefinite solutions of algebraic Riccati equations (ARE) corresponding to linear quadratic problems with an indefinite cost functional. The formulation of reasonable necessary and sufficient conditions for the existence of such solutions is a long-standing open problem. A central role is played by certain two-variable polynomial matrices associated with the ARE. Our main result characterizes all unmixed solutions of the ARE in terms of the Pick matrices associated with these two-variable polynomial matrices. As a corollary of this result, we find that the signatures of the extremal solutions of the ARE are determined by the signatures of particular Pick matrices.

Key words. algebraic Riccati equation, existence of semidefinite solutions, two-variable polynomial matrices, Pick matrices, dissipative systems

AMS subject classifications. 93C05, 93C15, 49N05, 49N10

PII. S036301290036851X

1. Introduction and problem statement.

Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ be such that (A, B) is a controllable pair. Let $Q \in \mathbb{R}^{n\times n}$ be symmetric and let $R \in \mathbb{R}^{m\times m}$ be nonsingular and symmetric. Finally, let $S \in \mathbb{R}^{m\times n}$. The quadratic equation

$$A^T K + KA + Q - (KB + S^T)R^{-1}(B^T K + S) = 0 \tag{1}$$

in the unknown $n \times n$ matrix K is called the (continuous-time) algebraic Riccati equation (the ARE). Since its introduction in control theory at the beginning of the sixties, the ARE has been studied extensively because of its prominent role in linear quadratic optimal control and filtering, H_∞-optimal control, differential games, and stochastic filtering and control. We refer to the papers collected in [2] for a discussion of the ARE and its applications and for an overview of the existing literature.

In this paper, we restrict ourselves to the case in which R is positive definite. However, the weighting matrix

$$M := \begin{pmatrix} Q & S^T \\ S & R \end{pmatrix}$$

is allowed to be indefinite. Our aim is to address a long-standing open problem concerning the ARE, namely, the problem of formulating reasonable necessary and sufficient conditions for the existence of at least one real positive semidefinite solution or of at least one real negative semidefinite solution. We want to stress that the main difficulty is the *indefiniteness of* M. For the case in which M is positive semidefinite, the problem is already well understood. For this case, necessary and sufficient conditions for the existence of at least one real positive semidefinite solution were obtained

*Received by the editors February 28, 2000; accepted for publication (in revised form) April 10, 2001; published electronically November 15, 2001.
http://www.siam.org/journals/sicon/40-3/36851.html

†Research Institute for Mathematics and Computer Science, P.O. Box 800, 9700 AV Groningen, The Netherlands (H.L.Trentelman@math.rug.nl).

‡Departmentof Mathematics, University of Maastricht, P.O. Box 616, 6200 MD Maastricht, The Netherlands (P.Rapisarda@math.unimaas.nl).

in [5] and [6]. Basically, these necessary and sufficient conditions can be formulated as follows: factor $M = (C\ \ D)^T(C\ \ D)$. Then the ARE (1) has at least one real positive semidefinite solution if and only if the system (A, B, C, D) is output stabilizable (see also [17], [23], or [24]).

For *indefinite* weighting matrices M, the problem was listed in [13] among a series of open problems in the field of systems and control. Partial results for this problem were obtained in [19, 20, 1, 7, 8]. For an overview and a discussion of these results, as well as their relation to the classical problem of the existence of nonnegative storage functions for dissipative systems, we refer to [13].

In the present paper we will present a solution to this open problem, under the assumption that the pair (A, B) is controllable. It will be proven that the signs of the smallest and largest real symmetric solution, respectively, depend on the signs of certain constant $n \times n$ matrices (so called *Pick matrices*, associated with the ARE), which are easily constructed from the parameters appearing in the ARE. A necessary and sufficient condition for the existence of a real symmetric positive semidefinite solution of the ARE (1) will turn out to be that (i) it has at least one real symmetric solution, and (ii) a suitable Pick matrix is negative semidefinite. Likewise, the existence of at least one negative semidefinite solution is determined by the positive semidefiniteness of a suitable Pick matrix. In the process of establishing these conditions we obtain a number of intermediate results, among which are a new characterization of all unmixed real symmetric solutions of the ARE, and a new characterization of the supremal and infimal real symmetric solutions, all in terms of the Pick matrices associated with the ARE.

A few words on notation are required at this point. In this paper we adopt the usual symbols $\mathbb{R}$ and $\mathbb{C}$ in order to denote the real and complex numbers, respectively. The open and closed right half-planes of $\mathbb{C}$ are denoted, respectively, by $\mathbb{C}_+^0$ and $\mathbb{C}_+$. Given $\lambda \in \mathbb{C}$, its complex conjugate is denoted by $\bar{\lambda}$. The space of n-dimensional real, respectively complex, vectors is denoted by $\mathbb{R}^n$, respectively $\mathbb{C}^n$, and the space of $m \times n$ real, respectively complex, matrices, by $\mathbb{R}^{m\times n}$, respectively $\mathbb{C}^{m\times n}$.

The symbol $\mathbb{R}^{\bullet\times n}$ denotes the space of real matrices with n columns, and $\mathbb{R}^{m\times\bullet}$ denotes the space of real matrices with m rows. Given two column vectors x and y, we denote with $\mathrm{col}(x, y)$ the vector obtained by stacking x over y. If $A \in \mathbb{R}^{m\times n}$, then $A^T \in \mathbb{R}^{n\times m}$ denotes its transpose, and if $A \in \mathbb{C}^{m\times n}$, then $A^* \in \mathbb{C}^{n\times m}$ denotes its conjugate transpose $\bar{A}^T$. If $A \in \mathbb{C}^{n\times n}$ is Hermitian, i.e., $A^* = A$, then we define the signature of A as the triple $\mathrm{sign}(A) = (n_-, n_0, n_+)$, where n_- is the number of negative eigenvalues of A, n_0 the algebraic multiplicity of 0 as an eigenvalue of A, and n_+ the number of positive eigenvalues of A.

The ring of polynomials with real coefficients in the indeterminate ξ is denoted by $\mathbb{R}[\xi]$; analogously, the ring of two-variable polynomials with real coefficients in the indeterminates ζ and η is denoted by $\mathbb{R}[\zeta, \eta]$. The space of all $n \times m$ polynomial matrices in the indeterminate ξ is denoted by $\mathbb{R}^{n\times m}[\xi]$, and that consisting of all $n \times m$ polynomial matrices in the indeterminates ζ and η by $\mathbb{R}^{n\times m}[\zeta, \eta]$. The space of polynomial matrices with real coefficients in the indeterminate ξ with n columns is denoted by $\mathbb{R}^{\bullet\times n}[\xi]$, and $\mathbb{R}^{m\times\bullet}[\xi]$ is the space of polynomial matrices with m rows. Given a matrix $R \in \mathbb{R}^{n\times m}[\xi]$, we define $R^\sim(\xi) := R^T(-\xi) \in \mathbb{R}^{m\times n}[\xi]$. If $F \in \mathbb{R}^{m\times n}[\xi]$, then F can be written as $F(\xi) = F_0 + F_1\xi + \cdots + F_L\xi^L$, where $F_j \in \mathbb{R}^{m\times n}$ for $j = 0, 1, \ldots, L$. We call the $m \times (L+1)n$ matrix $\tilde{F} := \begin{pmatrix} F_0 & F_1 & \ldots & F_L \end{pmatrix}$ the *coefficient matrix* of F. It is easy to see that

$$F(\xi) = \tilde{F} \begin{pmatrix} I_n \\ I_n\xi \\ \vdots \\ I_n\xi^L \end{pmatrix}.$$

For a given finite-dimensional Euclidean space X, we denote by $\mathfrak{C}^\infty(\mathbb{R}, X)$ the set of all infinitely differentiable functions from $\mathbb{R}$ to X, and by $\mathfrak{D}(\mathbb{R}, X)$ the subset of $\mathfrak{C}^\infty(\mathbb{R}, X)$ consisting of those functions having compact support. Finally, if K is a symmetric $n \times n$ matrix, the quadratic form on $\mathbb{R}^n$ defined by $x \mapsto x^T K x$ is denoted by $| x |_K^2$.

2. Linear differential systems and quadratic differential forms. In this section we give a brief review of the notion of linear differential systems. The reader is referred to the textbook [9] or to [21] for a thorough exposition. A linear differential system is a linear subspace $\mathcal{B}$ of $\mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^q)$ of all solutions w of a given system of linear, constant coefficient, higher order differential equations. Such a system of differential equations can always be represented as a single equation

$$R\left(\frac{d}{dt}\right) w = 0, \tag{2}$$

where $R \in \mathbb{R}^{\bullet \times q}[\xi]$ is a real polynomial matrix with q columns. The linear space $\mathcal{B}$ is called the behavior of the linear differential system, and (2) is called a kernel representation of $\mathcal{B}$. The variable w is called the manifest variable of $\mathcal{B}$. An alternative way to represent the behavior of a linear differential system is as an image representation. If $M \in \mathbb{R}^{q \times d}[\xi]$ and $\mathcal{B} = \{w \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^q) \mid \exists\ l \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^d)$ such that $w = M(\frac{d}{dt})l\}$, then we call

$$w = M\left(\frac{d}{dt}\right) l \tag{3}$$

an image representation of $\mathcal{B}$. Not all behaviors admit an image representation; indeed, a behavior can be represented in image form if and only if every one of its kernel representations is associated with a polynomial matrix $R \in \mathbb{R}^{\bullet \times q}[\xi]$ such that $\operatorname{rank}(R(\lambda))$ is constant for all $\lambda \in \mathbb{C}$, or equivalently, such that $\mathcal{B}$ is controllable. The image representation (3) of $\mathcal{B}$ is called *observable* if $(M(\frac{d}{dt})l = 0) \Longrightarrow (l = 0)$. It can be shown that this is the case if and only if the matrix $M(\lambda)$ has full column rank for all $\lambda \in \mathbb{C}$.

We proceed to review the notion of state maps introduced in [12]. We will consider only the case of image representations in this paper. Let (3) be an image representation of the behavior $\mathcal{B}$. A polynomial matrix $X \in \mathbb{R}^{n \times d}[\xi]$ is said to induce a *state map* for $\mathcal{B}$ (or, simply, for M) if the latent variable $x := X(\frac{d}{dt})l$ satisfies the *axiom of state*. This means that if we define the *full behavior* as

$$\mathcal{B}_{\text{full}} = \left\{ (w, x) \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^q \times \mathbb{R}^n) \mid \text{there exists } l \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^d) \text{ such that } w = M\left(\frac{d}{dt}\right) l,\ x = X\left(\frac{d}{dt}\right) l \right\},$$

then $(w_1, x_1), (w_2, x_2) \in \mathcal{B}_{\text{full}}$ and $x_1(0) = x_2(0)$ imply that $(w_1, x_1) \wedge (w_2, x_2)$, the concatenation of (w_1, x_1) and (w_2, x_2) at $t = 0$, belongs to the closure in the topology

of $\mathfrak{L}^1_{\rm loc}$ of $\mathcal{B}_{\rm full}$ (see [12]). Now assume that the image representation (3) is observable. Then a state map for the system can be computed as follows. If necessary, permute the components of w so that

$$M = \begin{pmatrix} U \\ Y \end{pmatrix} \tag{4}$$

with $U \in \mathbb{R}^{d\times d}[\xi]$, $\det(U) \neq 0$, and YU^{-1} is a proper rational matrix (it can be shown that such a permutation always exists). Now consider the set

$$\{r \in \mathbb{R}^{1\times d}[\xi] \mid \ rU^{-1} \text{ is strictly proper}\}. \tag{5}$$

It is not difficult to show that this set is a vector space over $\mathbb{R}$. It has been proved in [12] that X is a state map for (3) if and only if its rows (interpreted as elements of the vector space $\mathbb{R}^{1\times d}[\xi]$ over $\mathbb{R}$) span the vector space (5), and is a *minimal state map* (i.e., inducing a state variable of minimal possible dimension) if and only if its rows form a basis for (5). If this holds true, the number of rows of X is called the *McMillan degree* of M, denoted $n(M)$, or, referring to the behavior being represented in image form, *the McMillan degree of* $\mathcal{B}$, denoted $n(\mathcal{B})$. It can be shown (see Proposition 3.5.5 of [12]) that $n(M) = \deg(\det(U))$.

In many modeling and control problems it is necessary to study certain functionals of the system variables and their derivatives. In the context of linear systems these functionals are often taken to be quadratic. An efficient representation for such quadratic functionals by means of two-variable polynomial matrices has been proposed in [18]. In this section we review the definitions and results of such a two-variable polynomial framework, which are used in the rest of the paper.

Let $\Phi \in \mathbb{R}^{q_1\times q_2}[\zeta,\eta]$; then Φ can be written in the form

$$\Phi(\zeta,\eta) = \sum_{h,k=0}^{N} \Phi_{h,k}\zeta^h\eta^k,$$

where $\Phi_{h,k} \in \mathbb{R}^{q_1\times q_2}$ and N is an integer. The two-variable polynomial matrix Φ induces a bilinear functional acting on infinitely differentiable trajectories as follows:

$$L_\Phi : \mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^{q_1}) \times \mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^{q_2}) \longrightarrow \mathfrak{C}^\infty(\mathbb{R},\mathbb{R}),$$

$$L_\Phi(w_1,w_2) = \sum_{h,k=0}^{N} \left(\frac{d^h w_1}{dt^h}\right)^T \Phi_{h,k} \frac{d^k w_2}{dt^k}.$$

If Φ is a *symmetric two-variable polynomial matrix*, i.e., if $q_1 = q_2$ and $\Phi_{h,k} = \Phi^T_{k,h}$ for all h, k, then it induces also a quadratic functional $Q_\Phi : \mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^q) \longrightarrow \mathfrak{C}^\infty(\mathbb{R},\mathbb{R})$ defined by $Q_\Phi(w) := L_\Phi(w,w)$. We will call Q_Φ the *quadratic differential form* (QDF) associated with Φ. We denote the set of all symmetric $q\times q$ two-variable polynomial matrices by $\mathbb{R}^{q\times q}_s[\zeta,\eta]$. The QDF Q_Φ is called nonnegative, denoted $Q_\Phi \geq 0$, if $Q_\Phi(w) \geq 0$ for all $w \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^q)$.

With every $\Phi \in \mathbb{R}^{q\times q}_s[\zeta,\eta]$ we associate its *coefficient matrix*, which is defined as the infinite symmetric matrix with a finite number of nonzero elements, given by

$$\tilde{\Phi} := \begin{pmatrix} \Phi_{0,0} & \Phi_{0,1} & \dots & \Phi_{0,N} & \dots \\ \Phi_{1,0} & \Phi_{1,1} & \dots & \Phi_{1,N} & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \Phi_{N,0} & \Phi_{N,1} & \dots & \Phi_{N,N} & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}.$$

Clearly, $Q_\Phi \geq 0$ if and only if $\tilde{\Phi} \geq 0$.

The association of two-variable polynomial matrices with QDFs allows us to develop a calculus that has applications in stability theory, optimal control, and H_∞-control (see [18], [16] and [22]). We restrict our attention to a couple of concepts that are used extensively in this paper. One of them is the map $\partial : \mathbb{R}_s^{q\times q}[\zeta,\eta] \longrightarrow \mathbb{R}^{q\times q}[\xi]$, defined by

$$\partial\Phi(\xi) := \Phi(-\xi,\xi).$$

Observe that for every $\Phi \in \mathbb{R}_s^{q\times q}[\zeta,\eta]$, $\partial\Phi$ is para-Hermitian, i.e., $\partial\Phi = (\partial\Phi)^\sim$. Another feature of the calculus of QDFs that is used in this paper is the derivative of a QDF. Given a QDF Q_Φ we define its derivative as the QDF $\frac{d}{dt}Q_\Phi$ defined by $(\frac{d}{dt}Q_\Phi)(w) := \frac{d}{dt}(Q_\Phi(w))$. Q_Φ is called the derivative of Q_Ψ if $\frac{d}{dt}Q_\Psi = Q_\Phi$. In terms of the two-variable polynomial matrices associated with the QDFs, this relationship is expressed equivalently as $(\zeta+\eta)\Psi(\zeta,\eta) = \Phi(\zeta,\eta)$.

In this paper, we also use integrals of QDFs. In order to make sure that the integrals exist, we assume that the trajectories on which the QDF acts are of compact support, that is, they belong to $\mathfrak{D}(\mathbb{R},\mathbb{R}^q)$. Given a QDF Q_Φ, we define its integral as the functional

$$\int Q_\Phi : \mathfrak{D}(\mathbb{R},\mathbb{R}^q) \longrightarrow \mathbb{R},$$
$$\int Q_\Phi(w) = \int_{-\infty}^{+\infty} Q_\Phi(w)dt.$$

Questions such as when the integral of a QDF is a positive semidefinite operator arise naturally in the study of dissipativity. We call a QDF Q_Φ *average nonnegative* if $\int Q_\Phi \geq 0$, i.e., $\int_{-\infty}^{\infty} Q_\Phi(w)dt \geq 0$ for all $w \in \mathfrak{D}(\mathbb{R},\mathbb{R}^q)$. A QDF can be tested for average nonnegativity by analyzing the behavior of the para-Hermitian matrix $\partial\Phi$ on the imaginary axis. Indeed, it is proven in [18] that

$$\int Q_\Phi \geq 0 \Longleftrightarrow \partial\Phi(i\omega) \geq 0 \quad \forall\, \omega \in \mathbb{R}. \tag{6}$$

3. Storage functions and polynomial spectral factorization. In the context of dissipative systems, a QDF measures the power going into a system: its integral over the real line then measures the net flow of energy going into the system. The concept of storage function emerges in the framework of QDFs as follows. Let $\Phi \in \mathbb{R}_s^{q\times q}[\zeta,\eta]$; the QDF Q_Ψ is said to be a *storage function* for Q_Φ (or Ψ is a storage function for Φ) if the following *dissipation inequality* holds:

$$\frac{d}{dt}Q_\Psi \leq Q_\Phi.$$

Storage functions are related to dissipation functions, which we now define. A QDF Q_Δ is a *dissipation function* for Q_Φ (or Δ is a dissipation function for Φ) if $Q_\Delta \geq 0$ and $\int Q_\Phi = \int Q_\Delta$. There is a close relationship between storage functions, average nonnegativity, and dissipation functions.

PROPOSITION 1. *Let $\Phi \in \mathbb{R}_s^{q\times q}[\zeta,\eta]$. The following conditions are equivalent:*

1. $\int Q_\Phi \geq 0$*;*
2. Φ *admits a storage function;*
3. Φ *admits a dissipation function.*

Moreover, there exists a one-to-one relation between storage functions Ψ and dissipation functions Δ for Φ, defined by

$$\frac{d}{dt}Q_\Psi = Q_\Phi - Q_\Delta$$

or, equivalently,

$$(\zeta + \eta)\Psi(\zeta, \eta) = \Phi(\zeta, \eta) - \Delta(\zeta, \eta). \tag{7}$$

Since storage functions measure the energy stored inside a system, it is to be expected that they are related to the memory, to the state, of the system. This intuition has been formalized in [15] in more general terms than those needed in the rest of this paper. For our purposes, the following result from [15] will do.

PROPOSITION 2. *Let $\mathcal{B}$ be represented by $w = M(\frac{d}{dt})l$, and let $X \in \mathbb{R}^{n\times d}[\xi]$ induce a state map for $\mathcal{B}$. Let P be a symmetric $q \times q$ matrix, and define the two-variable polynomial matrix $\Phi(\zeta, \eta) = M^T(\zeta)PM(\eta)$. Let Q_Ψ be a storage function for Q_Φ. Then Q_Ψ is a quadratic function of the state, i.e., there exists a symmetric $n \times n$ matrix K such that $Q_\Psi(l) = \mid X(\frac{d}{dt})l \mid^2_K$ for all $l \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^d)$; equivalently, $\Psi(\zeta, \eta) = X^T(\zeta)KX(\eta)$.*

Given an average nonnegative QDF, in general there exist an infinite number of storage functions. It turns out that they all lie between two extremal storage functions.

PROPOSITION 3. *Let $\int Q_\Phi \geq 0$. Then there exist storage functions Ψ_- and Ψ_+ such that any other storage function Ψ for Φ satisfies*

$$Q_{\Psi_-} \leq Q_\Psi \leq Q_{\Psi_+}.$$

In the following we call Q_{Ψ_-} the *smallest* and Q_{Ψ_+} the *largest storage function* of Q_Φ.

In many cases it is of interest to compute explicitly a storage function for a given QDF. We review here a procedure to compute the extremal storage functions Q_{Ψ_-} and Q_{Ψ_+} introduced in Proposition 3. For this we need to introduce the notion of polynomial spectral factorization of a para-Hermitian polynomial matrix. Let P be a para-Hermitian polynomial matrix. A factorization $P = F^\sim F$, with F a real polynomial matrix, is called a *polynomial spectral factorization* of P, and F is called a *spectral factor* of P. The factorization is called *Hurwitz* if F is square and the roots of $\det(F)$ lie in $\mathbb{C}_-$. It is called *semi-Hurwitz* if the roots of $\det(F)$ lie in $\mathbb{C}^0_-$. The factorization is called *(semi-)anti-Hurwitz* if F is square and the roots of $\det(F)$ lie in $\mathbb{C}_+$ (respectively, in $\mathbb{C}^0_+$). It is well known (see, for example, [10]) that P has a semi-Hurwitz and a (semi-)anti-Hurwitz spectral factorization if and only if $P(i\omega) \geq 0$ for all $\omega \in \mathbb{R}$, and a Hurwitz and an anti-Hurwitz spectral factorization if and only if $P(i\omega) > 0$ for all $\omega \in \mathbb{R}$. The following result shows how to use semi-Hurwitz and semi-anti-Hurwitz polynomial spectral factorizations of $\partial\Phi$ to compute the extremal storage functions of Φ.

PROPOSITION 4. *Let $\Phi(\zeta, \eta) \in \mathbb{R}^{\bullet\times\bullet}_s[\zeta, \eta]$. Assume $\det(\partial\Phi) \neq 0$ and $\partial\Phi(i\omega) \geq 0$ for all $\omega \in \mathbb{R}$. Then the smallest and the largest storage functions Ψ_- and Ψ_+ of Φ can be constructed as follows. Let H and A be semi-Hurwitz, respectively semi-anti-*

Hurwitz, polynomial spectral factors of $\partial\Phi$. *Then*

$$\Psi_+(\zeta,\eta) = \frac{\Phi(\zeta,\eta) - A^T(\zeta)A(\eta)}{\zeta+\eta},$$
$$\Psi_-(\zeta,\eta) = \frac{\Phi(\zeta,\eta) - H^T(\zeta)H(\eta)}{\zeta+\eta}.$$

It also turns out that if P is para-Hermitian, and if $P(i\omega) > 0$ for all $\omega \in \mathbb{R}$, then for every factorization of the scalar polynomial $\det(P)$ as $\det(P) = f^\sim f$, where f and $f^\sim$ have no common roots, there exists a polynomial spectral factorization of P as $P = F^\sim F$, with $\det(F) = f$. This result is taken from [3].

PROPOSITION 5. *Let* $P \in \mathbb{R}^{m\times m}[\xi]$ *be para-Hermitian. Assume that* $P(i\omega) > 0$ *for all* $\omega \in \mathbb{R}$. *Then for every factorization* $\det(P) = f^\sim f$, *with* $f \in \mathbb{R}[\xi]$ *such that* f *and* $f^\sim$ *are coprime, there exists* $F \in \mathbb{R}^{m\times m}[\xi]$ *such that* $P = F^\sim F$ *and* $\det(F) = f$.

4. Pick matrices. In this section we discuss Pick matrices associated with average nonnegative quadratic differential forms. In the following, let $\Phi(\zeta,\eta) \in \mathbb{R}^{q\times q}[\zeta,\eta]$. Assume that $\partial\Phi(i\omega) > 0$ for all $\omega \in \mathbb{R}$. Since $\partial\Phi$ is para-Hermitian, the degree of the polynomial $\det(\partial\Phi)$ is even, say $2n$. Also, $\det(\partial\Phi(i\omega)) > 0$ for all $\omega \in \mathbb{R}$, so $\det(\Phi)$ can be factored as $f^\sim f$ with $f \in \mathbb{R}[\xi]$ such that f and $f^\sim$ are coprime. Of course, for a given Φ there are many f's that satisfy these properties. With any such f, we associate a *Pick matrix*, denoted by T_f.

Pick matrices are most easily introduced in the special case in which the singularities of $\partial\Phi$ are semisimple, i.e., every singularity λ of $\partial\Phi$ has the property that its algebraic multiplicity (its multiplicity as a root of $\det(\partial\Phi)$) is equal to its geometric multiplicity (i.e., $q - \operatorname{rank}(\partial\Phi(\lambda))$, the rank deficiency at λ). For the moment, assume this to be the case.

Let $f \in \mathbb{R}[\xi]$ be such that $\det(\partial\Phi) = f^\sim f$ and $(f, f^\sim)$ coprime. Let $\lambda_1, \lambda_2, \dots, \lambda_n$ be the roots of f. We use the convention that if the algebraic multiplicity of λ_i is m_i, then it appears in this list m_i times, and we have ordered the roots in such a way that $\lambda_1, \lambda_2, \dots, \lambda_{m_1}$ are equal, that $\lambda_{m_1+1}, \lambda_{m_1+2}, \dots, \lambda_{m_1+m_2}$ are equal, etc. Clearly, the other singularities of $\partial\Phi$ are then $-\lambda_1, -\lambda_2, \dots, -\lambda_n$, the roots of $f^\sim$. Now for $i = 1, 2, \dots, n$, let $v_i \in \mathbb{C}^q$ be such that

$$\partial\Phi(\lambda_i)v_i = 0,$$

and such that $v_1, v_2, \dots, v_n$ are linearly independent. The Pick matrix associated with f is now defined as the matrix

$$T_f := \begin{pmatrix} \frac{v_1^*\Phi(\bar\lambda_1,\lambda_1)v_1}{\bar\lambda_1+\lambda_1} & \frac{v_1^*\Phi(\bar\lambda_1,\lambda_2)v_2}{\bar\lambda_1+\lambda_2} & \cdots & \frac{v_1^*\Phi(\bar\lambda_1,\lambda_k)v_n}{\bar\lambda_1+\lambda_n} \\ \frac{v_2^*\Phi(\bar\lambda_2,\lambda_1)v_1}{\bar\lambda_2+\lambda_1} & \frac{v_2^*\Phi(\bar\lambda_2,\lambda_2)v_2}{\bar\lambda_2+\lambda_2} & \cdots & \frac{v_2^*\Phi(\bar\lambda_2,\lambda_k)v_n}{\bar\lambda_2+\lambda_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{v_n^*\Phi(\bar\lambda_n,\lambda_1)v_1}{\bar\lambda_n+\lambda_1} & \frac{v_n^*\Phi(\bar\lambda_n,\lambda_2)v_2}{\bar\lambda_n+\lambda_2} & \cdots & \frac{v_n^*\Phi(\bar\lambda_n,\lambda_n)v_n}{\bar\lambda_n+\lambda_n} \end{pmatrix}. \tag{8}$$

Note that $T_f = T_f^* \in \mathbb{C}^{n\times n}$, where $2n$ is the degree of $\det(\partial\Phi)$. Note that the n functions $e^{\lambda_1 t}v_1, e^{\lambda_2 t}v_2, \dots, e^{\lambda_n t}v_n$ span an n-dimensional subspace of the $2n$-dimensional complex linear space of solutions of the system of differential equations

$$(\partial\Phi)\left(\frac{d}{dt}\right)w = 0. \tag{9}$$

In the general case in which the singularities of $\partial\Phi$ are not all semisimple, the definition of T_f is also straightforward but notationally more involved. We will introduce the Pick matrix in the general case now.

The definition is most easily understood against the background of computing solutions to the system of differential equations (9). In general, a basis for the linear space of solutions of (9) is obtained by analyzing the structure of the singularities of the polynomial matrix $\partial\Phi$. Again let $\det(\partial\Phi) = f^{\sim}f$ be a given factorization, with $\deg(f) = n$. Let $\lambda_1, \lambda_2, \dots, \lambda_k$ be the roots of f. Again, this list of roots does not necessarily consist of distinct complex numbers. In fact, we use the convention that if a given root λ_i has geometric multiplicity n_i, then we include it n_i times in our list of roots. Hence, $\lambda_1, \lambda_2, \dots, \lambda_{n_1}$ are equal, $\lambda_{n_1+1}, \lambda_{n_1+2}, \dots, \lambda_{n_1+n_2}$ are equal, etc. It is well known that there exist integers $d_1, d_2, \dots, d_k \geq 1$ such that $d_1 + d_2 + \cdots + d_{n_1} = m_1$, the algebraic multiplicity of λ_1, $d_{n_1+1} + d_{n_1+2} + \cdots + d_{n_1+n_2} = m_2$, the algebraic multiplicity of λ_{n_1+1}, etc. The sum $\sum_i m_i$ of the algebraic multiplicities is equal to n, the degree of f.

The n-dimensional subspace of solutions of (9) with exponents in $\{\lambda_1, \lambda_2, \dots, \lambda_k\}$ is then computed as follows. Let $\partial\Phi^{(i)}$ be the ith derivative of $\partial\Phi$. For each $i = 1, 2, \dots, k$ there exist d_i complex vectors $a_{i,0}, a_{i,1}, \dots, a_{i,d_i-1} \in \mathbb{C}^q$ such that

(10)
$$\begin{pmatrix} \binom{0}{0}\partial\Phi^{(0)}(\lambda_i) & \binom{1}{0}\partial\Phi^{(1)}(\lambda_i) & \cdots & \cdots & \binom{d_i-1}{0}\partial\Phi^{(d_i-1)}(\lambda_i) \\ 0 & \binom{1}{1}\partial\Phi^{(0)}(\lambda_i) & \cdots & \cdots & \binom{d_i-1}{1}\partial\Phi^{(d_i-2)}(\lambda_i) \\ 0 & 0 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \binom{d_i-1}{d_i-1}\partial\Phi^{(0)}(\lambda_i) \end{pmatrix} \begin{pmatrix} a_{i,0} \\ a_{i,1} \\ \vdots \\ \vdots \\ a_{i,d_i-1} \end{pmatrix} = 0$$

and such that the n vectors $a_{i,j}$ are linearly independent. Using these vectors we form the matrices $V_i \in \mathbb{C}^{d_i q \times d_i}$ defined by

(11)
$$V_i := \begin{pmatrix} \binom{0}{0}a_{i,0} & \binom{1}{1}a_{i,1} & \cdots & \binom{d_i-2}{d_i-2}a_{i,d_i-2} & \binom{d_i-1}{d_i-1}a_{i,d_i-1} \\ \binom{1}{0}a_{i,1} & \binom{2}{1}a_{i,2} & \cdots & \binom{d_i-1}{d_i-2}a_{i,d_i-1} & 0 \\ \vdots & \vdots & & 0 & 0 \\ \binom{d_i-2}{0}a_{i,d_i-2} & \binom{d_i-1}{1}a_{i,d_i-1} & & \vdots & \vdots \\ \binom{d_i-1}{0}a_{i,d_i-1} & 0 & \cdots & 0 & 0 \end{pmatrix}.$$

For $i = 1, 2, \dots, k$, define the matrix function $W_i : \mathbb{R} \to \mathbb{R}^{q \times d_i}$ by

$$W_i(t) := e^{\lambda_i t} \begin{pmatrix} I_{q\times q} & tI_{q\times q} & \cdots & t^{d_i-1}I_{q\times q} \end{pmatrix} V_i,$$

and the matrix function $W : \mathbb{R} \to \mathbb{R}^{q \times n}$ by

$$W(t) := \begin{pmatrix} W_1(t) & W_2(t) & \cdots & W_k(t) \end{pmatrix}.$$

Then the columns of W form a basis for the n-dimensional subspace of solutions of (9) with exponents in $\{\lambda_1, \lambda_2, \dots, \lambda_k\}$.

We now introduce the Pick matrix T_f associated with Φ and the factorization

$\det(\partial\Phi) = f^{\sim}f$. For $i, j = 1, 2, \ldots, k$, define the nonsingular $d_j \times d_j$ matrix $\Lambda_{i,j}$ by

$$\Lambda_{i,j} = \begin{pmatrix} 1 & 0 & 0 & \cdots & \cdots & 0 \\ \frac{-1}{\bar{\lambda}_i+\lambda_j} & 1 & 0 & \cdots & \cdots & 0 \\ \frac{2!}{(\bar{\lambda}_i+\lambda_j)^2} & \frac{-2!}{\bar{\lambda}_i+\lambda_j} & 1 & 0 & \cdots & 0 \\ \frac{-3!}{(\bar{\lambda}_i+\lambda_j)^3} & \frac{3!}{(\bar{\lambda}_i+\lambda_j)^2} & \frac{-3!}{\bar{\lambda}_i+\lambda_j} & 1 & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & 0 \\ \frac{(-1)^{d_j-1}(d_j-1)!}{(\bar{\lambda}_i+\lambda_j)^{d_j-1}} & \cdots & \cdots & \frac{(d_j-1)!}{(\bar{\lambda}_i+\lambda_j)^2} & \frac{-(d_j-1)!}{\bar{\lambda}_i+\lambda_j} & 1 \end{pmatrix}. \tag{12}$$

Also, for $i, j = 1, 2, \ldots, k$ we define $\Theta_{i,j} \in \mathbb{C}^{d_i q \times d_j q}$ by

$$\Theta_{i,j} := \begin{pmatrix} \Phi(\bar{\lambda}_i, \lambda_j) & \frac{\partial\Phi}{\partial\eta}(\bar{\lambda}_i, \lambda_j) & \cdots & \frac{\partial^{d_j-1}\Phi}{\partial\eta^{d_j-1}}(\bar{\lambda}_i, \lambda_j) \\ \frac{\partial\Phi}{\partial\zeta}(\bar{\lambda}_i, \lambda_j) & \frac{\partial^2\Phi}{\partial\zeta\partial\eta}(\bar{\lambda}_i, \lambda_j) & \cdots & \frac{\partial^{d_j}\Phi}{\partial\zeta\partial\eta^{d_j-1}}(\bar{\lambda}_i, \lambda_j) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^{d_i-1}\Phi}{\partial\zeta^{d_i-1}}(\bar{\lambda}_i, \lambda_j) & \frac{\partial^{d_i}\Phi}{\partial\zeta^{d_i-1}\partial\eta}(\bar{\lambda}_i, \lambda_j) & \cdots & \frac{\partial^{d_i+d_j-2}\Phi}{\partial\zeta^{d_i-1}\partial\eta^{d_j-1}}(\bar{\lambda}_i, \lambda_j) \end{pmatrix}. \tag{13}$$

Here, $\frac{\partial^{i+j}\Phi}{\partial\zeta^i\partial\eta^j}(\zeta, \eta)$ denotes the (i, j)th partial derivative with respect to ζ and η of $\Phi(\zeta, \eta)$. We define the shift operator $\sigma : \mathbb{C}^{d_i q \times d_j q} \to \mathbb{C}^{d_i q \times d_j q}$ acting on matrices M that are partitioned into $q \times q$ blocks as follows: if

$$M = \begin{pmatrix} M_{1,1} & M_{1,2} & \cdots & M_{1,d_j} \\ M_{2,1} & M_{2,2} & \cdots & M_{2,d_j} \\ \vdots & \vdots & \ddots & \vdots \\ M_{d_i,1} & M_{d_i,2} & \cdots & M_{d_i,d_j}, \end{pmatrix},$$

then

$$\sigma(M) := \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & M_{1,1} & \cdots & M_{1,d_j-1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & M_{d_i-1,1} & \cdots & M_{d_i-1,d_j-1} \end{pmatrix}.$$

In terms of $\Theta_{i,j}$ and the shift-operator σ, for $i, j = 1, 2, \ldots, k$, we define the matrices $\Sigma_{i,j} \in \mathbb{C}^{d_i q \times d_j q}$ by

$$\begin{aligned} \Sigma_{i,j} := \frac{1}{\bar{\lambda}_i + \lambda_j}\Theta_{i,j} + \frac{1}{(\bar{\lambda}_i + \lambda_j)^2}\sigma(\Theta_{i,j}) + \frac{1}{(\bar{\lambda}_i + \lambda_j)^3}\sigma^2(\Theta_{i,j}) + \cdots \\ + \frac{1}{(\bar{\lambda}_i + \lambda_j)^{\max(d_i,d_j)-1}}\sigma^{\max(d_i,d_j)-1}(\Theta_{i,j}). \end{aligned} \tag{14}$$

Here, for a given M, $\sigma^2(M)$ is defined as $\sigma(\sigma(M))$, etc. The Pick matrix associated with Φ and the factorization $\det(\partial\Phi) = f^{\sim}f$ is now defined as the matrix $T_f \in \mathbb{C}^{n\times n}$, $T_f = (T_{i,j})_{i,j=1,2,\ldots,k}$, where the (i, j)th block is the complex $d_i \times d_j$ matrix given by

$$T_{i,j} := \Lambda_{j,i}^* V_i^* \Sigma_{i,j} V_j \Lambda_{i,j}. \tag{15}$$

Note that T_f is a Hermitian matrix.

For related material on Pick matrices, their application in interpolation problems, and connections with systems and control, see [4], [25].

5. The Riccati equation, linear matrix inequalities, and storage functions. In this section we study the connection between the existence of real symmetric solutions of the ARE and average nonnegativity of a given QDF associated with the ARE.

We associate with the ARE (1) the system with manifest variable $w = \mathrm{col}(x, u)$ represented by $\frac{d}{dt}x = Ax + Bu$, or equivalently

$$\begin{pmatrix} \frac{d}{dt}I_n - A & -B \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix} = 0. \tag{16}$$

Equation (16) constitutes a kernel representation of the behavior

$$\mathfrak{B} = \{\mathrm{col}(x, u) \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^n) \times \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^m) \mid \text{(16) is satisfied}\}. \tag{17}$$

A standing assumption in the remainder of this paper is that the pair (A, B) is controllable. Under this assumption, $\mathfrak{B}$ can be represented in image form. One such representation can be computed as follows. Let $X \in \mathbb{R}^{n\times m}[\xi]$ and $U \in \mathbb{R}^{m\times m}[\xi]$ induce a right coprime factorization of the rational matrix $(\xi I_n - A)^{-1}B$, i.e., $(\xi I_n - A)^{-1}B = X(\xi)U(\xi)^{-1}$ and

$$\mathrm{rank} \begin{pmatrix} X(\lambda) \\ U(\lambda) \end{pmatrix} = m$$

for all $\lambda \in \mathbb{C}$. Then $\mathfrak{B}$ is represented in observable image form as

$$\begin{pmatrix} x \\ u \end{pmatrix} = \begin{pmatrix} X(\frac{d}{dt}) \\ U(\frac{d}{dt}) \end{pmatrix} l. \tag{18}$$

Observe that any such X yields a minimal state map $X(\frac{d}{dt})$ for $\mathfrak{B}$.

Given the matrices $Q = Q^T \in \mathbb{R}^{n\times n}$, $R = R^T \in \mathbb{R}^{m\times m}$, and $S \in \mathbb{R}^{m\times n}$, and the polynomial matrices X and U, we define the symmetric $m\times m$ two-variable polynomial matrix Φ by

$$\Phi(\zeta, \eta) = \begin{pmatrix} X(\zeta)^T & U(\zeta)^T \end{pmatrix} \begin{pmatrix} Q & S^T \\ S & R \end{pmatrix} \begin{pmatrix} X(\eta) \\ U(\eta) \end{pmatrix}. \tag{19}$$

Note that if l and $\mathrm{col}(x, u)$ are related by (18), then the QDF Q_Φ associated with Φ satisfies

$$Q_\Phi(l) = \begin{pmatrix} x^T & u^T \end{pmatrix} \begin{pmatrix} Q & S^T \\ S & R \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix}.$$

Of course, $(\xi I_n - A)^{-1}B$ admits many right coprime factorizations. If $X_1U_1^{-1} = X_2U_2^{-1}$ are two right coprime factorizations, then they are related by a unimodular transformation: there exists a unimodular V such that $X_2 = X_1V$ and $U_2 = U_1V$. Hence the associated two-variable polynomial matrices are related by $\Phi_1(\zeta, \eta) = V^T(\zeta)\Phi_2(\zeta, \eta)V(\eta)$.

Example 6. In the Riccati equation (1), let $A = \left(\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right), B = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right), Q = \left(\begin{smallmatrix} 1 & a \\ a & 3 \end{smallmatrix}\right), R = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, and $S = \left(\begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix}\right)$. Here a is a parameter taking values in $\mathbb{R}$. Clearly, $(\xi I - A)^{-1}B = X(\xi)U(\xi)^{-1}$, with $X(\xi) = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, and $U(\xi) = \left(\begin{smallmatrix} \xi & 0 \\ 0 & \xi - 1 \end{smallmatrix}\right)$. The corresponding two-variable polynomial matrix is $\Phi(\zeta, \eta) = \left(\begin{smallmatrix} 1+\zeta\eta & a \\ a & 3+(\zeta-1)(\eta-1) \end{smallmatrix}\right)$.

The next result connects the average nonnegativity of the QDF associated with (19) with the existence of real symmetric solutions to the linear matrix inequality associated with the ARE (1) and with the existence of storage functions for Q_Φ.

THEOREM 7. *Let $\Phi(\zeta,\eta)$ be defined by (19), where X and U are such that $X(\xi)U(\xi)^{-1}$ is a right coprime factorization of $(\xi I_n - A)^{-1}B$. Then the following statements are equivalent:*

1. *$\int Q_\Phi \geq 0$;*
2. *there exists $K = K^T \in \mathbb{R}^{n\times n}$ such that $\mid X(\frac{d}{dt})l \mid_K^2$ is a storage function for Q_Φ;*
3. *there exists $K = K^T \in \mathbb{R}^{n\times n}$ such that the $(n+m)\times(n+m)$ symmetric matrix*

$$L(K) := \begin{pmatrix} Q - A^TK - KA & -KB + S^T \\ -B^TK + S & R \end{pmatrix}$$

satisfies $L(K) \geq 0$.

In fact, for every $K = K^T \in \mathbb{R}^{n\times n}$ there holds

$$\frac{d}{dt}\left|X\left(\frac{d}{dt}\right)l\right|_K^2 = Q_\Phi(l) - \left|\begin{pmatrix} X(\frac{d}{dt})l \\ U(\frac{d}{dt})l \end{pmatrix}\right|_{L(K)}^2 \tag{20}$$

for all $l \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^m)$; equivalently,

$$(\zeta+\eta)X^T(\zeta)KX(\eta) = \Phi(\zeta,\eta) - \begin{pmatrix} X(\zeta)^T & U(\zeta)^T \end{pmatrix} L(K) \begin{pmatrix} X(\eta) \\ U(\eta) \end{pmatrix}.$$

Consequently, for $K = K^T \in \mathbb{R}^{n\times n}$ the following statements are equivalent:

(i) *$L(K) \geq 0$;*

(ii) *$\mid X(\frac{d}{dt})l \mid_K^2$ is a storage function for Q_Φ;*

(iii)

$$\left|\begin{pmatrix} X(\frac{d}{dt})l \\ U(\frac{d}{dt})l \end{pmatrix}\right|_{L(K)}^2$$

is a dissipation function for Q_Φ.

Proof. We prove the equivalence of (i),(ii), and (iii). The first part of the theorem follows easily from this and from Proposition 1. We need the following lemma.

LEMMA 8. *Let $X \in \mathbb{R}^{n\times m}[\xi]$ and $U \in \mathbb{R}^{m\times m}[\xi]$ be such that $X(\xi)U(\xi)^{-1}$ is a right coprime factorization of $(\xi I - A)^{-1}B$. Then the mapping*

$$\begin{aligned} \mathcal{C}^\infty(\mathbb{R},\mathbb{R}^m) &\to \mathbb{R}^n \times \mathbb{R}^m, \\ l &\mapsto \begin{pmatrix} (X(\frac{d}{dt})l)(0) \\ (U(\frac{d}{dt})l)(0) \end{pmatrix} \end{aligned}$$

is surjective.

Proof of Lemma 8. Let $(x_0,u_0) \in \mathbb{R}^n \times \mathbb{R}^m$. Let $\tilde{u} \in \mathcal{C}^\infty(\mathbb{R},\mathbb{R}^m)$ be such that $\tilde{u}(0) = u_0$. Consider the differential equation $\dot{x} = Ax + B\tilde{u}$, $x(0) = x_0$, and let $\tilde{x} \in \mathcal{C}^\infty(\mathbb{R},\mathbb{R}^n)$ be its solution. Evidently,

$$\operatorname{col}(\tilde{x},\tilde{u}) \in \operatorname{im}\begin{pmatrix} X(\frac{d}{dt}) \\ U(\frac{d}{dt}) \end{pmatrix},$$

so there exists $l \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^m)$ such that

$$\begin{pmatrix} \tilde{x} \\ \tilde{u} \end{pmatrix} = \begin{pmatrix} X(\frac{d}{dt}) \\ U(\frac{d}{dt}) \end{pmatrix} l.$$

Consequently,

$$\begin{pmatrix} x_0 \\ u_0 \end{pmatrix} = \begin{pmatrix} \tilde{x}(0) \\ \tilde{u}(0) \end{pmatrix} = \begin{pmatrix} (X(\frac{d}{dt})l)(0) \\ (U(\frac{d}{dt})l)(0) \end{pmatrix}.$$

This concludes the proof of the lemma.

We resume the proof of Theorem 7. Let $K = K^T \in \mathbb{R}^{n\times n}$. We first prove that, for all l, (20) holds; equivalently,

$$(\zeta + \eta)X(\zeta)^T K X(\eta) = \Phi(\zeta, \eta) - \begin{pmatrix} X(\zeta)^T & U(\zeta)^T \end{pmatrix} L(K) \begin{pmatrix} X(\eta) \\ U(\eta) \end{pmatrix}. \tag{21}$$

Indeed, from the fact that $X(\xi)U(\xi)^{-1} = (\xi I - A)^{-1}B$ it follows that $\xi X(\xi) = AX(\xi) + BU(\xi)$. Consequently,

$$\begin{aligned}(\zeta + \eta)X(\zeta)^T K X(\eta) = {} & X(\zeta)^T A^T K X(\eta) + U(\zeta)^T B^T K X(\eta) \\ & + X(\zeta)^T K A X(\eta) + X(\zeta)^T K B U(\eta),\end{aligned}$$

which can be rewritten as

$$\begin{pmatrix} X(\zeta)^T & U(\zeta)^T \end{pmatrix} \begin{pmatrix} A^T K + KA & KB \\ B^T K & 0 \end{pmatrix} \begin{pmatrix} X(\eta) \\ U(\eta) \end{pmatrix}.$$

With $\Phi(\zeta, \eta)$ defined by (19), equation (21) then follows immediately. Since, by Lemma 8, the map

$$l \mapsto \begin{pmatrix} (X(\frac{d}{dt})l)(0) \\ (U(\frac{d}{dt})l)(0) \end{pmatrix}$$

is surjective, we have

$$\left| \begin{pmatrix} X(\frac{d}{dt})l \\ U(\frac{d}{dt})l \end{pmatrix} \right|^2_{L(K)} \geq 0 \quad \forall\, l \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^m)$$

if and only if $L(K) \geq 0$. Thus the equivalence of (i), (ii), and (iii) follows immediately from (20).

The equivalence of statements 1. and 2. follows from Propositions 1 and 2. The equivalence of 2. and 3. is an immediate consequence of the equivalence of (i) and (ii). □

If we assume that the matrix R is positive definite, the result of Theorem 7 can be sharpened, and a connection can be established between the QDF $\Phi(\zeta, \eta)$ defined in (19) and the ARE (1).

THEOREM 9. *Let $\Phi(\zeta, \eta)$ be defined by (19), where X and U are such that $X(\xi)U(\xi)^{-1}$ is a right coprime factorization of $(\xi I_n - A)^{-1}B$. Assume $R > 0$. Then the following statements are equivalent:*

1. $\int Q_\Phi \geq 0$;
2. *There exists a real symmetric solution to the* ARE.

In fact, for every $K = K^T \in \mathbb{R}^{n\times n}$ *the following conditions are equivalent:*

(i) $-K$ *satisfies the* ARE;

(ii) $\mid X(\frac{d}{dt})l \mid^2_K$ *is a storage function for* Q_Φ *with associated dissipation function*

$$\Delta(\zeta,\eta) = \begin{pmatrix} X(\zeta)^T & U(\zeta)^T \end{pmatrix} L(K) \begin{pmatrix} X(\eta) \\ U(\eta) \end{pmatrix} = F(\zeta)^T F(\eta),$$

where

$$F(\xi) := R^{-\frac{1}{2}}(-B^TK + S)X(\xi) + R^{\frac{1}{2}}U(\xi);$$

(iii) $\mid X(\frac{d}{dt})l \mid^2_K$ *is a storage function for* Q_Φ*, and the rank of the coefficient matrix of the QDF* $Q_\Phi(l) - \frac{d}{dt} \mid X(\frac{d}{dt})l \mid^2_K$ *is equal to* m.

Proof. We begin by proving the implication 1. $\Rightarrow$ 2. From condition 1. and the equivalence in (6), we obtain $\partial\Phi(i\omega) \geq 0$ for all $\omega \in \mathbb{R}$. Since $\det(\partial\Phi) \neq 0$, there exists a semi-Hurwitz factorization $\partial\Phi = H^\sim H$, with $\det(H) \neq 0$. According to Proposition 4, this yields the smallest storage function as induced by the two-variable polynomial matrix

$$\Psi_-(\zeta,\eta) = \frac{\Phi(\zeta,\eta) - H^T(\zeta)H(\eta)}{\zeta + \eta}.$$

By Proposition 2, there exists $K = K^T \in \mathbb{R}^{n\times n}$ such that $\Psi_-(\zeta,\eta) = X^T(\zeta)KX(\eta)$. We claim that $-K$ satisfies the ARE. Indeed, as in the proof of Theorem 7, we have

$$H^T(\zeta)H(\eta) = \begin{pmatrix} X(\zeta)^T & U(\zeta)^T \end{pmatrix} L(K) \begin{pmatrix} X(\eta) \\ U(\eta) \end{pmatrix}. \tag{22}$$

Since $\det(H) \neq 0$, the coefficient matrix $\tilde{H}$ of H has full row rank m. Since by Lemma 8 the mapping $l \mapsto \mathrm{col}((X(\frac{d}{dt})l)(0), (U(\frac{d}{dt})l)(0))$ is surjective, the coefficient matrix of $\mathrm{col}(X(\eta), U(\eta))$ has full row rank. Consequently, $L(K)$ has rank m. Since $R > 0$, $\mathrm{rank}(L(K)) = m$ if and only if the Schur complement of R in $L(K)$ is zero, that is, if and only if

$$Q - A^TK - KA - (-KB + S^T)R^{-1}(-B^TK + S) = 0,$$

in other words, if and only if $-K$ satisfies the ARE. This concludes the proof of the implication 1. $\Rightarrow$ 2. The implication 2. $\Rightarrow$ 1. follows from the implication (i) $\Rightarrow$ (ii) below.

Next, we prove the equivalence of (i), (ii), and (iii) of Theorem 9.

(i) $\Rightarrow$ (ii). Assume $-K$ satisfies the ARE. Then it is easily seen that

$$L(K) = \begin{pmatrix} R^{-1/2}(-B^TK + S) & R^{1/2} \end{pmatrix}^T \begin{pmatrix} R^{-1/2}(-B^TK + S) & R^{1/2} \end{pmatrix} \geq 0.$$

From Theorem 7 it then follows that $\mid X(\frac{d}{dt})l \mid^2_K$ is a storage function for Q_Φ, with associated dissipation function $F^T(\zeta)F(\eta)$, where $F(\xi) = R^{-1/2}(-B^TK + S)X(\xi) + R^{1/2}U(\xi)$.

(ii) $\Rightarrow$ (iii). According to (ii), $\mathrm{rank}(L(K)) = m$. The coefficient matrix of the QDF $Q_\Phi(l) - \frac{d}{dt} \mid X(\frac{d}{dt})l \mid^2_K$ is equal to

$$\begin{pmatrix} \tilde{X} \\ \tilde{U} \end{pmatrix}^T L(K) \begin{pmatrix} \tilde{X} \\ \tilde{U} \end{pmatrix},$$

with $\mathrm{col}(\tilde{X}, \tilde{U})$ the coefficient matrix of $\mathrm{col}(X, U)$. By Lemma 8 this coefficient matrix has full row rank. This proves the implication.

(iii) $\Rightarrow$ (i). Assume $L(K)$ has rank m. Since $\mathrm{rank}(R) = m$, this implies that the Schur complement of R is equal to zero, equivalently that $-K$ satisfies the ARE. □

Example 6, continued. For the Riccati equation of Example 6 we have $\partial\Phi(\xi) = \left(\begin{smallmatrix} 1-\xi^2 & a \\ a & 4-\xi^2 \end{smallmatrix}\right)$, so $\partial\Phi(i\omega) = \left(\begin{smallmatrix} 1+\omega^2 & a \\ a & 4+\omega^2 \end{smallmatrix}\right)$. By (6), the Riccati equation has a real symmetric solution if and only if $\partial\Phi(i\omega) \geq 0$ for all $\omega \in \mathbb{R}$. This holds if and only if $-2 \leq a \leq 2$.

Connections between the ARE and the linear matrix inequality in statement 3. of Theorem 7 are well-known. See, for example, Chap. 8 of [2], where solutions K of the linear matrix inequality such that $\mathrm{rank}(L(K)) = m$ are called *rank minimizing*. In a behavioral framework, connections between such concepts and storage functions were established in [14]; see also Chapter 5 of [11].

6. Pick matrices and the algebraic Riccati equation. In this section we derive the main result of this paper, a characterization of all unmixed real symmetric solutions of the ARE in terms of the Pick matrices associated with the two-variable polynomial matrix (19). As a corollary of this result, we obtain necessary and sufficient conditions for the existence of sign-definite solutions of the ARE. These conditions are in terms of the Pick matrices associated with the Hurwitz and anti-Hurwitz factorizations of $\det(\partial\Phi)$.

In this section, let $X(\xi)U(\xi)^{-1}$ be an arbitrary right coprime factorization of $(\xi I_n - A)^{-1}B$, and let the two-variable polynomial matrix Φ associated with the ARE be given by (19). From the fact that $\partial\Phi$ is para-Hermitian, we know that $\det(\partial\Phi)$ has even degree. In fact, the degree of $\det(\partial\Phi)$ is twice the dimension of the underlying state space system (17).

LEMMA 10. *Let $\Phi(\zeta, \eta)$ be defined as in* (19)*, and assume $R > 0$. Then the degree of* $\det(\partial\Phi)$ *is* $2n$.

Proof. Observe that $\partial\Phi = X^\sim QX + X^\sim S^T U + U^\sim SX + U^\sim RU$. Multiplying this equality on the right by U^{-1} and on the left by $(U^\sim)^{-1}$ yields $(U^\sim)^{-1}\partial\Phi U^{-1} = (U^\sim)^{-1}X^\sim QXU^{-1} + (U^\sim)^{-1}X^\sim S^T + SXU^{-1} + R$. Now observe that $X(\xi)U^{-1}(\xi) = (\xi I_n - A)^{-1}B$ is a matrix of strictly proper rational functions. It follows that $(U^\sim)^{-1}\partial\Phi U^{-1}$ is a matrix of proper rational functions and consequently $\deg(\det(\partial\Phi)) \leq \deg(\det(U)) + \deg(\det(U^\sim)) = 2n$. We now show that $\deg(\det(\partial\Phi)) = 2n$. Indeed, since $\lim_{|\lambda|\longrightarrow\infty}(U^\sim(\lambda))^{-1}\partial\Phi(\lambda)U(\lambda) = R > 0$, it follows that $(U^\sim)^{-1}\partial\Phi U$ has an inverse whose entries are also proper rational functions. Consequently $\deg(\det(\partial\Phi)) = 2n$. □

Assume now that $\int Q_\Phi \geq 0$, equivalently $\partial\Phi(i\omega) \geq 0$, for all $\omega \in \mathbb{R}$ (see (6)). According to Theorem 9 this is equivalent to the existence of a real symmetric solution of the ARE. Observe that every polynomial spectral factorization of $\partial\Phi$ as $\partial\Phi = F^\sim F$ with $F \in \mathbb{R}^{m\times m}[\xi]$ yields a factorization of $\det(\partial\Phi)$ as $\det(\partial\Phi) = f^\sim f$, with $f = \det(F)$ and $\deg(f) = n$. Let $\mathcal{F}$ be the set of all polynomials of degree n, with positive highest degree coefficient, that can occur as the determinant of a polynomial spectral factor of $\partial\Phi$:

$$\mathcal{F} := \{f \in \mathbb{R}[\xi] \mid f(\xi) = f_0 + f_1\xi + \cdots + f_n\xi^n, \quad f_n > 0,$$
$$\text{(23)} \quad \text{and there exists } F \in \mathbb{R}^{m\times m}[\xi] \text{ such that } \partial\Phi = F^\sim F \text{ and } \det(F) = f\}.$$

Also, let $\mathcal{S}$ be the set of all real symmetric solutions of the ARE:

$$\mathcal{S} := \{K \in \mathbb{R}^{n\times n} \mid K = K^T \text{ and } K \text{ satisfies the ARE}\}.$$

For any $K \in \mathcal{S}$, denote $A_K := A - BR^{-1}(B^TK + S)$ and let χ_{A_K} be the characteristic polynomial of A_K. Our basic result states that there is a one-to-one correspondence between $\mathcal{F}$ and $\mathcal{S}$.

THEOREM 11. *$\mathcal{S} \neq \emptyset$ if and only if $\partial\Phi(i\omega) \geq 0$ for all $\omega \in \mathbb{R}$. In that case there exists a bijection between $\mathcal{F}$ and $\mathcal{S}$. Such bijection* $\mathrm{Ric} : \mathcal{F} \to \mathcal{S}$ *is defined as follows. For any $f \in \mathcal{F}$, let $F \in \mathbb{R}^{m\times m}[\xi]$ be such that $f = \det(F)$ and $\partial\Phi = F^{\sim}F$. Then define* $\mathrm{Ric}(f) = K$, *where $K = K^T \in \mathbb{R}^{n\times n}$ is the unique solution of*

$$\frac{\Phi(\zeta,\eta) - F^T(\zeta)F(\eta)}{\zeta + \eta} = X^T(\zeta)(-K)X(\eta). \tag{24}$$

For any $K \in \mathcal{S}$ we have $\partial\Phi = (F_K)^{\sim}F_K$, where

$$F_K(\xi) := R^{-1/2}(B^TK + S)X(\xi) + R^{1/2}U(\xi).$$

Furthermore, for any $K \in \mathcal{S}$ we have $\det(F_K) = \sqrt{\det(R)}\ \chi_{A_K}$, *whence* $\det(\partial\Phi) = \det(R)\ (\chi_{A_K})^{\sim}\chi_{A_K}$ *and*

$$K = \mathrm{Ric}(\sqrt{\det(R)}\ \chi_{A_K}).$$

Proof. We begin by showing that the map $\mathrm{Ric} : \mathcal{F} \to \mathcal{S}$ is well defined. Let f_1 and f_2 be two elements of $\mathcal{F}$, and let F_1, $F_2 \in \mathbb{R}^{m\times m}[\xi]$ be such that $\partial\Phi = F_1^{\sim}F_1 = F_2^{\sim}F_2$ and $\det(F_i) = f_i$, $i = 1, 2$. It is well known (see, for example, Theorem 5.3 of [10]) that there exists an orthogonal $m \times m$ matrix L such that $F_2 = LF_1$. Now let K_1 and K_2 be $n \times n$ symmetric matrices such that

$$\frac{\Phi(\zeta,\eta) - F_i^T(\zeta)F_i(\eta)}{\zeta + \eta} = X^T(\zeta)(-K_i)X(\eta),$$

$i = 1, 2$. (Such matrices exist because of Theorem 2.) Then necessarily

$$X^T(\zeta)(-K_1)X(\eta) = X^T(\zeta)(-K_2)X(\eta).$$

It follows from the fact that $X(\frac{d}{dt})$ is a minimal state map that the map $l \to (X(\frac{d}{dt})l)(0)$ is surjective. Hence for all $x_0 \in \mathbb{R}^n$ there holds $x_0^T(-K_1)x_0 = x_0^T(-K_2)x_0$, which implies $K_1 = K_2$. This shows that Ric is well defined.

We proceed to show that Ric is bijective. We first prove that it is injective. Assume that $\mathrm{Ric}(f_1) = K_1 = \mathrm{Ric}(f_2) = K_2$. Let F_1 and F_2 be $m \times m$ polynomial matrices such that $\partial\Phi = F_1^{\sim}F_1 = F_2^{\sim}F_2$ and $\det(F_i) = f_i$, $i = 1, 2$. From the fact that $K_1 = K_2$ and from (24) it follows that $F_1^T(\zeta)F_1(\eta) = F_2^T(\zeta)F_2(\eta)$. This implies that

$$\det(F_1(\zeta))\det(F_1(\eta)) = \det(F_2(\zeta))\det(F_2(\eta)),$$

so that $f_1(\zeta)f_1(\eta) = f_2(\zeta)f_2(\eta)$. Given that the highest degree coefficient of f_1 and f_2 is positive (see (23)), we conclude that $f_1 = f_2$. This concludes the proof of the injectivity of Ric. In order to prove that Ric is surjective, let $K = K^T$ be a solution to the ARE. According to Theorem 9 there holds

$$(\zeta + \eta)X^T(\zeta)(-K)X(\eta) = \Phi(\zeta,\eta) - F_K(\zeta)^TF_K(\eta),$$

where $F_K \in \mathbb{R}^{m\times m}[\xi]$ is defined by

$$F_K(\xi) = R^{-\frac{1}{2}}(B^TK + S)X(\xi) + R^{\frac{1}{2}}U(\xi).$$

Note that $\partial\Phi = (F_K)^\sim F_K$. Define now $f := \det(F_K)$. Then $K = \mathrm{Ric}(f)$. This also proves the second statement of the theorem.

Next we prove that for all $K \in \mathcal{S}$ we have $\det(F_K) = \det(R^{1/2})\chi_{A_K}$. Consider the $(n+m)\times(n+m)$ polynomial matrix

$$P(\xi) := \begin{pmatrix} \xi I - A & B \\ -R^{-1/2}(B^TK + S) & R^{1/2} \end{pmatrix}.$$

Computing the determinant of P yields

$$\begin{aligned}\det(P(\xi)) &= \det(\xi I - A)\det(R^{1/2} + R^{-1/2}(B^TK + S)(\xi I - A)^{-1}B) \\ &= \det(R^{1/2})\det(\xi I - A + BR^{-1}(B^TK + S)).\end{aligned}$$

Using the fact that $X(\xi)U(\xi)^{-1}$ is a right coprime factorization of $(\xi I - A)^{-1}B$ and that (A, B) is a controllable pair, we have $\det(U(\xi)) = \det(\xi I - A)$, so we obtain $\det(R^{1/2})\chi_{A_K} = \det(F_K)$. The remaining statements of the theorem follow immediately from this. □

In the above, we have assumed that $\partial\Phi(i\omega) \geq 0$ for all $\omega \in \mathbb{R}$. In the case that, in addition, $\partial\Phi$ is nonsingular along the imaginary axis, equivalently $\partial\Phi(i\omega) > 0$ for all $\omega \in \mathbb{R}$, the one-to-one correspondence between polynomials and the set of real symmetric solutions of the ARE can be made even more explicit. This will be explained next.

Define $\mathcal{F}_{\mathrm{cop}}$ as the set of all real polynomials f such that the *determinant* of $\partial\Phi$ admits a factorization $f^\sim f$ such that f and $f^\sim$ are coprime:

$$\mathcal{F}_{\mathrm{cop}} = \{f \in \mathbb{R}[\xi] \mid f(\xi) = f_0 + f_1\xi + \cdots + f_n\xi^n, \quad f_n > 0,\ (f, f^\sim) \text{ coprime and } \det(\partial\Phi) = f^\sim f\}.$$

It is easily seen that if $\partial\Phi(i\omega) \geq 0$ for all $\omega \in \mathbb{R}$, then $\mathcal{F}_{\mathrm{cop}} \neq \emptyset$ if and only if $\partial\Phi(i\omega) > 0$ for all $\omega \in \mathbb{R}$. Hence it follows from Proposition 5 that $\mathcal{F}_{\mathrm{cop}} \subset \mathcal{F}$. In the remainder of this section we assume that $\partial\Phi(i\omega) > 0$ for all $\omega \in \mathbb{R}$.

Note that if $f \in \mathcal{F}_{\mathrm{cop}}$ and $K = \mathrm{Ric}(f)$, then, according to Theorem 11, $f = \sqrt{\det(R)}\chi_{A_K}$, so χ_{A_K} and $(\chi_{A_K})^\sim$ are coprime; equivalently, $\sigma(A_K) \cap \sigma(-A_K) = \emptyset$. If a solution K of the ARE satisfies this property, we call it *unmixed*. The set of all unmixed solutions of the ARE is denoted by $\mathcal{S}_{\mathrm{unm}}$. It follows immediately from Theorem 11 that Ric defines a bijection between $\mathcal{F}_{\mathrm{cop}}$ and $\mathcal{S}_{\mathrm{unm}}$.

We now explain the connection between the bijection Ric and the Pick matrices T_f associated with Φ. Recall that the bijection Ric between $\mathcal{F}_{\mathrm{cop}}$ and $\mathcal{S}_{\mathrm{unm}}$ is defined as follows. For a given $f \in \mathcal{F}_{\mathrm{cop}}$, let $F \in \mathbb{R}^{m\times m}[\xi]$ be such that $\partial\Phi = F^\sim F$ and $\det(F) = f$, and take $K = \mathrm{Ric}(f)$ to be the unique solution of (24). For the sake of exposition, assume for the moment that the singularities of $\partial\Phi$ are semisimple. We show how to compute, for $f \in \mathcal{F}_{\mathrm{cop}}$, the corresponding unmixed solution $K = \mathrm{Ric}(f)$, using the Pick matrix T_f.

Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the roots of f, with the convention that if a root has algebraic multiplicity m_i, then it appears in this list m_i times, and $\lambda_1, \lambda_2, \ldots, \lambda_{m_1}$ are equal, $\lambda_{m_1+1}, \lambda_{m_1+2}, \ldots, \lambda_{m_1+m_2}$ are equal, etc. Let $v_i \in \mathbb{C}^m$ be such that $\partial\Phi(\lambda_i)v_i =$

0 and $v_1, v_2, \dots, v_n$ are linearly independent. Evaluating (24) at $(\zeta, \eta) = (\bar{\lambda}_i, \lambda_j)$, premultiplying the result by v_i^* and postmultiplying it by v_j, we get

$$\frac{v_i^* \Phi(\bar{\lambda}_i, \lambda_j) v_j}{\bar{\lambda}_i + \lambda_j} - \frac{v_i^* F^T(\bar{\lambda}_i) F(\lambda_j) v_j}{\bar{\lambda}_i + \lambda_j} = -v_i^* X^T(\bar{\lambda}_i) K X(\lambda_j) v_j.$$

Note that, by coprimeness of f and $f^\sim$, $\bar{\lambda}_i + \lambda_j \neq 0$ for all (i, j). Now make the crucial observation that for all j

$$F(\lambda_j) v_j = 0.$$

Indeed, by definition of v_j we have $F^T(-\lambda_j) F(\lambda_j) v_j = 0$. Since, however, $F^T(-\lambda_j)$ is nonsingular (by coprimeness of f and $f^\sim$), the claim follows. Thus we immediately obtain

$$\frac{v_i^* \Phi(\bar{\lambda}_i, \lambda_j) v_j}{\bar{\lambda}_i + \lambda_j} = -v_i^* X^T(\bar{\lambda}_i) K X(\lambda_j) v_j,$$

which is equivalent to

$$T_f = -(S_f)^* K S_f,$$

where S_f is the *zero state matrix* associated with f, defined by

$$S_f := \begin{pmatrix} X(\lambda_1)v_1 & \dots & X(\lambda_n)v_n \end{pmatrix}.$$

For a motivation of the terminology zero state matrix, we refer to the proof of Theorem 12 below. Note that $S_f \in \mathbb{C}^{n \times n}$. In Theorem 12 we will prove that for any $f \in \mathcal{F}_{\text{cop}}$ the zero state matrix S_f is nonsingular. This immediately implies that the solution $K = \text{Ric}(f)$ is given by

$$K = \text{Ric}(f) = -(S_f^*)^{-1} T_f (S_f)^{-1}. \tag{25}$$

The above argument can be generalized to the case in which not all singularities of $\partial\Phi$ are semisimple. In the general case, the zero state matrix S_f associated with the polynomial factor f is defined in the following way. Let $\lambda_1, \lambda_2, \dots, \lambda_k$ be the roots of f. As in section 4, we use the convention that if a given root λ_i has geometric multiplicity n_i, then we include it n_i times in our list of roots. For $i = 1, 2, \dots, k$, let $V_i \in \mathbb{C}^{d_i m \times d_i}$ be defined by (10) and (11) (with $q = m$). Furthermore, define the $n \times d_i$ matrix S_i by

$$S_i := \begin{pmatrix} X(\lambda_i) & X^{(1)}(\lambda_i) & \dots & X^{(d_i-1)}(\lambda_i) \end{pmatrix} V_i,$$

where $X^{(j)}$ denotes the jth derivative of X. The zero state matrix in the general case is then defined by

$$S_f := \begin{pmatrix} S_1 & S_2 & \dots & S_k \end{pmatrix}. \tag{26}$$

Again, $S_f \in \mathbb{C}^{n \times n}$.

The following theorem is the main result of this paper. It yields the representation (25) of the bijection Ric in the general, not necessarily semisimple, case.

THEOREM 12. *Assume $\partial\Phi(i\omega) \geq 0$ for all $\omega \in \mathbb{R}$. Then the following three statements are equivalent:*

(i) $\partial\Phi(i\omega) > 0$ *for all* $\omega \in \mathbb{R}$*;*
(ii) $\mathcal{F}_{\rm cop} \neq \emptyset$*;*
(iii) $\mathcal{S}_{\rm unm} \neq \emptyset$.

Assume that this holds. Then $\mathrm{Ric} : \mathcal{F}_{\rm cop} \to \mathcal{S}_{\rm unm}$ *is a bijection. For all* $f \in \mathcal{F}_{\rm cop}$ *the zero state matrix* S_f *defined by* (26) *is nonsingular. Furthermore, for any* $f \in \mathcal{F}_{\rm cop}$*, the corresponding solution* $\mathrm{Ric}(f) \in \mathcal{S}_{\rm unm}$ *is given by*

$$\mathrm{Ric}(f) = -(S_f^*)^{-1} T_f S_f^{-1}. \tag{27}$$

Proof. The claim that conditions (i), (ii), and (iii) of Theorem 12 are equivalent, and the claim that under this condition Ric defines a bijection between $\mathcal{F}_{\rm cop}$ and $\mathcal{S}_{\rm unm}$, follow from Theorem 11.

We prove that the zero state matrix (26) is nonsingular. Let $F \in \mathbb{R}^{m\times m}[\xi]$ be such that $\det(\partial\Phi) = f^\sim f$ and $\det(F) = f$. Let $\xi = \mathrm{col}(\xi_1, \xi_2, \dots, \xi_k)$, with $\xi_i = \mathrm{col}(\xi_{i,1}, \xi_{i,2}, \dots, \xi_{i,d_i}) \in \mathbb{C}^{d_i}$, satisfy $S_f\xi = 0$, equivalently,

$$\sum_{i=1}^{k} \begin{pmatrix} X(\lambda_i) & X^{(1)}(\lambda_i) & \dots & X^{(d_i-1)}(\lambda_i) \end{pmatrix} V_i \xi_i = 0.$$

We will show that $\xi_i = 0$ for $i = 1, 2, \dots, k$.

Recall that the system $\frac{d}{dt}x = Ax + Bu$ has an observable image representation

$$\begin{pmatrix} x \\ u \end{pmatrix} = \begin{pmatrix} X(\frac{d}{dt}) \\ U(\frac{d}{dt}) \end{pmatrix} l, \tag{28}$$

and that $X(\frac{d}{dt})$ is a minimal state map for this system. Consider the extended system $\mathcal{B}_{\rm ext}$, obtained by including $f = F(\frac{d}{dt})l$ as a manifest variable, represented by the image representation

$$\begin{pmatrix} x \\ u \\ f \end{pmatrix} = \begin{pmatrix} X(\frac{d}{dt}) \\ U(\frac{d}{dt}) \\ F(\frac{d}{dt}) \end{pmatrix} l.$$

We claim that in the system $\mathcal{B}_{\rm ext}$, $\mathrm{col}(x, u)$ is output and f is input, and that $X(\frac{d}{dt})$ is a minimal state map also for $\mathcal{B}_{\rm ext}$.

To prove this, first note that

$$F^\sim F = X^\sim Q X + X^\sim S^T U + U^\sim S X + U^\sim R U.$$

Multiplying this equality on the right by U^{-1} and on the left by $(U^\sim)^{-1}$ yields

$$(U^\sim)^{-1} F^\sim F U^{-1} = (U^\sim)^{-1} X^\sim Q X U^{-1} + (U^\sim)^{-1} X^\sim S^T + S X U^{-1} + R.$$

Since XU^{-1} is strictly proper and $R > 0$, this implies that FU^{-1} is a proper rational matrix with nonsingular feedthrough term. This implies that also its inverse, UF^{-1}, is proper, and $XF^{-1} = XU^{-1}UF^{-1}$ is strictly proper. Since, therefore,

$$\begin{pmatrix} X \\ U \end{pmatrix} F^{-1}$$

is a proper rational matrix, in the system $\mathcal{B}_{\rm ext}$, $\mathrm{col}(x, u)$ is output and f is input.

Next we prove that $X(\frac{d}{dt})$ is a minimal state map for $\mathcal{B}_{\text{ext}}$. To prove this, we show that the rows of X form a basis for the real linear space $\mathcal{S}_1 = \{r \in \mathbb{R}^{1\times m}[\xi] \mid rF^{-1} \text{ is strictly proper}\}$. Since X induces a minimal state map for our original system (28), the rows of X form a basis for the real linear space $\mathcal{S}_2 = \{r \in \mathbb{R}^{1\times m}[\xi] \mid rU^{-1} \text{ is strictly proper}\}$. Since UF^{-1} and FU^{-1} are proper, rF^{-1} is strictly proper if and only if rU^{-1} is strictly proper. Hence the two linear spaces $\mathcal{S}_1$ and $\mathcal{S}_2$ coincide, so the rows of X indeed form a basis for $\mathcal{S}_1$.

Define a particular latent variable trajectory for $\mathcal{B}_{\text{ext}}$ by

$$\tilde{l}(t) = \sum_{i=1}^{k} e^{\lambda_i t} \begin{pmatrix} I_{m\times m} & tI_{m\times m} & \dots & t^{d_i-1}I_{m\times m} \end{pmatrix} V_i\xi_i.$$

Then we clearly have $\partial\Phi(\frac{d}{dt})\tilde{l} = 0$. Using the fact that none of the λ_i's is a singularity of $F^{\sim}$, this implies that $F(\frac{d}{dt})\tilde{l} = 0$. Our aim is to prove that $\tilde{l} = 0$. Indeed, look at the trajectory of the system $\mathcal{B}_{\text{ext}}$ corresponding to the choice of latent variable $\tilde{l}$. The input $f = F(\frac{d}{dt})\tilde{l}$ is equal to zero. Furthermore, a straightforward calculation shows that the value of the corresponding state trajectory at time $t = 0$ equals

$$\left(X\left(\frac{d}{dt}\right)\tilde{l}\right)(0) = \sum_{i=1}^{k} \begin{pmatrix} X(\lambda_i) & X^{(1)}(\lambda_i) & \dots & X^{(d_i-1)}(\lambda_i) \end{pmatrix} V_i\xi_i = 0.$$

Hence the output $(x,u) = (X(\frac{d}{dt})\tilde{l}, U(\frac{d}{dt})\tilde{l}$ of $\mathcal{B}_{\text{ext}}$ is zero. By observability of the image representation (28), this implies $\tilde{l} = 0$, as claimed.

Next, we prove that this implies $\xi_i = 0$ for all i. Indeed, since $\tilde{l} = 0$ we have

$$\tilde{l}(0) = \sum_{i=1}^{k} \begin{pmatrix} I_{m\times m} & 0 & \dots & 0 \end{pmatrix} V_i\xi_i = 0.$$

Consequently, $\tilde{l}(0) = \sum_{i=1}^{k}\sum_{j=0}^{d_i-1} a_{i,j}\xi_{i,j} = 0$. Since the vectors $a_{i,j}$ are linearly independent, this yields $\xi_{i,j} = 0$ for all $i = 1,2,\dots,k$ and $j = 1,2,\dots,d_i$. This proves that the zero state matrix S_f is nonsingular.

To prove (27) we use that $K_f = \text{Ric}(f)$ is uniquely defined by

$$\Phi(\zeta,\eta) - F^T(\zeta)F(\eta) = -(\zeta+\eta)X^T(\zeta)K_fX(\eta), \tag{29}$$

with $F \in \mathbb{R}^{m\times m}[\xi]$ such that $\partial\Phi = F^{\sim}F$ and $\det(F) = f$. The idea is to evaluate (29) and its partial derivatives with respect to ζ and η at the points $(\bar{\lambda}_i, \lambda_j)$. For all indices (r,s) we have

$$\begin{aligned}&\frac{\partial^{r+s}\Phi}{\partial\eta^r\partial\zeta^s}(\zeta,\eta) - F^{(s)T}(\zeta)F^{(r)}(\eta)\\ &= sX^{(s-1)T}(\zeta)K_fX^{(r)}(\eta) + rX^{(s)T}(\zeta)K_fX^{(r-1)}(\eta)\\ &\qquad\qquad + (\zeta+\eta)X^{(s)T}(\zeta)K_fX^{(r)}(\eta).\end{aligned}$$

Using this, for $i,j = 1,2,\dots,k$, we form the matrices $\Phi_{i,j}$ defined by (13). Next, with $\Sigma_{i,j}$ defined by (14) and $\Lambda_{i,j}$ defined by (12), a straightforward calculation shows that

$$\begin{aligned}&\Lambda_{j,i}^*V_i^*\Sigma_{i,j}V_j\Lambda_{i,j}\\ &= -V_i^*\begin{pmatrix} X^T(\bar{\lambda}_i)\\ X^{(1)T}(\bar{\lambda}_i)\\ \vdots\\ X^{(d_i-1)T}(\bar{\lambda}_i)\end{pmatrix} K_f \begin{pmatrix} X(\lambda_j) & X^{(1)}(\lambda_j) & \dots & X^{(d_j-1)}(\lambda_j)\end{pmatrix} V_j.\end{aligned} \tag{30}$$

The crucial point here is that the terms involving $F^{(r)T}(\bar{\lambda}_i)F^{(s)}(\lambda_j)$ vanish, since for $i = 1, 2, \ldots, k$ we have

$$(31)\qquad \begin{pmatrix} \binom{0}{0}\partial F^{(0)}(\lambda_i) & \binom{1}{0}\partial F^{(1)}(\lambda_i) & \cdots & \cdots & \binom{d_{i-1}}{0}\partial F^{(d_i-1)}(\lambda_i) \\ 0 & \binom{1}{1}\partial F^{(0)}(\lambda_i) & \cdots & \cdots & \binom{d_i-1}{1}\partial F^{(d_i-2)}(\lambda_i) \\ 0 & 0 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \binom{d_i-1}{d_i-1}\partial F^{(0)}(\lambda_i) \end{pmatrix} V_i = 0.$$

The latter follows from (10), combined with the fact that for $i = 1, 2, \ldots, k$ the matrices $F^T(-\lambda_i)$ are nonsingular. Since (30) holds for all $i, j = 1, 2, \ldots, k$, we obtain $T_f = S_f^* K_f S_f$. This completes the proof. □

This result yields a procedure for computing all unmixed solutions of the ARE (1). We sum up the steps that are required here.

1. Compute a right coprime factorization $X(\xi)U(\xi)^{-1}$ of $(\xi I - A)^{-1}B$.
2. Form the corresponding two-variable polynomial matrix Φ given by (19).
3. Check whether $\partial\Phi(i\omega) > 0$ for all $\omega \in \mathbb{R}$.
4. Factor $\det(\partial\Phi) = f^\sim f$ with f and $f^\sim$ coprime.

The following then computes the unique solution $K = K^T$ of the ARE such that its "closed loop characteristic polynomial" χ_{A_K} equals $\sqrt{\det(R)}\ f$.

5. Compute the zero state matrix S_f.
6. Compute the Pick matrix T_f.
7. Solve the equation $T_f = -S_f^* K S_f$.
8. Set $K = \mathrm{Ric}(f)$.

It is worthwhile to observe that similar results have been obtained in Chapter 5 of [11] for QDFs not necessarily associated with a state space representation (16). Note that the procedure circumvents the need to do a polynomial spectral factorization of $\partial\Phi$.

We now go back to the problem of establishing necessary and sufficient conditions for the existence of sign-definite solutions to the ARE. Our main result here is an immediate consequence of Theorem 12 and is based on the result of Proposition 3, namely, that the largest (smallest) storage function for Φ is associated with an anti-Hurwitz (Hurwitz) factorization of $\partial\Phi$. Let K_- and K_+ be the smallest, respectively the largest, real symmetric solution of the ARE.

COROLLARY 13. *Let $\Phi(\zeta, \eta)$ be defined as in* (19). *Assume that $\partial\Phi(i\omega) > 0$ for all $\omega \in \mathbb{R}$. Factor $\det(\partial\Phi) = (f_A)^\sim f_A = (f_H)^\sim f_H$, where f_A and f_H have their roots in the open right half plane and open left half plane, respectively. Then we have*

$$K_- = -(S_{f_A}^*)^{-1} T_{f_A} S_{f_A}^{-1},$$

$$K_+ = -(S_{f_H}^*)^{-1} T_{f_H} S_{f_H}^{-1}.$$

Consequently, $\mathrm{sign}(K_-) = -\mathrm{sign}(T_{f_H})$ *and* $\mathrm{sign}(K_+) = -\mathrm{sign}(T_{f_A})$. *In particular, the ARE* (1) *has a negative semidefinite (negative definite) solution if and only if the Pick matrix T_{f_A} is positive semidefinite (respectively, positive definite). It has a positive semidefinite (positive definite) solution if and only if the Pick matrix T_{f_H} is negative semidefinite (respectively, negative definite).*

Example 6, *continued.* For the Riccati equation of Example 6 we have $\partial\Phi(\xi) = \left(\begin{smallmatrix} 1-\xi^2 & a \\ a & 4-\xi^2 \end{smallmatrix}\right)$, and we have $\partial\Phi(i\omega) > 0$ for all $\omega \in \mathbb{R}$ if and only if $-2 < a < 2$. Assume

this to be the case. We have $\det(\partial\Phi(\xi)) = (1-\xi^2)(4-\xi^2)-a^2$. Set $k = 3+\sqrt{9+4a^2}$. The singularities of $\partial\Phi$ are then equal to $\lambda_1 = -\sqrt{1+\frac{1}{2}k}$, $\lambda_2 = -\sqrt{4-\frac{1}{2}k}$, $-\lambda_1$, and $-\lambda_2$. Clearly, $\det(\partial\Phi)$ can be factored as $f^{\sim}f$ with $(f^{\sim}, f)$ coprime in four different ways, and the Riccati equation has four real symmetric solutions, all of them unmixed. Here we compute the largest real symmetric solution, i.e., the solution K satisfying $\chi_{A_K} = f_H$, with $f_H(\xi) = (\xi + \sqrt{1+\frac{1}{2}k})(\xi + \sqrt{4-\frac{1}{2}k})$. Note that we are in the semisimple situation, i.e., the algebraic multiplicity of each singularity equals its corresponding rank deficiency. Solving $\partial\Phi(\lambda_1)v_1 = 0$ and $\partial\Phi(\lambda_2)v_2 = 0$ yields $v_1 = \left(\begin{smallmatrix} 2a/k \\ 1 \end{smallmatrix}\right)$ and $v_2 = \left(\begin{smallmatrix} 1 \\ -2a/k \end{smallmatrix}\right)$. The zero state matrix S_{f_H} is hence given by $S_{f_H} = \left(\begin{smallmatrix} 2a/k & 1 \\ 1 & -2a/k \end{smallmatrix}\right)$. Next we compute the Pick matrix corresponding to f_H. Clearly,

$$T_{f_H} = \begin{pmatrix} \frac{v_1^*\Phi(\lambda_1,\lambda_1)v_1}{2\lambda_1} & \frac{v_1^*\Phi(\lambda_1,\lambda_2)v_2}{\lambda_1+\lambda_2} \\ \frac{v_2^*\Phi(\lambda_2,\lambda_1)v_1}{\lambda_2+\lambda_1} & \frac{v_2^*\Phi(\lambda_2,\lambda_2)v_2}{2\lambda_2} \end{pmatrix},$$

which is equal to

(32)
$$T_{f_H} = \begin{pmatrix} \frac{1}{2\lambda_1}\left(\frac{4a^2}{k^2}(2+k) + \frac{2a^2}{k} + \frac{k}{2} + 5 - 2\lambda_1\right) & \frac{1}{\lambda_1+\lambda_2}\left(\frac{2a}{k}(1+\lambda_1\lambda_2) + a - \frac{4a^3}{k^2} - \frac{8a}{k} + \frac{2a}{k}(\lambda_1+\lambda_2)\right) \\ \frac{1}{\lambda_1+\lambda_2}\left(\frac{2a}{k}(1+\lambda_1\lambda_2) + a - \frac{4a^3}{k^2} - \frac{8a}{k} + \frac{2a}{k}(\lambda_1+\lambda_2)\right) & \frac{1}{2\lambda_2}\left(\frac{4a^2}{k^2}(8+2\lambda_2) - \frac{6a^2}{k} - \frac{k}{2} + 5\right) \end{pmatrix}.$$

This yields $K^+ = -(S_{f_H}^T)^{-1}T_{f_H}S_{f_H}$ as the solution corresponding to f_H. Note that this gives the largest real symmetric solution for each value of a between -2 and 2. For example, if $a = 0$, then $k = 6$, so $\lambda_1 = -2$ and $\lambda_2 = -1$. This yields $K^+ = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)\left(\begin{smallmatrix} 3 & 0 \\ 0 & 1 \end{smallmatrix}\right)\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right) = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 3 \end{smallmatrix}\right)$. Recall that $Q = \left(\begin{smallmatrix} 1 & a \\ a & 3 \end{smallmatrix}\right)$, so for $a = 0$ we have $Q > 0$. In this particular case it follows immediately that the ARE has a positive semidefinite solution (the corresponding linear quadratic problem is positive semidefinite). For values of a satisfying $-2 < a < -\sqrt{3}$ or $\sqrt{3} < a < 2$, Q is indefinite, so for this case it is a nontrivial matter to check whether the ARE has a positive (semi-) definite solution. According to Corollary 13, for a given $a \in (-2, 2)$ the ARE has a positive (semi-) definite solution if and only if for that value of a the Pick matrix (32) is negative (semi-) definite. As an example, take $a = 1.8$. In this case Q is indefinite. The Pick matrix corresponding to this value of a is computed as $T_{f_H} = \left(\begin{smallmatrix} -3.6835 & 0.3111 \\ 0.3111 & -0.2637 \end{smallmatrix}\right)$. The eigenvalues of T_{f_H} are computed as -3.7116 and -0.2356, so we conclude that for $a = 1.8$ our ARE has a positive definite solution. For $a = 1.98$ we compute $T_{f_H} = \left(\begin{smallmatrix} -3.7839 & 0.4375 \\ 0.4375 & 0.0894 \end{smallmatrix}\right)$, which has eigenvalues -3.8327 and 0.1382. For this value of a our ARE does not have a positive semidefinite solution.

In order to check whether for a given a the ARE of this example has at least one negative (semi-) definite solution, one should compute the Pick matrix T_{f_A} associated with the polynomial $f_A(\xi) = (\xi - \sqrt{1+\frac{1}{2}k})(\xi - \sqrt{4-\frac{1}{2}k})$, and check whether it is positive (semi-) definite.

7. Conclusions. In this paper we applied ideas from the calculus of two-variable polynomial matrices to the problem of characterizing all umixed solutions of the algebraic Riccati equation and formulating necessary and sufficient conditions for the existence of (semi) definite solutions.

We started from the two-variable polynomial matrix corresponding to the underlying quadratic functional, and associated with this a nonsingular one-variable

polynomial matrix. Then we showed that there is a bijection between the set of all scalar polynomial spectral factors of the determinant of this one-variable polynomial matrix and the set of all unmixed solutions of the ARE. For every such scalar polynomial spectral factor we defined a constant Hermitian matrix, called the Pick matrix, and we expressed the unmixed solution corresponding to this polynomial spectral factor in terms of its Pick matrix. This enabled us to conclude that the signatures of the extremal solutions of the ARE are determined by the Pick matrices corresponding to these solutions.

In this paper, we have restricted ourselves to the case in which (A, B) is a controllable pair, mainly in order to be able to use image representations. As a possible direction for future research, we mention the extension of our results to the noncontrollable case. Another interesting problem would be to generalize our results to the discrete-time algebraic Riccati equation.

REFERENCES

[1] B. D. O. Anderson, *Corrections to: Algebraic properties of minimal degree spectral factor*, Automatica J. IFAC, 11 (1975), pp. 321–322.

[2] S. Bittanti, A. J. Laub, and J. C. Willems, eds., *The Riccati Equation*, Springer-Verlag, Berlin, 1991.

[3] W. A. Coppel, *Linear Systems*, Notes on Pure Mathematics 6, Australian National University, Canberra, Australia, 1972.

[4] J. B. Garnett, *Bounded Analytic Functions*, Academic Press, New York, 1981.

[5] A. H. W. Geerts, *A necessary and sufficient condition for solvability of the linear-quadratic control problem without stability*, Systems Control Lett., 11 (1988), pp. 47–51.

[6] A. H. W. Geerts and M. L. J. Hautus, *The output stabilizable subspace and linear optimal control*, in Robust Control of Linear Systems and Nonlinear Control, Progr. Systems Control Theory 4, Birkhäuser Boston, Cambridge, MA, 1990, pp. 113–120.

[7] B. P. Molinari, *Conditions for non-positive solutions of the linear matrix inequality*, IEEE Trans. Automat. Control, AC-29 (1975), pp. 804–806.

[8] P. J. Moylan, *On a frequency domain condition in linear optimal control theory*, IEEE Trans. Automat. Control, AC-29 (1975), p. 806.

[9] J. W. Polderman and J. C. Willems, *Introduction to Mathematical System Theory: A Behavioral Approach*, Springer-Verlag, Berlin, 1997.

[10] A. C. M. Ran and L. Rodman, *Factorization of matrix polynomials with symmetries*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 845–864.

[11] P. Rapisarda, *Linear Differential Systems*, Ph.D. thesis, University of Groningen, Groningen, The Netherlands, 1998.

[12] P. Rapisarda and J. C. Willems, *State maps for linear systems*, SIAM J. Control Optim., 35 (1997), pp. 1053–1091.

[13] H. L. Trentelman, *When does the algebraic Riccati equation have a negative semi-definite solution?*, in Open Problems in Mathematical Systems and Control Theory, V. D. Blondel, E. D. Sontag, M. Vidyasagar, and J. C. Willems, eds., Springer-Verlag, New York, 1998, pp. 229–237.

[14] H. L. Trentelman and R. Van der Geest, *The Kalman–Yakubovic–Popov lemma in a behavioural framework*, Systems Control Lett., 32 (1997), pp. 283–290.

[15] H. L. Trentelman and J. C. Willems, *Every storage function is a state function*, Systems Control Lett., 32 (1997), pp. 249–260.

[16] H. L. Trentelman and J. C. Willems, *H_∞-control in a behavioral context: The full information case*, IEEE Trans. Automat. Control, 44 (1999), pp. 521–536.

[17] H. L. Trentelman, A. A. Stoorvogel, and M. L. J. Hautus, *Control Theory for Linear Systems*, Springer-Verlag, Berlin, London, Heidelberg, 2001.

[18] J. C. Willems and H. L. Trentelman, *On quadratic differential forms*, SIAM J. Control Optim., 36 (1998), pp. 1703–1749.

[19] J. C. Willems, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, AC-16 (1971), pp. 621–634.

[20] J. C. Willems, *On the existence of a nonpositive solution to the Riccati equation*, IEEE Trans. Automat. Control, AC-19 (1974), pp. 592–593.

[21] J. C. Willems, *Paradigms and puzzles in the theory of dynamical systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 259–294.
[22] J. C. Willems and H. L. Trentelman, *Synthesis of dissipative systems using quadratic differential forms, Parts* I *and* II, IEEE Trans. Automat. Control, to appear.
[23] H. K. Wimmer, *Decomposition and parametrization of semidefinite solutions of the continuous-time algebraic Riccati equation*, SIAM J. Control Optim., 32 (1994), pp. 995–1007.
[24] H. K. Wimmer, *Lattice properties of sets of semi-definite solutions of continuous-time algebraic Riccati equations*, Automatica J. IFAC, 31 (1995), pp. 173–182.
[25] D. C. Youla and M. Saito, *Interpolation with positive real-functions*, J. Franklin Inst., 284 (1967), pp. 77–108.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 4, pp. 993–1010

OUTPUT TRACKING THROUGH SINGULARITIES*

R. M. HIRSCHORN†

Abstract. Output tracking for nonlinear systems is complicated by the existence of "singular submanifolds." These are surfaces on which the decoupling matrix loses rank. To provide additional control action we identify a class of smooth vector fields whose integral curves can be *incrementally tracked* using rapidly switched piecewise constant controls. At discrete times the resulting piecewise smooth state trajectories approach the integral curve being tracked. These discontinuous controllers are applied to sliding mode control—we use incremental tracking to move the state toward the sliding surface. The resulting controller achieves approximate output tracking in situations where the usual approach to sliding mode control fails due to the loss of control action on the singular submanifold.

Key words. output tracking, sliding mode control, singularities, Lie brackets, discontinuous state feedback

AMS subject classification. 93C10

PII. S0363012999354879

1. Introduction.

Tracking in the case where the decoupling matrix loses rank on a "singular submanifold" has been considered by a number of authors (cf. [2, 5, 6, 7, 9, 15]). In [2] the problem of exact tracking is studied using results on singular ordinary differential equations and on the multiplicity of solutions. Conditions under which the singular tracking control is smooth or analytic are given in [9], assuming that the inputs and some of their derivatives are related to the outputs and their derivatives via a singular ordinary differential equation. In [7], output trajectories which the system can track using continuous open loop controls are identified for systems which satisfy a suitable observability condition, and a discontinuous feedback controller is introduced which achieves robust tracking in the face of perturbations. In [5] the relative order is locally increased by keeping the state trajectory near a codimension 1 submanifold. In some sense our approach takes the opposite point of view in that we seek to reduce the relative order by using vibratory controls. These switched controls allow motion in directions other than those of the drift vector field or vector fields in the Lie algebra generated by the control vector fields.

Recently there has been increased interest in the use of patterns in control. The pioneering work of Brockett [1], Pomet [12], Lui and Sussmann [10], and others looks at curves that can be approached by state trajectories of smooth affine systems. For single-input systems these results highlight the very limited class of smooth paths which can be closely approximated by the state trajectory. We introduce the notion of *incremental tracking* of smooth integral curves by state trajectories. The state trajectories are permitted to move far from the integral curve being tracked but are required to approach them arbitrarily closely arbitrarily often. This weaker notion of approximation by the state trajectory lends itself well to sliding mode control where we wish to steer the state to a sliding surface. This is a surface on which the state evolves so that the tracking errors go to zero. We are not concerned about the

*Received by the editors April 23, 1999; accepted for publication (in revised form) April 18, 2001; published November 28, 2001. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada.
http://www.siam.org/journals/sicon/40-4/35487.html
†Department of Mathematics and Statistics, Queen's University, Kingston, ON K7L 3N6, Canada (ron@mast.QueensU.CA).

path along which the trajectory approaches the sliding surface, as long as any large deviations take place in directions which are not seen directly by the output.

Sliding mode control utilizing discontinuous feedback controllers can achieve robust asymptotic output tracking (cf. [16, 13, 14] and the references therein) under the implicit assumption that the state trajectory can always be steered toward the "sliding surface." That is, the *decoupling matrix* is of full rank everywhere (cf. [8, 11]). In [6], sliding mode control is studied in the case where the decoupling matrix loses rank, and there exists a "singular submanifold" near which the state trajectory cannot be steered toward the sliding surface. For systems whose "singular submanifold" satisfies suitable transversality conditions, a class of smooth output functions y_d is identified which can be approximately tracked using a truncated sliding mode controller. For these outputs the state trajectory passes through the "singular submanifold" a finite number of times. There are, however, many simple systems in which truncated controllers cause the state trajectory to "stick" to the "singular submanifold," so that the state moves ever farther from the sliding surface. For such systems the standard approaches to output tracking are also not very successful. The following example illustrates the difficulties which can arise.

Example 1.1. Consider the affine nonlinear system in $\mathbb{R}^3$

$$\begin{aligned} \dot{x}_1 &= x_3^2 - x_2, \\ \dot{x}_2 &= x_3, \\ \dot{x}_3 &= u. \end{aligned} \tag{1.1}$$

Suppose that we wish to regulate the output $y = x_1$ so that $y(t)$ stays close to $y_d(t)$ while keeping the state vector bounded. If $s = \dot{e} + e$, where $e = y - y_d$, then we can regulate y by keeping the state trajectory on or near to the "sliding surface" $S_t^p = \{s = 0\} = \{x_1 + x_3^2 - x_2 = y_d + \dot{y}_d\}$. We note that without the term x_3^2 this system is linear with relative order 3, but here $\ddot{y} = -x_3 + 2x_3 u$ and the relative order of y is 2 (cf. [6, 8]). In particular,

$$\dot{s} = a(x) + b(x)u,$$

where $a(x,t) = x_3^2 - x_2 - x_3 - \dot{y}_d - \ddot{y}_d$ and $b(x) = 2x_3$. The natural sliding mode controller $u_{sm} = -(a + K)\ \mathrm{sign}(s)/b$ achieves $\dot{s} = -K\ \mathrm{sign}(s)$, whence $x(t)$ reaches S_t^p and stays in S_t^p after a finite time has elapsed (cf. [16], [13]). Inherent in this control scheme is the assumption that b does not vanish along the state trajectory. Of course, in our case b vanishes on the "singular manifold" $N = \{x_3 = 0\}$, and hence u_{sm} can become unbounded as $x(t)$ approaches N. One natural solution is to use the truncated controller $\min\{u_{sm}, L\ \mathrm{sign}(u_{sm})\}$ or the simpler controller

$$u_{sm}^L = -L\ \mathrm{sign}(sb). \tag{1.2}$$

For linear systems, such truncated controllers work on a neighborhood of the origin which expands as L grows. This is not the case here. In fact, suppose that we wish to track $y_d = 0$, where x_1 is positive, x_2 negative, and $x_3 = 0$ ($x \in N$, $s(x)$ is positive). If we perturb x_3 so that $x_3 = \epsilon > 0$, we have $u_{sm}^L < 0$; hence $\dot{x}_3 < 0$ and x returns to N. For $x_3 = -\epsilon$ we have $\dot{x}_3 = u_{sm}^L > 0$, and once again x returns to N. In essence the state trajectory will "stick" to the submanifold $N = \{x_3 = 0\}$. Of course, on N we have $\dot{s} = -x_2$, $\dot{x}_2 = 0$, so that $\dot{s} = -x_2 > 0$ and the state trajectory evolves on N in such a way that

$$s(e(t)) \longrightarrow \infty.$$

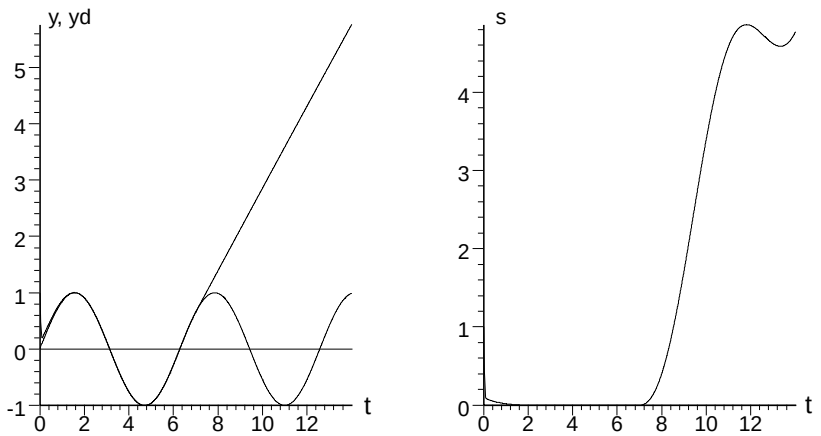


FIG. 1. *Tracking of* $\sin t$ *using a truncated sliding mode controller.*

We can track $y_d = 0$ using this approach if the initial state $x_2(0) > 0$. The larger $x_2(0)$ is, the more we can insulate the system from the above phenomena. On the other hand, if we track $y_d(t) = \sin t$, even with $x_2(0) > 0$, we will inevitably find that x_2 becomes negative and the above problem dominates. This phenomenon is illustrated by Figure 1, which shows the results of a simulation performed using SIMNON/PCW for Windows, Version 2.01 (SSPA Maritime Consulting AB, Sweden). If $x_2(0) < 0$, then the divergence of s and e is immediate. With $x(0) = (1, 11, 0)$ ($x_2(0) = 11$) and controller (1.2) with $L = 30$, the onset of this divergence is only delayed.

It is of interest to note that if we could enforce $s \equiv 0$ exactly in the case $y_d \equiv 0$, then $x_1, \dot{x}_1 \to 0$, $x_2 \to x_3^2 \geq 0$. Thus if $x_3 > 0$, then $\dot{x}_2 = x_3 = \sqrt{x_2}$, and the resulting "zero dynamics" are unstable.

The approximate input-output linearization scheme of [5] applied to this example has similar problems. Tracking schemes which are based on differentiating y until u appears come up against this same obstruction. Tomlin and Sastry have observed a similar phenomenon in the ball and beam example [15], where their switched control scheme is not effective. The above example presents similar obstructions.

Instead of taking more derivatives of s to deal with the singular submanifold N, we use fewer derivatives. As a result we lose direct control over s (as $\dot{s}$ is independent of u) but avoid the problems associated with the "singular manifold." We introduce a switched periodic controller which causes the state to "incrementally track" the integral curve of a vector field obtained from Lie brackets of the drift and control vector fields. The resulting continuous but nonsmooth state trajectory approaches the sliding surface. We will return to this example in section 4.

The rest of the paper is organized as follows. In section 2 we formulate the sliding mode control problem for single-input single-output affine nonlinear systems. In section 3 we introduce our switched controllers and present our results on approximate trajectory tracking for systems with drift. In section 4 we state and prove our main results—applications of incremental tracking to sliding mode control—and continue the above example. Finally, some concluding remarks are offered in section 5.

2. Output tracking and sliding surfaces. Suppose that M is a smooth manifold. Given a smooth function $h : M \to \mathbb{R}$ and a vector field $X(x)$ on M, $Xh(x) = dh_x X(x)$ denotes the *Lie derivative* of $h(x)$ along $X(x)$, and $X_t(x_0)$ the *integral curve* of X passing through x_0 at $t = 0$, so that $\frac{d}{dt}X_t(x_0) = X(X_t(x_0))$. If Y is a smooth vector field on M, then $[X,Y](x) = dY_x X(x) - dX_x Y(x)$ denotes the *Lie bracket* of X and Y, and $ad_X Y = [X,Y]$. Let $\{X,Y\}_{LA}$ denote the *Lie algebra* generated by $\{X,Y\}$, i.e., the smallest vector space containing X and Y and closed under Lie brackets. Suppose that N is a codimension 1 submanifold of M. A vector field X is *transversal to* N if $X(x) \notin T_x N \ \forall x \in N$, where $T_x N$ is the tangent space to N at x. If $P \subset Q$ is a submanifold and $f : M \to Q$ a smooth map of manifolds, then f *is transversal to* P if $\text{Image}(df_x) + T_{f(x)}P = T_{f(x)}Q$.

Consider the nonlinear control system model

$$\begin{aligned} \dot{x} &= f(x) + g(x)u, \qquad x(t_0) = x_0 \in M, \\ y &= h(x), \end{aligned} \tag{2.1}$$

where $M \subset \mathbb{R}^\ell$ is a smooth m-dimensional embedded submanifold of $\mathbb{R}^\ell$, $u : [t_0, \infty) \to \mathbb{R}$ is a piecewise smooth input, $f(x)$ and $g(x)$ are smooth vector fields on M, and h is a smooth output function on M. If $x \in M$, we denote by $||x||$ the norm on M which is induced by the standard norm on $\mathbb{R}^\ell$.

Suppose that $y_d : [t_0, \infty) \to \mathbb{R}$ is a smooth function which we wish the output y of (2.1) to track. The standard approach in sliding mode control (cf. [13, 16]) is to force the evolution of the *output tracking error* $e = y - y_d$ to be governed by a stable differential equation of the form $s(e^{\mathbf{p}}(t)) = 0$, where $e^{\mathbf{p}}(t) = (e(t), e^{(1)}(t), \ldots, e^{(p-1)}(t))$ and $s : \mathbb{R}^p \to \mathbb{R}$ is linear, so that

$$s(e^{\mathbf{p}}(t)) = e^{(p-1)}(t) + a_2 e^{(p-2)}(t) + \cdots + a_p e(t). \tag{2.2}$$

Definition 2.1. *The output of* (2.1) *can* approximately track y_d to degree p *if, given any* $\delta > 0$*, there exists an admissible input* u_δ *and time* $t_\delta > t_0$ *such that* $|s(e^{\mathbf{p}}(t))| \leq \delta$ *and the resulting state* $x(t)$ *is bounded on* $[t_\delta, \infty)$*. We say that* y asymptotically tracks y_d to degree p *if* $s(e^{\mathbf{p}}(t)) = 0$ *and* $x(t)$ *is bounded on* $[t_0, \infty)$*.*

The *relative degree* r of the output y is the least positive integer for which the derivative $y^{(r)}(t)$ is an explicit function of the input u. More precisely, r is the least positive integer for which $gf^{(r-1)}h \not\equiv 0$ (cf. [7, 8]). For single-input systems, the "decoupling matrix" is the 1×1 matrix whose entry is $gf^{(r-1)}h$. Thus the rank of the decoupling matrix changes where $gf^{(r-1)}h$ vanishes. We choose $p \leq r$ to avoid a possibly singular differential equation for u. Thus $y = h(x), y^{(1)} = fh(x), \ldots, y^{(p-1)} = f^{p-1}h(x)$. If we set $h^{\mathbf{p}} = (h, fh, \ldots, f^{p-1}h)$, then $s(e^{\mathbf{p}}(t)) = 0$ is equivalent to the requirement that $s^p(x(t), t) = 0$, where

$$\begin{aligned} s^p(x,t) &= s(h^{\mathbf{p}}(x) - y_d^{\mathbf{p}}(t)) \\ &= s(h^{\mathbf{p}}(x)) - s(y_d^{\mathbf{p}}(t)). \end{aligned} \tag{2.3}$$

In particular, if we let S_t^p denote the *sliding surface*

$$S_t^p = \{x | s^p(x,t) = 0\}, \tag{2.4}$$

then $x(t) \in S_t^p \ \forall t \geq t_f$ implies asymptotic tracking. Similarly, if

$$E_t^p = \{x | h^{\mathbf{p}}(x) = y_d^{\mathbf{p}}(t)\}, \tag{2.5}$$

then $E_t^p \subset S_t^p$, and $x(t) \in E_t^p \ \forall t \geq t_f$ implies $y^{\mathbf{p}} \equiv y_d^{\mathbf{p}}$ and perfect tracking. Our first assumption is that S_t^p is submanifold.

A1. S_t^p is an embedded codimension 1 submanifold of M for all $t \in [t_0, +\infty)$.

Remark 2.2. It is straightforward to show that A1 holds if the map $h^{\mathbf{p}}$ is transversal to the hyperplane $s^{-1}(0) + y_d^{\mathbf{p}}(t)$ (see [4, 6]).

The standard sliding mode controller approach (cf. [6, 13, 16]) is to pick $p = r$, the relative order of the output y. Then u appears explicitly in $\frac{d}{dt}s^r(x(t),t) = a(x(t),t) + b(x(t))u(t)$, where $a(x,t) = f^r h(x) - y_d^{(r)}(t) + \sum_{i=0}^{r-2} a_i(f^{i+1}h(x) - y_d^{(i+1)}(t))$ and $b(x) = gf^{r-1}h(x)$. The standard sliding mode controller takes the form $u_{sm}(x,t) = -(a(x,t) + K \text{ sign}(s^r(x,t)))/b(x)$, where $K > 0$. Using this control, $\frac{d}{dt}s^r(x(t),t) = -K \text{ sign}(s^r(x(t),t))$ and hence, after some finite time $t_f \geq t_0$, we will have $s^r(x(t),t) = 0 \, \forall t \geq t_f$. If, in addition, the system has bounded "zero dynamics" on E_t^p, then asymptotic tracking of an output y_d will be achieved (see [8]). We note that systems which fail to be strongly observable in the sense of [7] can have unstable zero dynamics (cf. [6, 15]). Of course, the assumption that b does not vanish along the state trajectory is strong. It holds in the linear case but is rarer in the nonlinear case. Typically b vanishes on the *singular submanifold* $N = \{gf^{r-1}h(x) = 0\}$, and u_{sm} becomes unbounded when the state trajectory reaches N. A natural solution is to use a truncated controller, but the resulting state trajectory can "stick" to N and evolve in such a way that one travels away from S_t^p on N (such is the case in Example 1.1). We now introduce switched controllers, which permit us to move toward the sliding surface even if $b(x)$ vanishes.

3. Incremental tracking. The set of curves which can be approximately tracked by the state trajectories of affine systems has been characterized in [12]. For single-input systems the state trajectory can only be made to stay close to integral curves of vector fields of the form $f + \alpha g$, where α is a smooth function on M. Thus to make the state approach the sliding surface S_t^r (where r is the relative degree of y) we are limited to the standard sliding mode controller and the problems associated with singular submanifolds. We seek instead to identify vector fields whose integral curves can be approached arbitrarily closely at discrete times by the state trajectory (see Figure 2). If the deviations from the integral curve are "parallel" to S_t^p for some $p \leq r$, we can use these state trajectories to implement sliding mode controllers for which singular manifolds do not pose a problem.

DEFINITION 3.1. *The integral curves of a smooth vector field X are said to be* incrementally tracked *by the state of* (2.1) *if there exist controllers $\{u_n\}$ with the following properties:*

(a) *each $u_n(x,t)$ is smooth with respect to x and is piecewise constant and periodic with respect to t with period $\tau_n = \frac{\beta_n}{n}$, where $0 < \beta_n \leq 1$;*

(b) *if $\alpha(t)$ is an integral curve of X on $[0,1]$, $x_n(t)$ the state trajectory when $u = u_n$, $x_n(0) = \alpha(0)$, and $\epsilon > 0$, then, for n sufficiently large,*

$$|| \, \alpha(k/n) - x_n(\beta_n k/n) \, || < \epsilon$$

for $k = 0, 1, \ldots, n$.

While not essential, we will assume that vector fields are complete. Let $\mathcal{I}$ denote the set of vector fields on M whose integral curves can be incrementally tracked by the state of system (2.1), and $\mathcal{I}_0$ the subset of $\mathcal{I}$ consisting of vector fields X with $\gamma X \in \mathcal{I}$ for all smooth functions $\gamma : M \to \mathbb{R}$.

THEOREM 3.2. *The set of vector fields $\mathcal{I}$ and $\mathcal{I}_0$ whose integral curves can be incrementally tracked by the state of* (2.1) *have the following properties:*

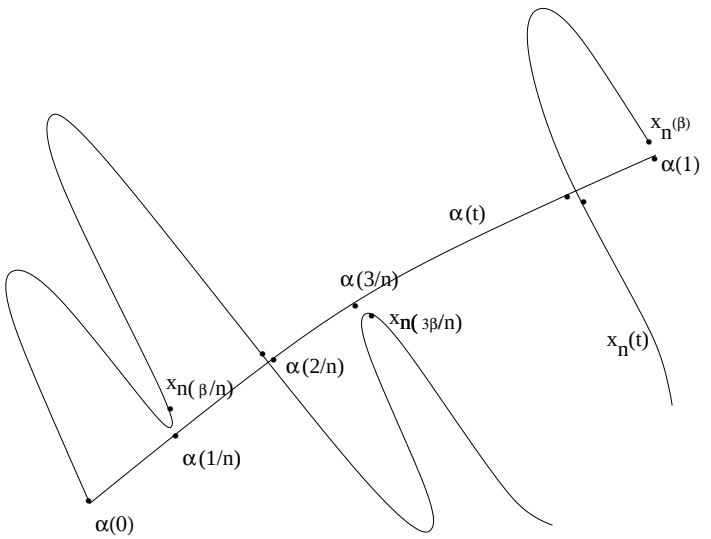


FIG. 2. *Incremental tracking of* $\alpha(t)$.

(i) $f \in \mathcal{I}$, $g \in \mathcal{I}_0$.
(ii) $\mathcal{I}_0$ *is a Lie algebra over* $\mathbb{R}$. *If* $X \in \mathcal{I}$ *and* $Y \in \mathcal{I}_0$, *then* $X + Y \in \mathcal{I}$.
(iii) *Suppose that* $Y \in \mathcal{I}$ *and* $X, ad_X^{k+1} Y \in \mathcal{I}_0$. *Then*
  (a) *if* $[ad_X^i Y, ad_X^j Y] = 0$ *for* $j \le k, i \ge 3k - j$, *then* $ad_X^k Y \in \mathcal{I}$ ($\mathcal{I}_0$ *if* k *is odd*);
  (b) *if* $ad_X^{2k} Y = 0$, *then* $ad_X^k Y \in \mathcal{I}$ ($\mathcal{I}_0$ *if* k *is odd*).
(iv) *If* $ad_g^{k+1} f = 0$, *then* $ad_g^k f \in \mathcal{I}$ ($\mathcal{I}_0$ *if* k *is odd*) *and* $ad_g^k f$ *can be incrementally tracked by the state of* (2.1) *using the periodic switched controllers* $\{u_n\}$ *defined by*

$$u_n(t - t_0) = \begin{cases} -n^{(2k+1)/k}, & 0 \le t - t_0 < 1/n^2, \\ 0, & 1/n^2 \le t - t_0 < 2/n^2, \\ n^{(2k+1)/k}, & 2/n^2 \le t - t_0 < 3/n^2, \end{cases}$$

and $u_n(t + 3/n^2) = u_n(t)$.

Remark 3.3. For the linear system model $\dot{x} = Ax + bu$, $x \in \mathbb{R}^n$, we have $f(x) = Ax \in \mathcal{I}$, $g(x) = b \in \mathcal{I}_0$, $ad_g f = Ab$, $ad_g^2 f = 0$. Thus (iii)(b) or (iv) of Theorem 3.2 implies that $ad_g f = Ab \in \mathcal{I}_0$. Repeating these steps with $ad_g f = Ab$ in place of $g = b$, etc., we find that b, $Ab, \ldots, A^{n-1} b \in \mathcal{I}_0$, and hence these constant vector fields can be incrementally tracked by the state. From this, one can deduce the standard linear result on controllability. We also note that (ii) above implies that incremental tracking of the drift vector field is preserved under smooth static state feedback. We also point out the fact that condition (iv) is nongeneric and will hold only for certain special systems.

We are interested in incremental tracking where large deviations of the state trajectory from the integral curve have only a small effect on the output of the system. We now make this notion more precise.

DEFINITION 3.4. *Suppose that* $\epsilon > 0$ *and* X *is a vector field on* M *whose integral curves can be incrementally tracked by the state* $\{x_n(t)\}$ *of* (2.1) *using controllers*

$\{u_n\}$. *If, for n sufficiently large,*

$$|| h^{\mathbf{p}}(\alpha(t)) - h^{\mathbf{p}}(x_n(\beta_n t)) || < \epsilon \ \ \forall t \in [0,1],$$

we say that the integral curves of X can be incrementally tracked preserving $h^{\mathbf{p}}$.

Let $\mathcal{I}^p$ denote the set of vector fields on M whose integral curves can be incrementally tracked preserving $h^{\mathbf{p}}$, and $\mathcal{I}_0^p$ the subset of $\mathcal{I}^p$ consisting of vector fields X with $\gamma X \in \mathcal{I}^p$ if $\gamma : M \to \mathbb{R}$ is smooth. We assume that $p \le r$.

THEOREM 3.5. *The set of vector fields $\mathcal{I}^p$ and $\mathcal{I}_0^p$ have the following properties:*

(i) $f \in \mathcal{I}^p$, $g \in \mathcal{I}_0^p$.

(ii) $\mathcal{I}_0^p$ *is a Lie algebra over* $\mathbb{R}$. *If* $X \in \mathcal{I}^p$ *and* $Y \in \mathcal{I}_0^p$, *then* $X + Y \in \mathcal{I}^p$.

(iii) *Suppose that* $Y \in \mathcal{I}^p$, $X, ad_X^{k+1} Y \in \mathcal{I}_0^p$, *and* $X h^{\mathbf{p}} = ad_X^{k+1} Y h^{\mathbf{p}} = 0$. *Then*

  (a) *If* $[ad_X^i Y, ad_X^j Y] = 0$ *for* $j \le k$, $i \ge 3k - j$, *then* $ad_X^k Y \in \mathcal{I}^p$ ($\mathcal{I}_0^p$ *if k is odd*).

  (b) *If* $ad_X^{2k} Y = 0$, *then* $ad_X^k Y \in \mathcal{I}^p$ ($\mathcal{I}_0^p$ *if k is odd*).

(iv) *If* $ad_g^{k+1} f = 0$ *and the output of system* (2.1) *has relative order* $r > p$, *then* $ad_g^k f \in \mathcal{I}^p$ ($\mathcal{I}_0^p$ *if k is odd*).

Example 3.6 (*Example* 1.1 *continued*). Here we have $f(x) = (x_3^2 - x_2, x_3, 0)$, $g(x) = (0,0,1)$, $ad_g^2 f(x) = (2,0,0)$, $ad_g^3 f(x) = (0,0,0)$, and $h(x) = x_1$, with $p = 1$ and relative order $r = 2$. Thus condition (iv) of Theorem 3.5 holds and $ad_g^2 f \in \mathcal{I}^p$.

Proof (Theorem 3.2*).*

(i) An integral curve $\alpha(t) = f_t(x)$ of f can be tracked exactly using $u_n = 0$, $\beta_n = 1$. In this case the corresponding state trajectory $x_n(t) = f_t(x) = \alpha(t)$; hence $f \in \mathcal{I}$. Now let $\gamma : M \to \mathbb{R}$ be smooth, and set $t_k = k/n$, $\alpha(t) = \gamma(x) g_t(x)$, $u_n(x,t) = \gamma(x) n$, and $\beta_n = 1/n$. Then

$$x_n(\beta_n t_k) = x_n\left(\frac{k}{n^2}\right) = (f + \gamma n g)_{\frac{k}{n^2}}(x) = \left(\frac{1}{n} f + \gamma g\right)_{\frac{k}{n}}(x),$$

which approximates $\alpha(t_k)$. In particular, we can guarantee that $|| \alpha(t_k) - x_n(\beta_n t_k) || < \epsilon$ for $k = 0, 1, \ldots, n$ and n sufficiently large. This means that $\gamma g \in \mathcal{I}$, hence $g \in \mathcal{I}_0$. Note that in both of the above cases $x_n(t)$ stays close to $\alpha(t)$ $\forall t \in [0,1]$.

(ii) Suppose that $X, Y \in \mathcal{I}_0$, $\alpha(t)$ is an integral curve for $X + Y$ on $[0,1]$, and $\epsilon > 0$. Then $2X, 2Y \in \mathcal{I}_0$, and if $\ell > 0$, we define the "switched integral curve" $\gamma(t) = 2Y_t(x)$ for $0 \le t < 1/2\ell$, and $\gamma(t) = 2X_t(2Y_{1/2\ell}(x))$ for $1/2\ell \le t < 1/\ell$. It follows that

$$\gamma(1/\ell) = (2X)_{1/2\ell}((2Y)_{1/2\ell}(x)) = X_{1/\ell}(Y_{1/\ell}(x)) = Z(\ell)_{1/\ell}(x) = (Z(\ell)/\ell)_1(x),$$

where $Z(\ell) = (X + Y) + \frac{1}{2\ell}[X, Y] + \cdots$ (cf. [17, 18]). Continuing to switch between integral curves of X and Y we get

$$\gamma(2/\ell) = X_{1/\ell}(Y_{1/\ell}(X_{1/\ell}(Y_{1/\ell}(x)))) = (Z(\ell)/\ell)_1^2(x), \cdots, \gamma(k/\ell) = (Z(\ell)/\ell)_1^k(x).$$

Here $Z(\ell) \to X + Y$ as $\ell \to \infty$; hence $(Z(\ell)/\ell)_1^\ell(x) \to (X + Y)_1(x)$ as $k \to \ell$ and $\ell \to \infty$ (cf. [17]). In particular, for ℓ sufficiently large, $|| \alpha(t_k) - \gamma(t_k) || < \epsilon/2$, where $t_k = k/\ell$ and $k = 0, 1, \ldots, \ell$. Since $2Y \in \mathcal{I}_0$ we know that given $\epsilon' > 0$ there exist piecewise constant periodic w.r.t. t controllers $\{u_n\}$, with period $\tau_n = \beta_n/n$ ($0 < \beta_n \le 1$), such that the integral curves of $2Y$ are incrementally tracked by the corresponding state trajectory $x_n(t)$. Thus we have $|| 2Y_{k/n}(\alpha(0)) - x_n(\beta_n k/n) || < \epsilon'/2$ for $k = 0, 1, \ldots, n$ and n sufficiently large. In particular, if $p = 2Y_{1/2\ell}(\alpha(0))$,

we can arrange that $|| p - x_n(\beta_n k/n) || < \epsilon'/2$ for some k and n sufficiently large. Similarly, $2Y \in \mathcal{I}_0$, so there exist controllers $\{u'_n\}$ with period $\tau'_n = \beta'_n/n$ such that $|| 2X_{1/2\ell}(p) - x'_n(\beta'_n k'/n') || < \epsilon'/2$ for some k', n'. Thus this concatenation of $\{u'_n\}$ and $\{u_n\}$ results in a piecewise smooth state trajectory $\tilde{x}_n$ which achieves $|| \alpha(t_1) - \tilde{x}_n(\tilde{\beta}_n t_1) || < \epsilon'$ for appropriate $\tilde{\beta}_n$ and n sufficiently large. Now we repeat the pattern (u_n followed by u'_n) to generate a piecewise smooth state trajectory $\tilde{x}_n$ for which $|| \alpha(t_k) - \tilde{x}_n(\tilde{\beta}_n t_k) || < k\epsilon'$ holds (k applications of the triangle inequality). Thus we can choose $\epsilon' = \epsilon/n$ to achieve incremental tracking of $X + Y$, hence $X + Y \in \mathcal{I}$. Now we can repeat the above argument using $\alpha X, \alpha Y$ to conclude that $\alpha(X+Y) \in \mathcal{I}$, hence $X + Y \in \mathcal{I}_0$. To show that $[X, Y] \in \mathcal{I}_0$ we argue as above. If $\ell > 0$, then $\sqrt{\ell}X, \sqrt{\ell}Y \in \mathcal{I}$. Consider the "switched integral curve" $\gamma(t)$ produced by following the integral curve for $-\sqrt{\ell}Y$ for $1/4\ell$ units of time, then the integral curve for $-\sqrt{\ell}X$ for $1/4\ell$ units of time, then the integral curve for $\sqrt{\ell}Y$ followed by that of $\sqrt{\ell}X$. Then

$$\gamma(1/\ell) = 4\sqrt{\ell}X_{1/4\ell}(4\sqrt{\ell}Y_{1/4\ell}(4\sqrt{\ell}X_{-1/4\ell}(4\sqrt{\ell}Y_{-1/4\ell}(x)))),$$

so that $\gamma(1/\ell) = X_{1/\sqrt{\ell}}(Y_{1/\sqrt{\ell}}(X_{-1/\sqrt{\ell}}(Y_{-1/\sqrt{\ell}}(x)))) = (Z(\ell)/\ell)_1(x)$, where $Z(\ell) \to [X, Y]$ as $\ell \to \infty$, hence $(Z_\ell/\ell)^\ell_1(x) \to [X, Y]_1(x)$ as $\ell \to \infty$, assuming X, Y, x fixed (cf. [18]). Continuing to switch between these integral curves, we generate $\gamma(k/\ell)$ for $k = 0, \ldots, \ell$. Thus, for ℓ sufficiently large,

$$|| \alpha(t_k) - \gamma(t_k) || < \epsilon/2,$$

where $\alpha(t)$ is an integral curve for $[X, Y]$ on $[0, 1]$, $t_k = k/\ell$, and $k = 0, 1, \ldots, \ell$. Since $4\sqrt{\ell}X, 4\sqrt{\ell}Y \in \mathcal{I}_0$, they can be incrementally tracked using periodic switched controllers $\{u_n\}$ and $\{u'_n\}$. We then argue as above to show that $[X, Y] \in \mathcal{I}$. Repeating these steps with $\sqrt{a}X, \sqrt{a}Y$ shows that $a[X, Y] \in \mathcal{I}$, hence $[X, Y] \in \mathcal{I}_0$. Finally, suppose that $X \in \mathcal{I}$, $Y \in \mathcal{I}_0$. Let $\alpha(t)$ be an integral curve for $X + Y$ on $[0, 1]$, m a positive integer, and $\epsilon > 0$. Then $mY \in \mathcal{I}$, and we define the "switched integral curve" $\gamma(t) = mY_t(x)$ for $0 \le t < 1/m\ell$, and $\gamma(t) = X_t(mY_{1/m\ell}(x))$ for $1/m\ell \le t < (m + 1)/m\ell$. Thus

$$\gamma\left(\frac{m+1}{m\ell}\right) = (X)_{1/\ell}((mY)_{1/m\ell}(x)) = X_{1/\ell}(Y_{1/\ell}(x)) = (Z(\ell)/\ell)_1(x),$$

where $Z(\ell) = (X + Y) + \frac{1}{2\ell}[X, Y] + \cdots$ (cf. [17, 18]). Continuing to switch between integral curves of X and Y, we get

$$\gamma\left(\frac{2(m+1)}{m\ell}\right) = X_{1/\ell}(Y_{1/\ell}(X_{1/\ell}(Y_{1/\ell}(x)))) = \left(\frac{Z(\ell)}{\ell}\right)^2_1(x), \ldots, \gamma\left(\frac{k}{\ell}\right) = \left(\frac{Z(\ell)}{\ell}\right)^k_1(x).$$

Then $Z(\ell) \to X + Y$ as $k \to \ell$ and $\ell \to \infty$ [17]. Now $\frac{k(m+1)}{m\ell} \to \frac{k}{\ell}$ as $m \to \infty$ so, for ℓ and m sufficiently large, $|| \alpha(t_k) - \gamma(t_k) || < \epsilon/2$, where $t_k = k/\ell$ and $k = 0, 1, \ldots, \ell$. Now repeat the argument used to show that $\mathcal{I}_0$ is closed under sums to conclude that $X + Y \in \mathcal{I}$.

(iii)(a) Suppose that $Y \in \mathcal{I}$, $X, ad^{k+1}_X Y \in \mathcal{I}_0$, and $[ad^i_X Y, ad^j_X Y] = 0$ for $j \le k$, $i \ge 3k - j$. Set $X^+ = n^{(2k+1)/k}X$, $X^- = -n^{(2k+1)/k}X \in \mathcal{I}_0$, and denote by $\psi(t)$ the switched integral curve which results from following the integral curve for X^- for $1/n^2$ units of time, where $\psi(0) = x \in M$, then following the integral curve for Y for $1/n^2$ units of time, and finally following the integral curve for X^+ for $1/n^2$ units of time. By construction,

$$\psi(3/n^2) = X_{n^{1/k}}(Y_{1/n^2}(X_{-n^{1/k}}(x))).$$

Noting that

$$X_t(Y_s(X_{-t}(x))) = s\sum_{n=0}^{\infty}\frac{t^n}{n!}ad_X^iY(x),$$

an absolutely convergent series for all t (cf. [17, 18]), we see that

$$\psi(3/n^2) \quad = \quad (G(n)+B(n))_{\frac{1}{n}}(x),$$

where

$$\begin{array}{lcl} G(n) & = & \frac{1}{n}Y + \frac{n^{1/k}}{n}ad_XY + \cdots + ad_X^kY, \\ B(n) & = & n^{1/k}ad_X^{k+1}Y + \cdots + n^{\ell/k}ad_X^{k+\ell}Y + \cdots. \end{array} \tag{3.1}$$

Since X and $ad_X^{k+1}Y \in \mathcal{I}_0$, a Lie algebra from (ii) above, it follows that $B(n) \in \mathcal{I}_0\ \forall n > 0$. In particular, the integral curve for $-B$ (writing B, G for $B(n), G(n)$) can be incrementally tracked by the state. This means that the switched integral curve $\gamma(t)$ which results from following the integral curve for $-nB$ for time $1/n^2$, followed by the switched integral curve $\psi(t)$, results in

$$\gamma(4/n^2) = (G+B)_{\frac{1}{n}}((B)_{-\frac{1}{n}}(x)). \tag{3.2}$$

Using the Baker–Campbell–Hausdorff formula [17], which converges for n sufficiently large, we have

$$\gamma\left(\frac{1}{n^2}\right) = \left(G + \frac{1}{2n}[G,B] + \frac{1}{12n^3}\{2[B,[G,B]] + [G,[G,B]]\} + \cdots\right)_{\frac{1}{n}}(x).$$

From the definitions for $G(n)$ and $B(n)$ and in light of *hypothesis* of Theorem 3.2(iii)(a) we have

$$\frac{1}{2n}[G(n),B(n)] = \sum_{j=0}^{k}\sum_{\ell=1}^{k-j-1}\frac{n^{\frac{j+\ell}{k}}}{n}[ad_X^jY, ad_X^{k+\ell}Y];$$

hence $\frac{1}{2n}[G(n),B(n)] \to 0$ as $n \to \infty$. Tedious applications of the Jacobi identity show that $\frac{1}{12n^3}\{2[B(n),[G(n),B(n)]] + [G(n),[G(n),B(n)]]\} \to 0$ as $n \to \infty$ as a consequence of Theorem 3.2(iii)(a), and the same conclusion applies to the higher order terms in the Baker–Campbell–Hausdorff series. In particular, we see that $\gamma(4/n^2) = (Z(n)/n)_1(x)$, where $Z(n) \to ad_X^kY$ as $n \to \infty$. Repeating ℓ times the switched integral curves used to generate $\gamma(t)$, we arrive at the state $\gamma(4\ell/n^2)$ and observe that $\gamma(4/n) = (Z(n)/n)_1^n(x) \to ad_X^kY_1(x)$ as $n \to \infty$. Thus $\gamma(t)$ is a switched integral curve of vector fields which can be incrementally tracked by the state of system (2.1). Furthermore, if $\beta_n = 4/n$ and $t_\ell = \ell/n$, then $\gamma(\beta_n t_\ell) \to (ad_X^\ell Y)_{t_\ell}(x)$ as $n \to \infty$. If $\alpha(t)$ is the integral curve for ad_X^kY with $\alpha(0) = x$, we have $\| \alpha(t_\ell) - \gamma(\beta_n t_\ell) \| < \epsilon/2$ for n sufficiently large and $\ell = 0, 1, \ldots, n$. Here γ switched between integral curves of vector fields which can be incrementally tracked. Thus we can repeat the argument used in (ii) above to show that there exist piecewise constant periodic controllers $\{u_n\}$ with periods $\tau_n = \beta_n/n$, where $\beta = 4/n$ such that $\| \alpha(t_\ell) - x_n(\beta_n t_\ell) \| < \epsilon$ for n sufficiently large and $\ell = 0, 1, 2, \ldots n$. This implies that $ad_X^kY \in \mathcal{I}$. If k is odd, we replace X with $-X$ and proceed as above to conclude that $-ad_X^kY \in \mathcal{I}$, from which we deduce that $ad_X^kY \in \mathcal{I}_0$.

(iii)(b) This is a particular case of (iii)(a).

(iv) This result is a consequence of (i) and (iii)(a). If we set $Y = f$ and $X = g$, then $Y \in \mathcal{I},\ X \in \mathcal{I}_0$ as a consequence of (i). Since $ad_X^{k+1}Y = 0$, it follows that $ad_X^{2k}Y = 0$; hence $ad_x^i Y = 0$ for $i \geq 3k - j$ and $j \leq k$, and (iii)(a) holds (and also (iii)(b)). In particular, we can conclude that $ad_X^k Y = ad_g^k f \in \mathcal{I}$. One can check that the controller u_n defined in (iv) is precisely the one used in the proof of (iii)(a). A more direct approach to the proof of (iv) is illuminating and is outlined below. Using the control $u_n(t)$ defined in (iv) (and $t_0 = 0$, to save accounting) we have

$$x_n(3/n^2) = (f + n^{(2k+1)/k} g)_{1/n^2}(f_{1/n^2}((f - n^{(2k+1)/k} g)_{1/n^2}(x_0))).$$

Applying the Baker–Campbell–Hausdorff formula (cf. [17]) two times, we can write $x_n(3/n^2) = (X(n)/n)_1(x_0)$. In the case $k = 2$ this yields (with the help of MAPLE V, Version 5.00 (Waterloo Maple, Waterloo, ON)) the expression

$$\begin{aligned} X(n) \;=\; & \tfrac{3}{n} f + \tfrac{2}{n^{1/2}}[g,f] + \tfrac{5}{6}[g,[g,f]] + \tfrac{1}{144n^4}[[f,g],[f,[f,g]]] \\ & + \tfrac{1}{72n^{3/2}}[[f,g],[g,[f,g]]] + \tfrac{1}{96n^{7/2}}[[f,[f,g]],[g,[f,g]]] \\ & + \tfrac{1}{576n^{11/2}}[[f,g],[[f,g],[f,[f,g]]]] - \tfrac{1}{576n^3}[[f,g],[[f,g],[g,[f,g]]]] \\ & + \tfrac{1}{3456n^5}[[g,[f,g]],[[f,g],[f,[f,g]]]] \\ & - \tfrac{1}{3456n^{5/2}}[[g,[f,g]],[[f,g],[f,[f,g]]]] + \cdots . \end{aligned}$$

Because $ad_g^3 f = 0$ it is not hard to show that all terms in $X(n)$, other than $ad_g^2 f = 0$, are multiplied by negative powers of n. In particular, $\lim_{n\to\infty} X(n) = \frac{5}{6}[g,[g,f]]$. A similar situation holds for other values of k, that is, $\lim_{n\to\infty} X(n) = c_k[g,[g,f]]$, where $c_k > 0$. Repeating the above, we find that $x_n(3\ell/n^2) = (X(n)/n)_1^\ell(x_0)$ and $x_n(3/n) = (X(n)/n)_1^n(x_0)$. But $\lim_{n\to\infty}(X(n)/n)^n(x_0) = \lim_{n\to\infty} X(n)_1(x_0) = c_k ad_g^k f_1(x_0)$, and hence $ad_g^k f$ is incrementally tracked by the state of system (2.1). □

Proof (*Theorem* 3.5).

(i) We can track an integral curve $\alpha(t)$ of f exactly using $u_n = 0$; thus $|| h^{\mathbf{P}}(\alpha(t)) - h^{\mathbf{P}}(x_n(t)) || = 0 \quad \forall t \in [0,1]$. This means that $f \in \mathcal{I}^p$. As noted in the proof of Theorem 3.2, we can find controllers u_n such that the corresponding state trajectory x_n closely follows the integral curves for γg for all $t \in [0,1]$ (not just for discrete times). Since the state trajectory x_n makes no large deviation from the integral curve of γg we have incremental tracking preserving $h^{\mathbf{P}}$.

(ii) In the proof of Theorem 3.2(ii) we saw that an integral curve $\alpha(t)$ of $X + Y$ can be tracked by switched integral curves of X and Y which stay close to $\alpha(t)$ for all $t \in [0,1]$. Since $X, Y \in \mathcal{I}^p$, we can find switched controllers $u_n,\ u'_n$ such that the corresponding state trajectories $x_n,\ x'_n$ incrementally track the integral curves of X and Y while preserving $h^{\mathbf{P}}$. Thus the image under $h^{\mathbf{P}}$ of the concatenation of $x_n,\ x'_n$ used to incrementally track $\alpha(t)$ will stay close to $h^{\mathbf{P}}(\alpha(t))$, and we will have incremental tracking of $X + Y$ preserving $h^{\mathbf{P}}$. The same situation holds in the case of $[X, Y]$.

(iii)(a) In the proof of Theorem 3.2(iii)(a), we constructed switched integral curves of X and Y which incrementally track integral curves of $ad_X^k Y$ but need not closely approximate these curves except at a discrete set of times. Thus the controllers u_n produce state trajectories x_n which incrementally track the integral curve $\alpha : t \mapsto (ad_X^k Y)(x)$ while making frequent and large deviations from $\alpha(t)$. By construction, these large motions are along integral curves for the vector fields X and $B(n) = \sum_{\ell=1}^{\infty} n^{\ell/k} ad_X^{k+\ell} Y$. Since $Xh = ad_X^{k+1} Y h^{\mathbf{P}} = 0$, we have $(X ad_X^{k+1} Y h^{\mathbf{P}} -$

$ad_X^{k+1}YXh^{\mathbf{p}}) = [X, ad_X^{k+1}Y]h^{\mathbf{p}} = 0$, and hence $\forall Z \in \{X, ad_X^{k+1}Y\}_{LA}$ we have $Zh^{\mathbf{p}} = 0$. In particular, $B(n)h^{\mathbf{p}} = 0$, and it follows that the large motions of the state trajectory x_n are in directions in which $h^{\mathbf{p}}$ does not vary. Thus we achieve incremental tracking of $\alpha(t)$ preserving $h^{\mathbf{p}}$.

(iii)(b) This is a particular case of (iii)(a).

(iv) Since $ad_g^{k+1}f = 0$ we have $g, ad_g^{k+1}f \in \mathcal{I}_0^p$. Clearly $ad_g^{k+1}fh^{\mathbf{p}} = 0$ and $gh^{\mathbf{p}} = (gh, gfh, \cdots, gf^{p-1}h)$. Since $p < r$ and $gf^ih = 0$ if $i \leq r-2$ (definition of relative order) we have $gh^{\mathbf{p}} = 0$, and the result follows from (iii)(b) above. □

4. Incremental sliding mode controllers. In the nonsingular case, the simple sliding mode controller (1.2) gives rise to vector fields $f + Lg$ and $f - Lg$ with several noteworthy properties. Given any compact subset C there exists $L > 0$, $\sigma_1, \sigma_2, \delta > 0$ such that

(i) on the set C

$$\begin{aligned}(f - Lg)s(h^{\mathbf{r}}) &< -\delta - \sigma_1,\\ (f + Lg)s(h^{\mathbf{r}}) &> +\delta + \sigma_2;\end{aligned}$$

(ii) $(f \pm Lg)h^{\mathbf{r-1}}(x) = fh^{\mathbf{r-1}}(x)$ if $r > 1$.

Remark 4.1. Suppose that y_d is a smooth function satisfying

$$-\sigma_1 \leq \frac{d}{dt}s(y_d^{\mathbf{r}}(t)) \leq \sigma_2$$

on $[t_0, +\infty)$. Then condition (i) implies that if the state stays in C, then the output will asymptotically track y_d using the simplified controller (1.2). In particular, if $s^p(x(t), t) > 0$ (so that we are "above" the sliding surface $\{s^p(x,t) = 0\}$) and $u = -L$, we have

$$\frac{d}{dt}s^p(x(t), t) = \frac{d}{dt}s(h^p(x(t)) - \frac{d}{dt}s(y_d^{\mathbf{r}}(t)).$$

From the definition of s (s is linear) we have $\frac{d}{dt}s(y^p(t)) = s(\dot{y}^p(t))$. Furthermore, $\frac{d}{dt}s(h^p(x(t))) = d(s\circ h^p)_{x(t)}\dot{x}(t) = d(s\circ h^p)_{x(t)}(f - Lg)(x(t)) = (f - Lg)(s\circ h^p)(x(t)) = (f - Lg)s(h^p(x(t))) < -\delta - \sigma_1$. This, combined with our assumption that $-\sigma_1 \leq s(\dot{y}_d^{\mathbf{r}}(t))$ or $-\frac{d}{dt}s(y_d^{\mathbf{r}}(t)) \leq \sigma_1$, yields $\frac{d}{dt}s^p(x(t), t) \leq -\delta_1 - \sigma_1 + \sigma_1 = -\delta_1$. In particular, the state trajectory returns to the sliding surface $\{s^p(x,t) = 0\}$. A similar situation results when $s^p(x(t), t) < 0$ and $u = L$.

Remark 4.2. Condition (ii) follows from the definition of the relative order r, since $gf^ih = 0$ for $i < r - 1$. That this is important in sliding mode control can be seen as follows: when the state "slides" on the sliding surface S_t^r, the trajectory is the integral curve of the "equivalent vector field" on S_t^r, which has the form $X = \alpha(f + Lg) + (1 - \alpha)(f - Lg)$, where $|\alpha(x)| \leq 1$ and $Xs(h^{\mathbf{r}}) = 0$ (cf. [3]). Note that $Xh^{\mathbf{r-1}} = (\alpha f + (1-\alpha)f)h^{\mathbf{r-1}} = fh^{\mathbf{r-1}}$. As a consequence, along this integral curve the tracking error satisfies the stable differential equation $s(e^{\mathbf{r}}(t)) = 0$, where s is defined by (2.2).

We seek to weaken the above in several ways. First we use the sliding surface S_t^p, where p is allowed be smaller than the relative order r of y. As a consequence of Theorem 3.5, $f \pm Lg \in \mathcal{I}^p$. We relax (ii) by allowing vector fields of the form $d^+, d^- \in \mathcal{I}^p$ such that $d^{\pm}h^{\mathbf{p-1}} = fh^{\mathbf{p-1}}$ and only require (i) above to hold on an open subset of Z of M which is invariant under the integral curves of d^+, d^-. We summarize these observations as follows.

DEFINITION 4.3. *Let X be a vector field on M. An open subset Z of M is said to be* invariant *with respect to a vector field X if, for all $x \in Z$, the integral curve $t \mapsto X_t(x)$ stays in Z.*

A2. There exists an open subset $Z \subset M$, invariant with respect to vector fields $d^+, d^- \in \mathcal{I}^p$, and constants $\alpha_1, \alpha_2 \in \mathbb{R}$, $\sigma_1, \sigma_2, \delta > 0$ such that

(i) on Z

$$\begin{array}{rcll} d^- s(h^{\mathbf{p}}) & < & -\delta - \sigma_1, & \text{when } \{s(h^{\mathbf{p}}) \geq \alpha_1\}, \\ d^+ s(h^{\mathbf{p}}) & > & +\delta + \sigma_2, & \text{when } \{s(h^{\mathbf{p}}) \leq \alpha_2\}; \end{array}$$

(ii) $d^{\pm} h^{\mathbf{p-1}}(x) = f h^{\mathbf{p-1}}(x)$ if $p > 1$.

If A2 holds for constants $\alpha_1, \alpha_2, \sigma_1, \sigma_2$, we define the following restricted class of desired output functions:

$$\mathcal{Y}_d = \{y_d \mid \alpha_1 \leq s(y_d^{\mathbf{p}}(t)) \leq \alpha_2,\ -\sigma_1 \leq s(\dot{y}_d^{\mathbf{p}}(t)) \leq \sigma_2 \ \ \forall t \geq t_0\}.$$

We will show that these outputs can be approximately tracked. We note that in the nonsingular case, A2 holds with $Z = M$, $d^+ = f + Lg$, and $d^- = f - Lg$ for L sufficiently large. If A2 holds with $d^{\pm} \in \mathcal{I}^p$, we define the set-valued map $F_d(x,t)$ by

$$F_d(x,t) = \begin{cases} d^+(x), & x \in \{s^p(x,t) < 0\}, \\ d^-(x), & x \in \{s^p(x,t) > 0\}, \\ \overline{\text{co}}\,\{d^+(x), d^-(x)\}, & x \in \{s^p(x,t) = 0\}, \end{cases} \tag{4.1}$$

where $\overline{\text{co}}\,\{d^+(x), d^-(x)\}$ is the closed convex hull generated by the $\{d^+(x), d^-(x)\}$.

THEOREM 4.4. *Suppose* A1, A2 *hold for system* (2.1). *Then there exist d^+, $d^- \in \mathcal{I}^p$ and an open subset $Z \subset M$ such that, for all smooth functions $y_d \in \mathcal{Y}_d$,*

(i) *the differential inclusion $\dot{x} \in F_d(x,t)$ with $x(t_0) \in Z$ has a unique solution $x_F(t) \in Z$ defined on $[t_0, +\infty)$;*

(ii) *for any solution x_f to $\dot{x} \in F_d(x,t)$ there exists $t_f \geq t_0$ such that $x_F(t) \in Z \cap S_t^p$ on $[t_f, +\infty)$;*

(iii) *for $t \geq t_f$ the curve $t \mapsto y_F(t) = h(x_F(t))$ is a smooth function of t which satisfies $s(y_F^{\mathbf{p}}(t) - y_d^{\mathbf{p}}(t)) = 0$. In particular, $\lim_{t\to\infty}(y_F^{\mathbf{p}}(t) - y_d^{\mathbf{p}}(t)) = 0$.*

Proof. By construction, $F_d(x,t)$ is nonempty, compact, and convex, and it is straightforward to show that F_d is upper semicontinuous with respect to x, t. Thus the *basic conditions* of [3, p. 76] are satisfied; the proof that local solutions to the *differential inclusion* $\dot{x}(t) = F_d(x(t),t)$ exist can be found in [3, pp. 67–68 and pp. 77–78] and is omitted here. That solutions stay in Z follows from A2, i.e., the assumption that Z is strongly invariant with respect to d^+, d^-. To establish uniqueness we note that both d^+ and d^- are transversal to $S_t^P \cap Z$ as a consequence of A2(i). Furthermore, the limiting vector fields on $S_t^P \cap Z$, which result from $d^+(x)(d^-(x))$ when x approaches $\{s^p = 0\}$ from $\{s^p > 0\}(\{s^p < 0\})$, define the opposite orientations on $S_t^P \cap Z$. Thus [3, Corollary 2, p. 108] implies that there is exactly one solution to this differential inclusion starting at $x(t_0) \in Z$.

(ii) Suppose that $y_d \in \mathcal{Y}_d$ and $s^p(x_F(t_0), t_0) < 0$. Then, from the definition of $\mathcal{Y}_d$, we have $-\sigma_1 \leq s(\dot{y}_d^{\mathbf{p}}(t)) \leq \sigma_2$. But $d^+ s(h^{\mathbf{p}}(x_F(t))) > \delta + \sigma_2$ by A2(i). Thus $\frac{d}{dt} s^p(x_F(t),t) = d^+ s(h^{\mathbf{p}}(x_F(t))) - s(\dot{y}_d^{\mathbf{p}}(t)) \geq \delta + \sigma_2 - \sigma_2 = \delta > 0$, and $s^p(x,t)$ is strictly increasing along integral curves of d^+ if $\{s^p < 0\}$. Similarly, $s^p(x,t)$ is strictly decreasing along integral curves of d^- in $\{s^p > 0\}$. Since $x_F(t) \in Z\ \forall t \geq t_0$ by (i), we have established (ii).

(iii) For $t \geq t_f$, x_F is a smooth integral curve for the equivalent vector field X defined in Remark 4.2. Here $Xh^{\mathbf{p-1}} = fh^{\mathbf{p-1}}$ as a consequence of A2(ii), hence $y^{(i)} = X^i h = f^i h$ for $i = 1, 2, \ldots, p-1$. From section 2 we know that if $y^{(i)} = f^i h$ for $i = 1, 2, \ldots, p-1$, then $s(y^{\mathbf{p}}(t) - y_d^{\mathbf{p}}(t)) = s(e^{\mathbf{p}}(t)) = 0$ is equivalent to $s^p(x(t), t) = 0$ or $x(t) \in S_t^p$. In particular, since $x_F(t) \in S_t^p$ from (ii), we have $s(e^{\mathbf{p}}(t)) = 0 \; \forall t > t_f$ and $\lim_{t\to\infty}(y_F^{\mathbf{p}}(t) - y_d^{\mathbf{p}}(t)) = 0$. □

A necessary condition for approximate tracking of y_d is that both $y_d^{\mathbf{p}}$ and the state trajectory remain bounded. In the nonsingular case the state trajectory and the solution to the differential inclusion $\dot{x} \in F_d(x, t)$ are identical, and it suffices to ensure that solutions to $\dot{x} \in F_d(x, t)$ with $x(t_0) \in E_t^p \cap Z$ remain bounded. In our case the same assumption suffices.

A3. Suppose that A2 holds for system (2.1) and $y_d \in \mathcal{Y}_d$. Then solutions to the differential inclusion $\dot{x} \in F_d(x, t)$, with initial state $x(t_0) \in E_t^p \cap Z$, remain bounded for $t \in [t_0, +\infty)$.

Remark 4.5. Note that in light of Theorem 4.4 (ii) it suffices to study the trajectory on S_t^p. Since there is a unique vector field $G(x, t)$ in $\overline{\text{co}}\,\{d^+(x), d^-(x)\}$ that makes $\frac{\partial}{\partial t} + G(x, t)$ tangent to S_t^p, it suffices to check that this one integral curve is bounded. A sufficient (but far from necessary) condition for A3 to hold is that $Z \cap S_t^p$ be bounded.

Suppose that A1, A2, A3 hold for system (2.1) with initial state $x(t_0) \in Z$, where Z is an open subset of M invariant with respect to vector fields $d^+,\ d^- \in \mathcal{I}^p$. If we could make the state of (2.1) exactly track the solution $x_F(t)$ to $\dot{x} \in F_d(x, t)$, then Theorem 4.4 would imply asymptotic tracking of y_d. We now describe a "digital controller" which allows us to incrementally track x_F and approximately track y_d. We are motivated by the typical "sample and hold" digital controller with fixed sample rate T. That is, if $u(x, t)$ is a smooth function of x and t, the digital controller $u_k(t)$ takes the form

$$\begin{aligned} u_k(t) &= u(x_k, t_k) \quad \text{for } t_k \leq t < t_{k+1}, \\ t_{k+1} &= t_k + T, \\ x_k &= x(t_k), \end{aligned}$$

where $x(t_k)$ is the state at time t_k which results from using the control u_k on the time interval $[t_{k-1}, t_k)$. We have controllers $u_n^+(x, t)$ and $u_n^-(x, t)$, which are piecewise constant periodic functions of t with periods $\tau_n^+ = \beta_n^+/n$ and $\tau_n^- = \beta_n^-/n$, respectively, and which cause the state of (2.1) to incrementally track integral curves of $d^+,\ d^-$, respectively. Thus we require a digital controller with variable sampling rate. We define our *digital controller* for system (2.1) as follows:

$$\begin{aligned} u_k(x, t) &= \begin{cases} u_n^+(x, t) & \text{for } t_k \leq t < t_{k+1},\ s^p(x_k, t_k) < 0, \\ u_n^-(x, t) & \text{for } t_k \leq t < t_{k+1},\ s^p(x_k, t_k) \geq 0, \end{cases} \\ t_{k+1} &= \begin{cases} t_k + \tau_n^+ & \text{if } s^p(x_k, t_k) < 0, \\ t_k + \tau_n^- & \text{if } s^p(x_k, t_k) \geq 0, \end{cases} \\ x_k &= x(t_k). \end{aligned} \tag{4.2}$$

We observe that while $u_k(x, t)$ is not constant with respect to t over $[t_k, t_{k+1})$, it is piecewise constant due to the piecewise constant time dependence of u_n^+ and u_n^-.

THEOREM 4.6. *Under assumptions* A1, A2, A3*, the switched controller* (4.2) *achieves the following property for the closed loop system: if* $x(t_0) = x_0 \in Z$ *and*

$y_d \in \mathcal{Y}_d$, then, for n sufficiently large, the output y of (2.1) approximately tracks y_d to degree p.

Proof. Let $x_F(t)$ denote the solution to the differential inclusion $\dot{x} \in F_d(x,t)$; $x(t_0) = x_0 \in Z$ and let $\epsilon > 0$. From Theorem 4.4 there exists $t_f \geq t_0$ such that $x_F(t) \in Z \cap S_t^p \ \forall t \geq t_f$ and $\lim_{t\to\infty}(y_F^{\mathbf{p}}(t) - y_d^{\mathbf{p}}(t)) = \lim_{t\to\infty}(h^{\mathbf{p}}(x_F(t)) - y_d^{\mathbf{p}}(t)) = 0$. This implies that $x_F(t) \to E_t^p \cap Z$ as $t \to \infty$ and, in light of A3, $x_F(t)$ is a bounded function of t. We first consider the case where $s^p(x_0,t_0) < 0$. Then for $t_0 \leq t < t_f$ we have $s^p(x_F(t),t) < 0$, $F_d(x_F(t),t) = d^+(x_F(t))$, $u_k = u_n^+$, $t_1 = t_0 + \tau_n^+ = t_0 + \beta_n^+/n$, $t_2 = t_0 + 2\beta_n^+/n, \ldots,$ and $t_k = t_0 + k\beta_n^+/n$. The vector field d^+ is incrementally tracked by the state trajectory x_n^+ produced by u_n^+. We now calculate the rate of change of $s^p(x(t),t)$ when $x(t)$ is the integral curve $x_F(t)$ of d^+ but time t is rescaled to match the time rescaling which occurs in incremental tracking. For $t < t_f$ we have, from A2 and the linearity of s,

$$\begin{aligned}
\tfrac{d}{dt}s^p(x_F(t_0+t), t_0+\beta_n^+ t) &= d^+ s(h^{\mathbf{p}}(x_F(t_0+t))) - \tfrac{d}{dt}s(y^{\mathbf{p}}(t_0+\beta_n^+ t)) \\
&\geq \delta + \sigma_2 - s(\beta_n^+ \dot{y}^{\mathbf{p}}(t_0+\beta_n^+ t)) \\
&= \delta + \sigma_2 - \beta_n^+ s(\dot{y}^{\mathbf{p}}(t_0+\beta_n^+ t)) \\
&\geq \delta + \sigma_2 - \beta_n^+ \sigma_2 \\
&\geq \delta
\end{aligned}$$

as $0 < \beta_n \leq 1$. Thus there is some least time $t_1 > 0$ such that $s^p(x_F(t_0+t_1), t_0 + \beta_n^+ t_1) = 0$ and some positive integer k_1 (depending on n) such that $x_F(t_0 + k/n) \leq x_F(t_0+t_1) \leq x_F(t_0+(k+1)/n)$. We can make $||x_F(t_0+t_1) - x_F(t_0+(k_1+1)/n)||$ arbitrarily small by increasing n, and hence

$$||s^p(x_F(t_0+(k_1+1)/n), t_0+\beta_n^+(k_1+1)/n)|| < \epsilon/4$$

for n sufficiently large. Since $x_F(t)$ is incrementally tracked by $x_n^+(t)$, we have $||x_F(t_0+k/n) - x_n^+(t_0+\beta_n^+ k/n)|| \to 0$ as $n \to \infty$. Therefore by picking n large enough we ensure that $||s^p(x_n^+(t_0+\beta_n^+ k/n), t_0+\beta_n^+ k/n)|| < \epsilon/2$ for $k = 0, \ldots, n$. In particular, using the "digital" controller (4.2) results in a state trajectory $x_n^+(t)$ for (2.1) with the property that, for n sufficiently large, $s^p(x_n^+(t_k), t_k) < 0$ for $k = 0, 1, \ldots, \ell_1 - 1$, $s^p(x_n^+(t_{\ell_1}), t_{\ell_1}) > 0$, and $||s^p(x_n^+(t_{\ell_1}), t_{\ell_1})|| < \epsilon/2$. Thus $u_k = u_n^+$ for $k = 0, \ldots, \ell_1 - 1$ and $u_{\ell_1} = u_n^-$. We can now repeat the above, starting from the initial state $x(t_{\ell_1}) = x_n^+(t_{\ell_1})$. Since $s^p(x_n^+(t_{\ell_1}), t_{\ell_1}) > 0$ we now incrementally track the integral curve of d^- until $s^p(x_n^-(t_{\ell_2}), t_{\ell_2}) < 0$ and $||s^p(x_n^-(t_{\ell_2}), t_{\ell_2})|| < \epsilon/2$ (increasing n if necessary). Because the integral curve $x_F(t)$ is bounded, we can choose n sufficiently large to continue the above switching and ensure that the state trajectory x_n resulting from the controller (4.2) satisfies

$$||s^p(x_n(t_k), t_k)|| < \epsilon/2 \quad \forall k \geq \ell_1.$$

Incremental tracking ensures that x_n is close to x_F at discrete times $\{t_k\}$, but for $t_k < t < t_{k+1}$ we may have $x_n(t)$ far from $x_F(t)$. We now use the fact that $d^{\pm} \in \mathcal{I}^p$, and thus are incrementally tracked preserving $h^{\mathbf{p}}$, to show that s^p is unaffected by these deviations. In particular, on $[t_0, t_f]$ we have $||h^{\mathbf{p}}(x_F(t_0+t)) - h^{\mathbf{p}}(x_n^+(t_0+\beta_n t))|| \to 0$ as $n \to \infty$, by definition of incrementally tracking. This allows us to ensure that $||s^p(x_n(t),t)|| < \epsilon \ \forall t \geq t_0 + \ell_1/n$ for n sufficiently large. Because $\beta \leq 1$ this implies that, for n sufficiently large,

$$||s^p(x_n(t),t)|| < \epsilon \quad \forall t \geq t_f,$$

and $x_n(t)$ is bounded; hence the output y of (2.1) approximately tracks y_d to degree p. □

Let $\mathcal{R}(x_0)$ denote the set of states which can be reached from the initial state $x(t_0) = x_0$. Theorem 4.6 ensures approximate tracking if $x_0 \in Z$, and so it is natural to look for a controller which steers x_0 to the open set Z in finite time. It will often be the case that $\mathcal{R}(x_0) \cap Z \neq \emptyset$. In particular, we need to use the above theorem when the state trajectory "sticks" to the singular submanifold under the naive truncated sliding mode controller. Thus if Z intersects the singular submanifold it is likely reachable from the initial state. Suppose that C is compact, Z is an open subset of M, and $u_0(x,t)$ is a controller for system (2.1) which transfers the state from $x(t_0) = x_0 \in M$ to $x(t_1) = x_1 \in Z \cap C$. Define the hybrid controller as

$$u_d(x,t) = \begin{cases} u_0(x,t), & t \in [t_0, t_1), \\ u_k(x,t), & t \in [t_k, t_{k+1}), \end{cases} \tag{4.3}$$

where u_k is the digital controller (4.2), and $k \geq 1$.

THEOREM 4.7. *Suppose that* A1, A2, A3 *hold, $C \subset M$ is compact, and there exists a controller $u_0(x,t)$ which transfers the state of system* (2.1) *to $Z \cap C$ at time $t_1 \geq t_0$. Then, for n sufficiently large, the hybrid switched controller* (4.3) *achieves the following property for the closed loop system: if $y_d \in \mathcal{Y}_d$, then the output y of* (2.1) *approximately tracks y_d to degree p.*

Proof. For an initial state $x(t_1) = x_1 \in Z \cap C$, Theorem 4.6 implies that, for n_1 sufficiently large, the controller (4.3) achieves approximate tracking of y_d. From the continuity of solutions to $\dot{x} \in F_d(x,t)$ with respect to the initial conditions (cf. [3]), we have approximate tracking of y_d for any initial state in some open neighbourhood U_1 of x_1. Because $Z \cap C$ is compact we can obtain a finite open covering $\cup_{i=1}^{m} U_i$ of $Z \cap C$ by such open sets. Thus the hybrid switched controller (4.3) with $n \geq max\{n_i | i = 1, \ldots, m\}$ results in approximate tracking of y_d. □

Remark 4.8. We note that the hypotheses of Theorem 4.7 are satisfied for affine systems whose singular set $\{gf^{r-1}h(x) = 0\}$ is empty. In this case we use $p = r$, $Z = M$, and $d^{\pm} = f \pm Lg$, and u_0 is not needed. To verify the hypotheses of Theorem 4.7 for a given system model, one could start by using Theorems 3.2 and 3.5 to find vector fields which the state trajectory can incrementally track. If the natural sliding mode controller has a singular submanifold N, check to see if the vector fields which can be incrementally tracked preserving the output map are sufficient for A2 to hold. Then, if A3 holds as well (see Remark 4.5), Theorems 4.6 and 4.7 yield a controller. Example 1.1 is a case in point.

Example 4.9 (*Example* 1.1 *continued*). We have seen that $f, ad_g^2 f \in \mathcal{I}^p$ for $p = 1$, and so it is natural to choose a sliding surface with $p = 1$. We set $s(e^{\mathbf{p}}) = e$, where $e = y - y_d$. Then $p = 1$, $s^1(x,t) = x_1 - y_d(t)$, $s(h^{\mathbf{p}}(x)) = x_1$, and $S_t^1 = E_t^1 = \{x_1 = y_d(t)\}$. Clearly the set S_t^1 is an embedded submanifold (2-dimensional) for each fixed time t, so that A1 holds. Here $gs(h^{\mathbf{p}}) = 0$, but $fs(h^{\mathbf{p}}(x)) = x_3^2 - x_2$ and $(ad_g^2 f)s(h^{\mathbf{p}}(x)) = 2$. To satisfy A2(i) we want $(ad_g^2 f)s(h^{\mathbf{p}}) > 0$ and $fs(h^{\mathbf{p}}) < 0$ on some open set Z, so it is natural to look for a set invariant with respect to $f, ad_g^2 f$ and on which $x_2 \approx q_2 > 0$, $x_3 \approx 0$. For many systems a systematic approach to finding a suitable subset Z may not be possible, but for the example under consideration x_2, x_3 satisfy a linear differential equation. Thus we can find such a set by constructing a Lyapunov function. In particular, let $z_1(x) = x_2 - q_2$, $z_2(x) = x_3$, and $u = -a_0 z_1 - a_1 z_2$ $(a_0, a_1 > 0)$.

Then $\dot{z} = Az$ for

$$A = \begin{bmatrix} 0 & 1 \\ -a_0 & -a_1 \end{bmatrix}.$$

We can find a Lyapunov function $V = z^T Q z$ for $\dot{z} = Az$ by solving Lyapunov equations $A^T Q + QA = -I$ for the positive definite matrix

$$Q = \begin{bmatrix} a & b \\ b & c \end{bmatrix},$$

where $a = (a_1^2 + a_0(a_0+1))/(2a_0a_1)$, $b = 1/2a_0$, $c = (a_0+1)/(2a_0a_1)$. Then for $q_0 > 0$ we define

$$Z(q_0) = \{x \in \mathbb{R}^3 \mid z(x)^T Q z(x) < q_0\} = \{a(x_2 - q_2)^2 + 2b(x_2 - q_2)x_3 + cx_3^2 < q_0\},$$

where $q_0 > 0$. By construction, $Z(q_0)$ is invariant with respect to $f - (a_0 z_1 + a_1 z_2)g = f + \gamma g$, where $\gamma(x) = -a_0(x_2 - q_2) - a_1 x_3$. Since $g \in \mathcal{I}_0^p$ and $f \in \mathcal{I}^p$, we have $d^- = f + \gamma g \in \mathcal{I}^p$ as a consequence of Theorem 3.2. Since, by construction, V is decreasing along the integral curves of d^-, we have $Z(q_0)$ invariant with respect to d^-. Because $Z(q_0)$ puts no restrictions on x_1 it is also invariant with respect to $d^+ = q_1 a d_g^2 f \in \mathcal{I}^p$, where $q_1 > 0$. To verify that A2 holds we first note that $s(h^{\mathbf{p}}(x)) = x_1$; hence $d^- s(h^{\mathbf{p}}(x)) = (f + \gamma g)x_1 = x_3^2 - x_2$ and $d^+ s(h^{\mathbf{p}}(x)) = q_1 a d_g^2 f x_1 = 2q_1 > 0$. Note that by shrinking q_0 we can ensure that in the set $Z(q_0)$ we have x_3 arbitrarily close to 0 and x_2 arbitrarily close to q_2. In particular, given *any* constants $\alpha_1, \alpha_2 \in \mathbb{R}$, $\sigma_1, \sigma_2, \delta > 0$, we can choose q_0, q_1, q_2 such that on the set $Z = Z(q_0)$ we have

$$\begin{array}{lcll} d^- s(h^{\mathbf{p}}) & < & -\delta - \sigma_1 & \text{when } \{s(h^{\mathbf{p}}) \geq \alpha_1\}, \\ d^+ s(h^{\mathbf{p}}) & > & +\delta + \sigma_2 & \text{when } \{s(h^{\mathbf{p}}) \leq \alpha_2\}. \end{array}$$

Thus A2(i) holds and A2(ii) holds automatically as $p = 1$. In light of Remark 4.5, assumption A3 will hold if $Z \cap S_t^p$ is bounded. Here $Z = Z(q)$ is a bounded set by construction, and hence A3 holds. Thus A1, A2, A3 hold and Theorem 4.7 implies that we can approximately track to degree 1 the set of output paths

$$\mathcal{Y}_d = \{y_d \mid \alpha_1 \leq s(y_d^{\mathbf{p}}(t)) \leq \alpha_2,\ -\sigma_1 \leq s(\dot{y}_d^{\mathbf{p}}(t)) \leq \sigma_2 \ \ \forall t \geq t_0\}.$$

The construction of the controller u_0 which moves the state into Z is simplified here because Z is the level set $\{V(z(x)) = q_0\}$ of a Lyapunov function for $\dot{z} = Az$, where $u = \gamma g$. We set $u_0(x,t) = \gamma(x)$. For any $x(t_0)$ there will be $t_1 \geq t_0$ such that $x(t_1) \in Z$. We incrementally track d^- using $u_n^-(x,t) = \gamma(x)$ ($\beta_n^- = 1$) and incrementally track d^+ using the controller from Theorem 3.2(iv) with $k = 2$, namely,

$$u_n^+(x,t) = \begin{cases} -n^{5/2}, & t_k \leq t < t_k + 1/n^2, \\ 0, & t_k + 1/n^2 \leq t < t_k + 2/n^2, \\ n^{5/2}, & t_k + 2/n^2 \leq t < t_k + 3/n^2, \end{cases}$$

where $\beta_n^+ = 3/n$, $\tau_n^- = 1/n$, and $\tau_n^+ = 3/n^2$. If we want $y_d(t) = \sin t \in \mathcal{Y}_d$, we can define Z by choosing $q_0 = 0.5$, $q_1 = 1$, $q_2 = 2$, $a_1 = 40$, $a_2 = 400$. To ensure close tracking we pick $\epsilon = 0.1$. Figure 3 shows a SIMNON simulation using the controller (4.3) with $n = 5$ and $x(0) = (2, 2, 0) \in Z$. The tracking performance is not particularly sensitive to variations in these parameters. Increasing n gives tighter tracking but requires more control effort.

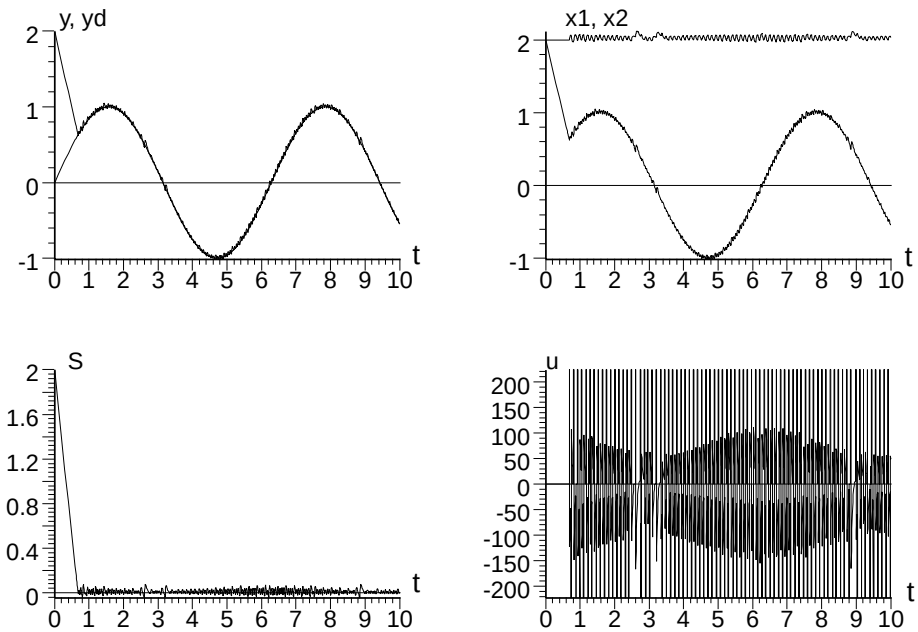


FIG. 3. *Approximate tracking of a* $y_d(t) = \sin t$ *with* $x_0 \in Z$.

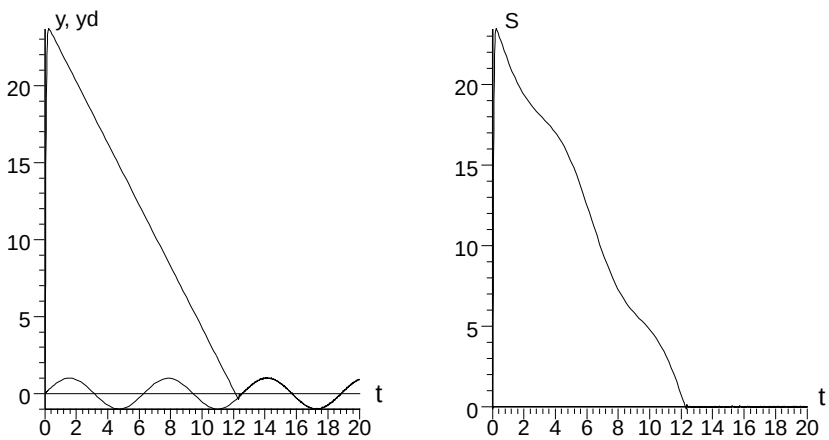


FIG. 4. *Approximate tracking of a* $y_d(t) = \sin t$ *with* $x_0 \notin Z$.

In Figure 4 we show the effect of an initial state which is initially well outside of Z ($x_2(0) = -0.1 < 0$).

We note that in this situation state trajectories resulting from controllers based on relative degree will stick to the singular manifold $N = \{x_3 = 0\}$ and send $s(e^{\mathbf{r}}(t)) \to \infty$. Our approach has the state passing back and forth across N. The initial delay is due to the requirement that the state must enter Z before our switched controller can act to reduce s.

5. Conclusions. There are situations in which it is useful to be able to control the state of a system so that it closely approaches a given curve at discrete times. We have introduced the concept of *incremental tracking* of integral curves, where the state trajectory (with reparametrized time) closely approaches an integral curve at discrete times. These controllers were then applied to sliding mode control, where the state trajectory used to reach the sliding surface is not very critical. Our discontinuous "digital sliding mode controller" achieved approximate tracking in situations where the natural truncated sliding mode controller (and the natural truncated smooth controller based on inversion) fails.

REFERENCES

[1] R. W. Brockett, *Characteristic phenomena and model problems in nonlinear control*, in Proceedings of the IFAC Congress, San Francisco, CA, 1996, Vol. G, pp. 135–140.

[2] P. E. Crouch, I. Ighneiwa, and F. Lamnabhi-Lagarrigue, *On the singular tracking problem*, Math. Control Signals Systems, 4 (1991), pp. 341–362.

[3] A. F. Filippov, *Differential Equations with Discontinuous Righthand Sides*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1988.

[4] V. Guillemin and A. Pollack, *Differential Topology*, Prentice–Hall, Englewood Cliffs, NJ, 1974.

[5] J. Hauser, S. Sastry, and P. Kokotovic, *Nonlinear control via approximate input-output linearization: The ball and beam example*, IEEE Trans. Automat. Control, 37 (1992), pp. 392–398.

[6] R. M. Hirschorn, *Singular sliding mode control*, IEEE Trans. Automat. Control, 46 (2001), pp. 276–286.

[7] R. M. Hirschorn and E. Aranda-Bricaire, *Global approximate output tracking for nonlinear systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 1389–1398.

[8] A. Isidori, *Nonlinear Control Systems*, 2nd ed., Springer-Verlag, Heidelberg, Germany, 1989.

[9] B. Jakubczyk and F. Lamnabhi-Lagarrigue, *On tracking through singularities: Regularity of the control*, Systems Control Lett., 21 (1993), pp. 271–276.

[10] W. Liu and H. J. Sussmann, *Limits of highly oscillatory controls and the approximation of general paths by admissible trajectories*, in Proceedings of the 30th IEEE CDC, IEEE Publications, Piscataway, NJ, 1991, pp. 437–442.

[11] H. Nijmeijer and A. van der Schaft, *Nonlinear Dynamical Control Systems*, Springer-Verlag, New York, 1990.

[12] J.-B. Pomet, *On the curves that may be approached by trajectories of a smooth affine system*, Systems Control Lett., 36 (1999), pp. 143–149.

[13] J.-J. E. Slotine, *Applied Nonlinear Control*, Prentice–Hall, Englewood Cliffs, NJ, 1991.

[14] J.-J. E. Slotine, *Sliding controller design for nonlinear systems*, Internat. J. Control, 40 (1984), pp. 421–434.

[15] C. J. Tomlin and S. S. Sastry, *Switching through singularities*, in Proceedings of the 36th IEEE CDC, San Diego, CA, IEEE Publications, Piscataway, NJ, 1997, pp. 1–6.

[16] V. I. Utkin, *Sliding Modes in Control and Optimization*, Communications and Control Engineering Series, Springer-Verlag, Berlin, 1992.

[17] V. S. Varadarajan, *Lie Groups, Lie Algebras, and Their Representations*, Prentice–Hall, Englewood Cliffs, NJ, 1974.

[18] F. W. Warner, *Foundations of Differentiable Manifolds and Lie Groups*, Scott, Foresman, Glenview, IL, 1971.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 4, pp. 1011–1041

RATE OF CONVERGENCE FOR CONSTRAINED STOCHASTIC APPROXIMATION ALGORITHMS*

ROBERT BUCHE† AND HAROLD J. KUSHNER†

Abstract. There is a large literature on the rate of convergence problem for general unconstrained stochastic approximations. Typically, one centers the iterate θ_n about the limit point $\bar{\theta}$ and then normalizes by dividing by the square root of the step size ϵ_n. Then some type of convergence in distribution or weak convergence of U_n, the centered and normalized iterate, is proved. For example, one proves that the interpolated process formed by the U_n converges weakly to a stationary Gaussian diffusion, and the variance of the stationary measure is taken to be a measure of the rate of convergence. See the references in [A. Benveniste, M. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximation*, Springer-Verlag, Berlin, New York, 1990; L. Gerencér, *SIAM J. Control Optim.*, 30 (1992), pp. 1200–1227; H. J. Kushner and D. S. Clark, *Stochastic Approximation for Constrained and Unconstrained Systems*, Springer-Verlag, Berlin, New York, 1978; H. J. Kushner and G. Yin, *Stochastic Approximation Algorithms and Applications*, Springer-Verlag, Berlin, New York, 1997; M. T. Wasan, *Stochastic Approximation*, Cambridge University Press, Cambridge, UK, 1969] for algorithms where the step size either goes to zero or is small and constant. Large deviations provide an alternative approach to the rate of convergence problem [P. Dupuis and H. J. Kushner, *SIAM J. Control Optim.*, 23 (1985), pp. 675–696; P. Dupuis and H. J. Kushner, *SIAM J. Control Optim.*, 27 (1989), pp. 1108–1135; P. Dupuis and H. J. Kushner, *Probab. Theory Related Fields*, 75 (1987), pp. 223–244; A. P. Korostelev, *Stochastic Recurrent Processes*, Nauka, Moscow, 1984; H. J. Kushner and G. Yin, *Stochastic Approximation Algorithms and Applications*, Springer-Verlag, Berlin, New York, 1997]. When the iterates of the algorithm are constrained to lie in some bounded set, the limit point is frequently on the boundary. With the exception of the large deviations type [P. Dupuis and H. J. Kushner, *SIAM J. Control Optim.*, 23 (1985), pp. 675–696; P. Dupuis and H. J. Kushner, *Probab. Theory Related Fields*, 75 (1987), pp. 223–244], the rate of convergence literature is essentially confined to the case where the limit point is not on a constraint boundary.

When the limit point is on the boundary of the constraint set the usual steps are hard to carry out. In particular, the stability methods which are used to prove tightness of the normalized iterates cannot be carried over in general, and there is the problem of proving tightness of the normalized process and characterizing the limit process.

This paper develops the necessary techniques and shows that the stationary Gaussian diffusion is replaced by an appropriate stationary reflected linear diffusion, whose variance plays the same role as a measure of the rate of convergence. An application to constrained function minimization under inequality constraints $q^i(x) \leq 0, i \leq p$, is given, where both the objective function and the constraints are observed in the presence of noise. The iteration is on both the basic state variable and a Lagrange multiplier, which is constrained to be nonnegative. If a limit multiplier value for an active constraint is zero, then the classical method for computing the rate cannot be used, but (under appropriate conditions) it is a special case of our results. Rate of convergence results are important because, among other reasons, they immediately yield the advantages of iterate averaging methods, as noted in [H. J. Kushner and G. Yin, *Stochastic Approximation Algorithms and Applications*, Springer-Verlag, Berlin, New York, 1997].

Key words. stochastic approximation, constrained stochastic approximation, rate of convergence, recursive algorithms, weak convergence

AMS subject classifications. 60F17, 62L20, 93E15, 93E20, 93E35

PII. S0363012999361639

*Received by the editors September 20, 1999; accepted for publication (in revised form) May 12, 2001; published electronically November 28, 2001.

http://www.siam.org/journals/sicon/40-4/36163.html

†Division of Applied Mathematics, Brown University, Providence, RI 02912 (buche@cfm.brown.edu, hjk@dam.brown.edu). The research of the first author was supported in part by NSF grant ECS 9703895 and ARO contract DAAG55-98-1-0158. The research of the second author was supported in part by NSF grant ECS 9703895 and ARO contract DAAD19-99-1-0-223.

1. Introduction. There is an extensive theory concerning the rate of convergence of the SA (stochastic approximation)-type algorithms of the forms

$$\theta_{n+1}^\epsilon = \theta_n^\epsilon + \epsilon Y_n^\epsilon, \text{ constant step size,} \tag{1.1}$$

and

$$\theta_{n+1} = \theta_n + \epsilon_n Y_n, \text{ decreasing step size, } \epsilon_n \to 0. \tag{1.2}$$

Here the adjustable parameter or state θ is in $\mathbb{R}^r$, Euclidean r-space. See [21] for a comprehensive development of rate results under both probability one and weak convergence assumptions on the $\{\theta_n^\epsilon\}$ or $\{\theta_n\}$. The theory also covers correlated and state dependent noise. See also [2, 14]. Let us write

$$\begin{aligned} Y_n^\epsilon &= g(\theta_n^\epsilon) + \xi_n^\epsilon, \\ Y_n &= g(\theta_n) + \xi_n, \end{aligned} \tag{1.3}$$

where $g(\cdot)$ plays the role of a "mean" or "centering" function and ξ_n^ϵ, ξ_n are the so-called "noises" on which we will make further assumptions below.

In proving rate of convergence results, one usually starts by assuming some appropriate type of convergence of θ_n or θ_n^ϵ to a limit point $\bar\theta$. For θ_n, this convergence might be in either the probability one or in the weak convergence senses, and for θ_n^ϵ it is in the weak convergence sense. Define the matrix $g_\theta(\cdot)$, whose ith *row* is the gradient of the ith component of the "centering" vector $g(\cdot)$ with respect to θ. Define $A = g_\theta(\bar\theta)$. The usual procedure is to work with the normalized iterates defined by $U_n^\epsilon = (\theta_n^\epsilon - \bar\theta)/\sqrt{\epsilon}$ and $U_n = (\theta_n - \bar\theta)/\sqrt{\epsilon_n}$, resp. Define the interpolated processes $U^\epsilon(\cdot)$ by $U^\epsilon(t) = U_n^\epsilon$ for $t \in [n\epsilon, n\epsilon + \epsilon)$. For the decreasing step case, define $t_n = \sum_{i=0}^{n-1} \epsilon_i$. Then, define $U^n(\cdot)$ by $U^n(0) = U_n$ and, for $t > 0$, $U^n(t) = U_{n+i}$ for $t \in [t_{n+i} - t_n, t_{n+i+1} - t_n)$. Let $m(t)$ denote the unique value of n such that $t_n \le t < t_{n+1}$.

One starts the proof of the "rate" result for the unconstrained case by proving tightness of the set of interest (under appropriate conditions, and where n_ϵ might have to go to infinity as $\epsilon \to 0$)

$$\{U_n^\epsilon; \epsilon > 0, n \ge n_\epsilon\} \tag{1.4a}$$

or of

$$\{U_n, n < \infty\}. \tag{1.4b}$$

Given this tightness, one continues the proof by proving the weak convergence of either $U^\epsilon(t_\epsilon + \cdot)$ (as $\epsilon \to 0$, with t_ϵ going to infinity fast enough) or of $U^n(\cdot)$ (as $n \to \infty$) to a stationary diffusion process of the type

$$dU = AU dt + \sigma dw \tag{1.5a}$$

or of the type

$$dU = \left[A + \frac{I}{2}\right] U dt + \sigma dw. \tag{1.5b}$$

Here σ is a constant matrix, and $w(\cdot)$ is a standard vector-valued Wiener process. Equation (1.5a) is the goal under (2.5a) or if the step size is ϵ. Equation (1.5b) is

the goal under (2.5b). In (1.5a) it is assumed that A is Hurwitz, and in (1.5b) that $A + I/2$ is Hurwitz. The stationary covariance Σ_U of $U(\cdot)$ is taken to be the desired measure of the rate of convergence. Note that the result implies that U_n converges in distribution to a normally distributed random variable with mean zero and covariance Σ_U, and similarly for U_n^ϵ if $n \to \infty$ fast enough as $\epsilon \to 0$. If A (resp., $A + I/2$) is Hurwitz, then the theory is well known under quite general conditions.

The constrained problem. The SA algorithm is often constrained by some mechanism that keeps the iterates in some desired set H, by a projection or other means. The convergence of the θ_n or θ_n^ϵ for constrained versions of the algorithms (1.1) and (1.2) is also well treated in the literature. Convergence with probability one for constrained forms of (1.2) is treated in [21, Chapter 5] under a martingale noise condition. In [21, Chapter 6] more general noise conditions are used. The essential condition is that there be a point $\bar{\theta}$ which is asymptotically stable for the constrained ODE, analogously to what is required for the unconstrained problem. This reference also contains more general results concerning convergence to local minima or chain recurrent points. One can then treat the rate of convergence to such points. A quite general result for probability one convergence for the constrained problem is in [10, Section 8]. That section verifies a basic condition (Assumption 2.2) of [10] for constrained algorithms under a variety of conditions on the noise (those in [10, Section 5], which cover a large proportion of those commonly used). Then under the above-mentioned stability condition, the probability one convergence theorem [10, Theorem 3.1] holds. The classical reference [18, Chapter 5] also contains probability one convergence results for the constrained problem. Additionally, [21, 18] contain extensive results concerning weak convergence for the constrained forms of both (1.1) and (1.2), under very general conditions.

There is virtually nothing available concerning the rate of convergence for the constrained problem. Reference [21] did deal with this rate problem but where the limit point $\bar{\theta}$ was interior to the constraint set H. In this case, the results are the same as for the unconstrained case. When $\bar{\theta}$ is on the boundary of H, then additional problems arise. In this paper, we will give a fairly complete treatment for a large class of systems, when $\bar{\theta} \in \partial H$. We write the ith component of a vector x as x^i. For an $\mathbb{R}^r$-valued function $g(\cdot) = \{g^i(\cdot), i \le r\}$ on $\mathbb{R}^r$, we write $g_\theta(\cdot)$ as the matrix whose ith row is the gradient of $g^i(\cdot)$ with respect to θ.

Unless noted otherwise, we will suppose that the physical constraint set is

$$H = \left\{\theta : 0 \le \theta^i \le b^i, \text{ some subset of components } i\right\}, \tag{1.6}$$

where $0 < b^i \le \infty$. Thus, some components might not be constrained. The constraint is enforced by using an orthogonal projection onto H if the iterate attempts to leave H [21]; i.e., the iterate is returned to the closest point in H. To simplify notation, it will always be assumed that if $\bar{\theta}^i$ is at the end of its allowed interval, then $\bar{\theta}^i = 0$.

By simple affine transformations, coordinate by coordinate, the constraint set includes the cases where the ith components $\theta_n^{i,\epsilon}$ and θ_n^i of the iterates are constrained to lie in some finite interval $[a^i, b^i]$. Keep in mind that, under Assumption 2.4 (below) or its weak convergence counterpart, all that matters is *the shape of the constraint set in a small neighborhood of the limit point* $\bar{\theta}$. This local description will be referred to as L. For example, in the two dimensional case where the physical constraint set is the bounded box $[a_1, b_1] \times [a_2, b_2]$ and $\bar{\theta}^1 = a_1, a_2 < \bar{\theta}^2 < b_2$, the local description about $\bar{\theta}$, after an affine change in each coordinate, is the half plane $L = \{x : x^1 \ge 0, x^2 \text{ unconstrained}\}$.

The constrained form of the algorithms can be written as

$$\theta_{n+1}^{\epsilon} = \theta_n^{\epsilon} + \epsilon Y_n^{\epsilon} + \epsilon Z_n^{\epsilon}, \tag{1.7}$$

$$\theta_{n+1} = \theta_n + \epsilon_n Y_n + \epsilon_n Z_n, \tag{1.8}$$

where ϵZ_n^{ϵ} and $\epsilon_n Z_n$ are the correction terms due to the projection back onto H, if any.

The methods and results follow the general outline used for the case where boundaries are not a factor and (1.5a) and (1.5b) are replaced by a stationary reflected linear diffusion process. There are additional complications in the tightness proofs and in characterizing the mean drift at the limit point. The basic issues are well illustrated by two dimensional problems, with martingale difference noise. Thus for maximal clarity we start with those cases and discuss the extensions afterwards. For appropriate definitions of the matrices A and $\Sigma = \sigma\sigma'$, $\sigma = \{\sigma_{ij}; i, j\}$, the weak sense limit of the $U^n(\cdot)$ will be a stationary solution to a Skorohod problem (reflected diffusion process) of either the form

$$dU = AUdt + \sigma dw + dz, \text{ when } \epsilon_n = \epsilon \text{ or } \epsilon_n \to 0, \text{ under (2.5a)} \tag{1.9a}$$

or of the form

$$dU = \left[A + \frac{I}{2}\right] Udt + \sigma dw + dz, \text{ when } \epsilon_n \to 0, \text{ under (2.5b)}, \tag{1.9b}$$

where $z(\cdot)$ is the reflection term which keeps the values in the correct set. Given $A, \sigma, w(\cdot), U(0)$, and the reflection directions (orthogonal to the boundary faces), there are unique strong sense and adapted solutions $U(\cdot), z(\cdot)$ to (1.9a) and (1.9b), and they are continuous. The process $z^i(\cdot)$ can increase only at t where $U^i(t) = 0$, and it is the minimum such process which forces $U^i(t) \geq 0$ [7]. The precise definition of the Skorohod problem is given below (see (6.5)). The exact forms of the limit equation will be given in section 5 for the various cases.

As is usual in rate of convergence studies, we make an assumption about convergence. In order not to overencumber the development, unless otherwise noted we will work with the form (1.8) with probability one convergence assumed. Much effort was spent in [21] on the weak convergence case as well. The theory of convergence for this case is usually much simpler than the probability one case, particularly when the noise structure is complicated, and it contains virtually the same information (see [21]). Additionally, it is the only type of convergence that can be used with (1.7), and even with (1.8), if the step sizes go to zero slowly enough, or if the noise structure is complex. With probability one convergence, by starting at a large enough time, we can suppose that the iterate is always in an arbitrarily small neighborhood of the limit point, and this facilitates the proofs of the rate results. This is not necessarily true under weak convergence. However, in [21], it is shown that (under weak convergence and with a probability arbitrarily close to unity) the iterate remains in an arbitrarily small neighborhood of the limit point for a long enough time before possibly leaving so that the probability one "localization" technique can still be used. Such a method will work here as well but is omitted due to lack of space and because the adaptation is similar. Details will be found in [5]. Thus the entire theory holds for the "small constant" step size algorithm (1.7).

Section 6 deals with various results in ergodic theory that are needed to complete the proof that the weak sense limit processes are stationary. For simplicity throughout

the paper, the main development assumes martingale difference noise. Section 7 shows the changes that are needed when more general noise is used. An application to a Lagrangian algorithm is given in section 8. Section 9 contains a few comments concerning generalizing the constraint set. Under suitable conditions the basic ideas carry over, but verification of some conditions becomes more complicated. Since one cannot readily compute the stationary covariance matrix for reflected diffusions, even of the simple type which occurs here, combined analysis/simulation was used to get some feeling for the effect of the constraint on the asymptotic variances. A few comments appear in section 10.

The proof of the tightness of $\{U_n\}$ is one of the crucial steps of the development. In general, this requires a special Liapunov function which accounts for the constraint or reflection. Its construction motivated by [11] and the necessary changes in their proof to get the form that we need are in the appendix.

2. Two dimensional problems: Martingale difference noise; assumptions. Let E_n denote the expectation conditioned on $\{\theta_0, Y_i, i < n\}$.

ASSUMPTION 2.1. *There is $\bar{\theta}$ such that $\theta_n \to \bar{\theta}$ with probability one.*

ASSUMPTION 2.2. *There is an $\mathbb{R}^r$-valued "centering" function $g(\cdot)$ on $\mathbb{R}^r$ whose partial derivatives up to second order are continuous in some neighborhood of $\bar{\theta}$, and, for small $\rho > 0$ and some positive definite symmetric matrix $\Sigma = \{\Sigma_{ij}; i, j\}$,* (1.3) *holds and*

$$E_n \xi_n = 0, \tag{2.1}$$

$$E_n \left[\xi_n \xi_n' - \Sigma\right] I_{\{|\theta_n - \bar{\theta}| \le \rho\}} \to 0 \tag{2.2}$$

in the mean as $n \to \infty$.

Write $\Sigma = \sigma\sigma'$, where σ is a square root of Σ.

ASSUMPTION 2.3. *Suppose that, for each small $\rho > 0$,*

$$\left\{|\xi_n|^2 I_{\{|\theta_n - \bar{\theta}| < \rho\}}, n < \infty\right\} \text{ is uniformly integrable,} \tag{2.3}$$

$$\sup_n E_n |\xi_n|^2 I_{\{|\theta_n - \bar{\theta}| < \rho\}} < \infty \text{ with probability 1.} \tag{2.4}$$

ASSUMPTION 2.4. *$\epsilon_n > 0, \epsilon_n \to 0$, $\sum_n \epsilon_n = \infty$, and either*

$$\sqrt{\frac{\epsilon_n}{\epsilon_{n+1}}} = 1 + o(\epsilon_n), \tag{2.5a}$$

or

$$\epsilon_n = 1/n. \tag{2.5b}$$

Even for the martingale difference noise case, [21] uses weaker conditions (second order asymptotic stationarity not required) on the noise for the case where $\bar{\theta} \in H^0$, the interior of H, and those same conditions will work here. Essentially, one needs only that (5.5) converges weakly to a Wiener process. Comments on the correlated noise case are in section 7. Expand

$$g(\theta_n) = g(\bar{\theta}) + g_\theta(\bar{\theta})(\theta_n - \bar{\theta}) + \mu_n, \tag{2.6}$$

where

$$\mu_n = \int_0^1 \left[g_\theta(\bar\theta + s(\theta_n - \bar\theta)) - g_\theta(\bar\theta)\right](\theta_n - \bar\theta)ds.$$

In the classical unconstrained case where $\bar\theta \in H^0$, we must have $g(\bar\theta) = 0$. Then, under (2.5a) and appropriate conditions on the noise, the method in [18, 21] shows that $U^n(\cdot)$ converges weakly to the stationary Gauss–Markov process satisfying (1.5a) and then computes the stationary variance [21]. Under (2.5b), and with $A + I/2$ assumed Hurwitz, the limit is the stationary solution to (1.5b).

It is not necessarily the case that $g(\bar\theta) = 0$ when $\bar\theta \in \partial H$, the boundary of H. Of course, in any case one must have $g^i(\bar\theta) \le 0, i \le r$, since otherwise the limit point cannot be on the boundary. If $g^i(\bar\theta) < 0$, we say either that there is a *forcing term to the boundary* or that *coordinate i has a forcing term to the boundary.* There are several natural divisions of the possibilities, depending on whether there are boundary forcing terms and whether $\bar\theta^i > 0$ for any $i \le r$.

To gain insight into the various issues, in the remainder of this section and in the next, we confine ourselves to the two dimensional problem. In all cases, $b^i > 0$. The general dimensional problem is dealt with in subsequent sections. The first case to be treated is for H having either of the forms $H = \{\theta : b^i \ge \theta^i \ge 0, i \le 2\}$ or $H = \{\theta : b_1 \ge \theta^1 \ge 0\}$ and where

$$g^1(\bar\theta) < 0,\ g^2(\bar\theta) = 0 \text{ and } \bar\theta^1 = 0,\ b_2 > \bar\theta^2 > 0. \tag{2.7}$$

Here $\bar\theta$ is on an open face of H, coordinate 1 has a forcing term to the boundary, and the limit for coordinate 2 is interior to its constraint set. The second case is for $H = \{\theta : b^i \ge \theta^i \ge 0, i \le 2\}$ and

$$g^1(\bar\theta) < 0,\ g^2(\bar\theta) = 0 \text{ and } \bar\theta^1 = \bar\theta^2 = 0. \tag{2.8}$$

Here $\bar\theta$ is in a corner of H, so that both coordinates are on the ends of their constraint sets, and coordinate 1 has a forcing term to the boundary. In the third and fourth cases, defined by (2.9) and (2.10), respectively, there are no forcing terms to the boundary:

$$g(\bar\theta) = 0 \text{ and } \bar\theta^1 = 0,\ \bar\theta^2 > 0, H = \{\theta : b^i \ge \theta^i \ge 0, i \le 2\}, \text{ or } \{\theta : b^1 \ge \theta^1 \ge 0\}. \tag{2.9}$$

$$g(\bar\theta) = 0 \text{ and } \bar\theta = 0, H = \{\theta : \theta^i \ge 0, i \le 2\}. \tag{2.10}$$

The results for the remaining possibilities can be read off from the results for these.

The crucial problem in the proofs of rate of convergence is the proof of tightness (equivalently, boundedness in probability) of the set (1.4b). This will be dealt with in section 3 for the above cases and in section 4 for the general case. Owing to the fact that $\bar\theta \in \partial H$ for all cases, the traditional Liapunov function cannot always be used for the proof. Then, after showing the tightness of (1.4b), section 5 treats the weak convergence of $U^n(\cdot)$ to a solution of (1.9) and characterizes the limit reflection term. The stationarity of the limit is easy to show when $\bar\theta \in H^0$, since the initial time for (1.5) is arbitrary in that we can work with $U^n(-T + \cdot)$ for arbitrary large positive T and use the linearity of (1.5). This is harder for (1.9) and is dealt with in section 6.

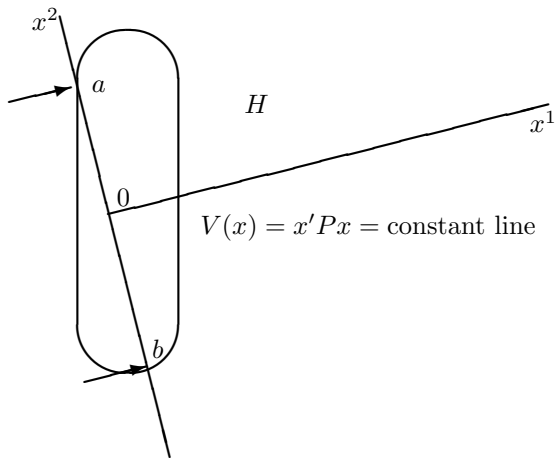


FIG. 3.1. *Projection to b increases* $V(\cdot)$.

3. Tightness of $\{U_n, n < \infty\}$. Part I. In this section, we work with a two dimensional problem, where the Liapunov function $V(U) = |U|^2/2$ can be used. The same methods work in any dimension where this Liapunov function works. This would be the case if the centering function $g(\cdot)$ is the negative of the gradient of a smooth function $f(\cdot)$, which we are minimizing via the SA, and the Hessian of $f(\cdot)$ at $\bar{\theta}$ is positive definite. The problem with a quadratic form Liapunov function that is not of the type $V(U) = |U|^2/2$ is that the reflection or projection onto H will increase its value in part of the state space. This is illustrated in Figure 3.1. Reflection to point b increases $V(\theta)$. The simpler cases of this section more clearly illustrate the role of the forcing term to the boundary. When the Liapunov function $|U|^2/2$ cannot be used, we need to use a Liapunov function which accounts for the boundary behavior (i.e., the reflection or projection). This is much more complicated and is the subject of the next section.

3.1. Cases (2.7), (2.8). Let

$$cg^1(\bar{\theta}) < 0,\ g^2(\bar{\theta}) = 0,\ \bar{\theta}^1 = 0,\ b_2 > \bar{\theta}^2 > 0, \tag{3.1}$$

$$g^2_{\theta^2}(\bar{\theta}) < 0 \text{ under (2.5a)}, \text{ and } g^2_{\theta^2}(\bar{\theta}) < -1/2 \text{ under (2.5b)}.$$

THEOREM 3.1. *Assume that Assumptions* 2.1–2.4 *and* (3.1) *hold. Then* $\{U_n, n < \infty\}$ *is tight. The tightness also holds for the case* (2.8).

The localization method. The proof uses the *localization method* [21], which is defined as follows, and will be used frequently in the subsequent analysis. Since $\theta_n \to \bar{\theta}$ with probability one, for any $\delta > 0, \rho > 0$,

$$P\left\{\sup_{m \geq n} |\theta_m - \bar{\theta}| \geq \rho\right\} \leq \delta$$

for large n. Thus, for the purposes of proving tightness and characterizing the limits of $U^n(\cdot)$ (which involves only the "tail" of the sequence $\{\theta_n\}$), without loss of generality we can modify the process on a set of arbitrarily small measure and reset the time origin so that we can suppose that

$$|\theta_n - \bar{\theta}| \leq \rho \quad \text{for all } n$$

for any desired $\rho > 0$. The value of ρ will usually be chosen small enough so that errors due to linearization are dominated by the linear terms.

Proof. The development will be for the case (2.5a). The proof under (2.5b) differs only in the following expansion: Under (2.5a), $\sqrt{\epsilon_n}/\sqrt{\epsilon_{n+1}} = 1 + o(\epsilon_n)$. Under (2.5b), the ratio equals $1 + \epsilon_n/2 + o(\epsilon_n)$.

The case (2.7). The proof for case (2.7) is essentially classical and uses the Liapunov function $V(U) = U'U/2$. Since $\bar{\theta}^2 > 0$, by the localization argument, we can suppose that $Z_n^2 = 0$. By the hypotheses, $\theta_n^1 - \bar{\theta}^1 \geq 0,\ U_n^1 \geq 0$. Define

$$\tilde{U}_{n+1} = \frac{\sqrt{\epsilon_n}}{\sqrt{\epsilon_{n+1}}} \left[U_n + \sqrt{\epsilon_n} Y_n\right].$$

Then, by centering (1.8) at $\bar{\theta}$ and dividing each side by $\sqrt{\epsilon_{n+1}}$, we can write

$$U_{n+1} = \tilde{U}_{n+1} + \frac{\sqrt{\epsilon_n}}{\sqrt{\epsilon_{n+1}}} \sqrt{\epsilon_n} Z_n. \tag{3.2}$$

Note that $\tilde{U}_{n+1}$ is the normalized value before projection back onto H (or is U_{n+1} if no projection is needed).

Define $\delta_n = \theta_n - \bar{\theta}$ and expand (via a truncated Taylor series)

$$g(\bar{\theta} + \delta_n) - g(\bar{\theta}) = \begin{pmatrix} g_{\theta^1}^1(\bar{\theta})\delta_n^1 + g_{\theta^2}^1(\bar{\theta})\delta_n^2 \\ [g_\theta^2(\bar{\theta})]'\delta_n \end{pmatrix} + y_n\delta_n = g_\theta(\bar{\theta})\delta_n + y_n\delta_n, \tag{3.3}$$

where

$$y_n\delta_n = \int_0^1 \left[g_\theta(\bar{\theta} + s\delta_n) - g_\theta(\bar{\theta})\right] \delta_n ds.$$

Using the above expansion and (2.5a), write

$$\tilde{U}_{n+1} = U_n + \sqrt{\epsilon_n} g(\bar{\theta}) + \epsilon_n \bar{A} U_n + \epsilon_n y_n U_n + \sqrt{\epsilon_n}\xi_n + o(\epsilon_n)\left[g(\theta_n) + \xi_n + U_n\right], \tag{3.4}$$

where $\bar{A} = g_\theta(\bar{\theta})$. By the localization hypothesis and the continuity of $g(\cdot)$ at $\bar{\theta}$, without loss of generality we can suppose that $|\epsilon_n y_n U_n| + o(\epsilon_n)|U_n| \leq c_1\epsilon_n|U_n|$, where c_1 is as small as desired, and that the $g(\theta_n)$ are bounded. The $o(\epsilon_n)$ in (3.4) is due to the expansion $\epsilon_n/\sqrt{\epsilon_{n+1}} = \sqrt{\epsilon_n} + o(\epsilon_n)$, under (2.5a).

Using the localization argument again, and Assumption 2.3, yields

$$E_n|\tilde{U}_{n+1}|^2/2 - |U_n|^2/2 = \sqrt{\epsilon_n} U_n^1 g^1(\bar{\theta}) + \epsilon_n U_n' \left[g_\theta(\bar{\theta}) U_n\right] + o(\epsilon_n)\left|U_n\right|^2 + O(\epsilon_n). \tag{3.5}$$

Since the projection onto H, if any, does not increase the norm $|u|$ defined by the Liapunov function,

$$E_n\left|U_{n+1}\right|^2/2 - \left|U_n\right|^2/2 \leq \text{ right-hand side of (3.5)}. \tag{3.6}$$

By the localization argument, without loss of generality, for any $K < \infty$ we can suppose that

$$\sqrt{\epsilon_n} U_n^1 g^1(\bar{\theta}) \leq -\epsilon_n K [U_n^1]^2. \tag{3.7}$$

The second term on the right-hand side of (3.5) is

$$\epsilon_n \left[[U_n^1]^2 g_{\theta^1}^1(\bar{\theta}) + U_n^1 U_n^2 \left[g_{\theta^1}^2(\bar{\theta}) + g_{\theta^2}^1(\bar{\theta})\right] + [U_n^2]^2 g_{\theta^2}^2(\bar{\theta})\right]. \tag{3.8}$$

The first term of (3.8) is dominated by $\epsilon_n K[U_n^1]^2/8$ (see (3.7)) since K can be supposed to be arbitrarily large. The middle term of (3.8) is dominated in absolute value by $|g^2_{\theta^1}(\bar\theta) + g^1_{\theta^2}(\bar\theta)| = K_1$ times

$$[U_n^1]^2/c + c[U_n^2]^2$$

for any $c > 0$. Choose c small, but so that $K_1/c < K/8$. Write $\gamma = -g^2_{\theta^2}(\bar\theta) > 0$. Then, summarizing, we have

$$E_n\left|U_{n+1}\right|^2/2 - \left|U_n\right|^2/2 \le -\epsilon_n K|U_n^1|^2/2 - \epsilon_n(\gamma-\delta)|U_n^2|^2 + O(\epsilon_n), \tag{3.9}$$

where K is as large and δ is as small as desired. Equation (3.9) implies that, under the localization method, $\sup_n E|U_n|^2 < \infty$, which yields the tightness (see [21, Chapter 10]).

The case (2.8). The main difference between the treatment of (2.7) and (2.8) is that now we can no longer assume that $Z_n^2 = 0$, since $\bar\theta^2 = 0$. However, the projection still does not increase the norm $|u|$ defined by the Liapunov function, so the analysis for (2.7) carries over with no change. □

Note that $g_1(\bar\theta) < 0$ created a "force" pushing U_n^1 to zero (hence the appellation "forcing term to the boundary"). This simplified the analysis, since it dominated the interactions between U_n^1 and U_n^2.

3.2. A simple form of the cases (2.9) and (2.10).

THEOREM 3.2. *Assume that Assumptions* 2.1–2.4 *hold. Let* $A = g_\theta(\bar\theta)$ *be Hurwitz, and suppose that* $A + I/2$ *is Hurwitz, under* (2.5b). *Let* $V(x) = |x|^2$ *be a Liapunov function for the ODE* $\dot x = Ax$, *in that* $A + A'$ *is negative definite. Then* $\{U_n, n < \infty\}$ *is tight for the cases of* (2.9) *and* (2.10).

Comment on the proof. Under the localization method, the right-hand side of (3.5) can be written as

$$\epsilon_n\left[A + A'\right]U_n/2 + o(\epsilon_n)|U_n|^2 + O(\epsilon_n).$$

Since the projection does not increase the norm defined by the Liapunov function, a standard argument of the type used for the unconstrained case [21, Chapter 10] shows the tightness of $\{U_n, n < \infty\}$. □

A comment on the general case of (2.9) or (2.10). Suppose that $A + A'$ is not negative definite. Then, given any positive definite and symmetric C, there is a positive definite symmetric P such that $A'P + PA = -C$. Thus $x'Px$ is a Liapunov function for the ODE $\dot x = Ax$ and could be used to get tightness if $\bar\theta \in H^0$ for the classical unconstrained case. However, for the constrained case, under the norm defined by $x'Px$ (unless P is diagonal), some of the possible projections will not be norm reducing, in the sense that we will not always have $U'_{n+1}PU_{n+1} \le \tilde U'_{n+1}P\tilde U_{n+1}$.[1] See Figure 3.1.

Thus $U'PU$ cannot be used as a Liapunov function for the general constrained problem. When $|U|^2/2$ fails, we need to construct a Liapunov function which takes the projection (equivalently, the boundary behavior) into account. This is another major distinction between the constrained and unconstrained cases. The construction of the needed Liapunov function is given in the appendix and is based on that of [11]. The result is stated and used in the next section. This same Liapunov function will serve as the basis of the stability argument for the more general correlated noise case.

[1] As noted earlier, if $g(\cdot)$ is the negative of the gradient of a smooth function $f(\cdot)$, and the Hessian of $f(\cdot)$ is positive definite at $\bar\theta$, then $|U|^2/2$ can be used.

4. A general Liapunov function. Tightness, part II. *A high dimensional extension of the cases of section* 3. There is one important extension of the results of section 3 to an arbitrary number of dimensions. It is convenient to partition the coordinates to separate out those which have forcing terms to the boundary, i.e., those for which $g^i(\bar{\theta}) < 0$. Thus, suppose that $g^i(\bar{\theta}) = 0, i \leq k$, $g^i(\bar{\theta}) < 0, i = k+1, \ldots, r$. Let $g_a(\cdot), \bar{\theta}_a, U_{a,n}, U_a, \theta_{a,n}$, etc., denote the vectors composed of the first k components. Let $g_b(\cdot)$, etc., denote the vectors composed of the last $r-k$ components. Of course, we could have $k = r$.

ASSUMPTION 4.1. *Define A to be the $k \times k$ matrix whose ith row is the gradient of $g^i(\cdot)$ with respect to θ_a at $\theta = \bar{\theta}$ for $i \leq k$. Suppose that A is Hurwitz under* (2.5a) *and $A + I/2$ is Hurwitz under* (2.5b). *Finally, suppose that $\bar{\theta}^i > 0, i \leq k$, and $\bar{\theta}^i = 0$ for the other coordinates. Let H be the set where $\theta^i \in [0, b^i]$, $i = k+1, \ldots, r$, and either the other coordinates are unconstrained or else $\bar{\theta}^i$ lies in the interior of $[0, b^i]$.*

THEOREM 4.1. *Assume that Assumptions* 2.1–2.4 *and Assumption* 4.1 *hold. Then $\{U_n, n < \infty\}$ is tight.*

Proof. Assumption 4.1 says that the only components of $\bar{\theta}$ which lie on the boundary are those associated with forcing terms to the boundary. In this case, using the localization argument, we can suppose that $\theta_n^i, i > k$, are always within their constraints, hence never projected.

For some positive definite symmetric matrix P, let $\theta_a' P \theta_a$ be a Liapunov function for $\dot{\theta}_a = A\theta_a$, with $PA + A'P = -C$ (or for $\dot{\theta}_a = (A + I/2)\theta_a$, according to the case), where C is positive definite and symmetric. Now with the Liapunov function

$$V(U) = |U_b|^2/2 + U_a{}'PU_a/2,$$

we follow the analysis of Theorem 3.1 to get the desired result. Actually the method is that used in [21], with the addition of the method of Theorem 3.1 to exploit the negativity of the $g^i(\bar{\theta})$ for the i associated with forcing terms to the boundary. □

The general case. We now turn to the general case. First, we state a result (Theorem 4.2) concerning the existence of a Liapunov function for a deterministic Skorohod problem. The construction and proof are based on [11] and are given in the appendix. The state space here is assumed to satisfy the following condition. The integer k is as in Assumption 4.1. As noted in section 1, the set L in Assumption 4.2 is supposed to represent the "local structure" of H about the point $\bar{\theta}$ for those components that do not have forcing terms to the boundary.

ASSUMPTION 4.2. *There is an integer $\nu \leq k$ such that*

$$L = \{x : x^i \geq 0, i \leq \nu\} \subset \mathbb{R}^k.$$

Thus the first ν coordinates are constrained, and the last $k-\nu$ are not. Let ∂L_i denote the faces of L, and n_i the interior normal to ∂L_i. Interior to ∂L_i, the reflection direction is denoted by the unit vector d_i, and $\langle d_i, n_i \rangle > 0$ for each i. The reflection directions on the intersections of the ∂L_i are arbitrary vectors in the convex hull of the directions on the adjoining faces. At each edge and corner, there is a convex combination of the directions on the adjoining faces that points strictly interior to L.

The conditions are more general than we need for our original SA problem, where the reflection directions are orthogonal to the boundary faces, but are useful for the extension in which the reflections are oblique to the boundary faces. In any case, it is the form set up in [11]. For $x \in \mathbb{R}^k$, let $\text{In}(x)$ denote the indices $i \leq \nu$ such that $x^i \leq 0$. For an index set I, let $\text{cone}\{d_i, i \in I\}$ denote the closed infinite cone generated by the linear combinations (with nonnegative coefficients) of the vectors $d_i, i \in I$. Under the

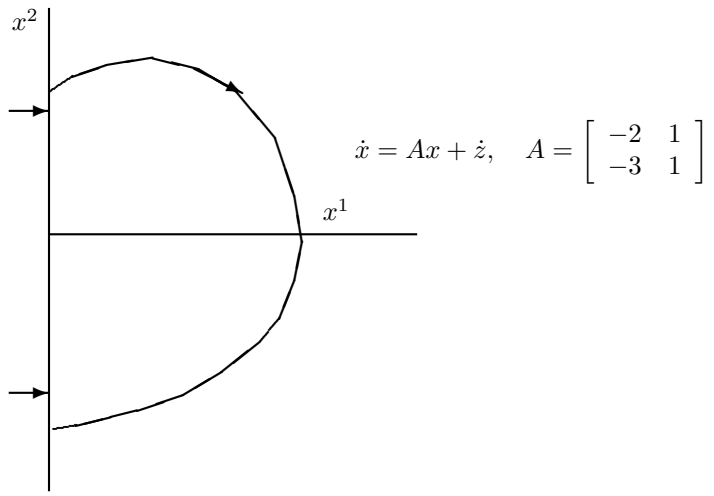


FIG. 4.1. *Flow for Hurwitz A, but nonconvergence to the origin.*

conditions on the reflection directions in Assumption 4.3, there are unique continuous solutions $(x(\cdot), z(\cdot))$ to (4.1a) and (4.1b) for each initial condition [7].

ASSUMPTION 4.3. *Consider the deterministic Skorohod problem for $x \in \mathbb{R}^k$:*

$$dx = Axdt + dz,\ x(t) \in L,\ \textit{under}\ (2.5a), \tag{4.1a}$$

$$dx = \left[A + \frac{I}{2}\right] xdt + dz,\ x(t) \in L,\ \textit{under}\ (2.5b), \tag{4.1b}$$

where $z(\cdot)$ is the reflection term, which can change only when $x(t) \in \partial L$. For almost all t, $\dot{z}(t)$ takes values in the convex cone cone$\{d_i, i \in In(x(t))\}$. For each initial condition $x(0) \in L$, the corresponding solution $x(\cdot)$ converges to zero as $t \to \infty$.

Note on Assumption 4.3 *and the convergence of the SA algorithm.* The key part of the assumption is the convergence of $x(\cdot)$ to zero. This is not necessarily guaranteed (for (4.1a)) even if A is Hurwitz, L an orthant, and the reflection directions normal to the faces. Refer to Figure 4.1 for an illustration of nonconvergence with Hurwitz A. Figure 4.2 illustrates a convergent case.

The hypothesis of convergence Assumption 2.1 would not hold under the situation in Figure 4.1, where A corresponds to the coordinates without forcing terms to the boundary, but it would hold for the case of Figure 4.2. A crucial difference between the unconstrained and constrained cases is that stability of the linearization of the averaged dynamics does not necessarily imply even local convergence of the algorithm.

The proof of the next theorem is in the appendix. For $\epsilon > 0$, let $N_\epsilon(B)$ denote the ϵ-neighborhood of the set $B \in \mathbb{R}^k$.

THEOREM 4.2. *Assume that Assumptions* 4.2 *and* 4.3 *hold. Then there exists a real-valued function $V(\cdot)$ on $\mathbb{R}^k - \{0\}$ with the following properties. It is continuous, together with its partial derivatives up to second order. There is a (twice continuously differentiable) surface ∂S such that any ray from the origin crosses ∂S only once and for a scalar $\alpha > 0$ and $x \in \partial S$, $V(\alpha x) = \alpha V(x)$.*[2] *Thus, the second partial derivatives*

[2]Thus the gradient is the same at all points on any ray from the origin.

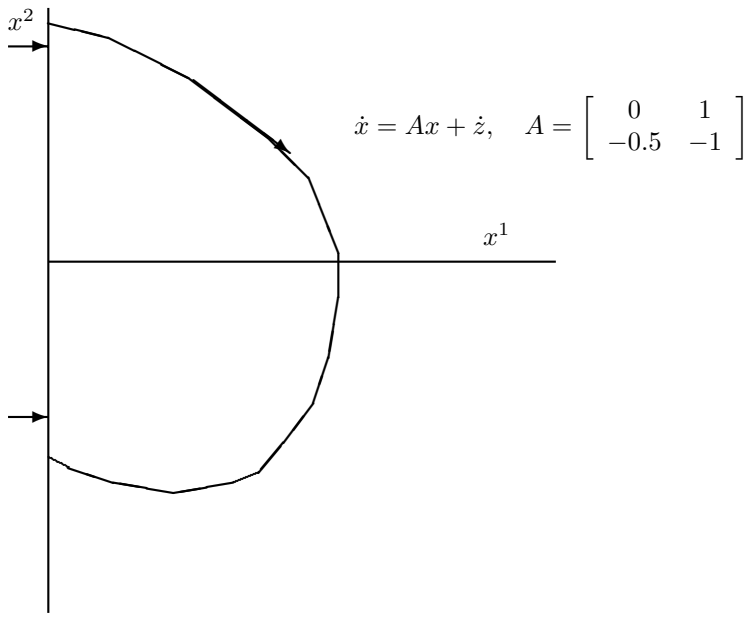


FIG. 4.2. *Flow for Hurwitz A, and convergence to the origin.*

are of the order of $1/|x|$ *as* $|x| \to \infty$ *and*

$$V_v(v)\Big|_{v=\alpha x} = V_x(x).$$

Also, for some real c_i, d_i

$$c_1|x| + c_2 \geq V(x) \geq d_1|x|.$$

There are $c > 0$ *and* $\epsilon > 0$ *such that, for* $x \in \mathbb{R}^k - N_\epsilon(0)$, $V'_x(x)Ax \leq -c|x|$. *For* $x \in \mathbb{R}^k - L - N_\epsilon(0)$, $V'_x(x)d_i \leq -1, i \in In(x)$. *If* $\tilde{x} \in \mathbb{R}^k - L - \{0\}$, *and* x *is the closest point on* ∂L *in a reflection direction in cone*$\{d_i, i \in In(\tilde{x})\}$, *then* $V(x) \leq V(\tilde{x})$. *Define* $V(0) = 0$. *Then* $V(\cdot)$ *is globally Lipschitz continuous.*

The Liapunov function $V(\cdot)$ can be applied to our problem. This will be done in several steps, depending on whether there are forcing terms to the boundary.

No forcing terms to the boundary: $g(\bar{\theta}) = 0$. Here, $k = r$. Some components θ_n^i of θ_n are constrained to the interval $[0, b^i]$, and others are unconstrained. In the proofs of the results of this section, the localization argument will be used without specific mention. Now, $L \in \mathbb{R}^r$ is the intersection of the half spaces corresponding to $x^i \geq 0$ for those i (and only those i) for which θ_n^i is constrained to $[0, b^i]$ and $\bar{\theta}^i = 0$.

THEOREM 4.3. *Let* A *be the matrix whose* i*th row is the gradient of* $g^i(\cdot)$ *with respect to* θ *at* $\theta = \bar{\theta}$. *Assume that* $g(\bar{\theta}) = 0$, *Assumptions* 2.1–2.4 *hold, and Assumption* 4.3 *holds. Let* $d_i = n_i$, *an orthogonal reflection. Then* $\{U_n, n < \infty\}$ *is tight.*

Proof. Only the case (2.5a) will be dealt with. We first work with U_n such that $|U_n|$ is larger than some arbitrarily large K_1, and suppose (without loss of generality) that $E|U_0| < \infty$. Recall the definition of $\tilde{U}_{n+1}$ above (3.2). For the $V(\cdot)$ of Theorem 4.2, we can write

$$E_n V(\tilde{U}_{n+1}) - V(U_n) = \epsilon_n V'_x(U_n) g(\theta_n) + \epsilon_n O(1)\, |V_{xx}(\theta_n)| + O(\epsilon_n). \tag{4.2}$$

By expanding $g(\theta_n)$ as in Theorem 3.1, we dominate the right-hand side by

$$\epsilon_n V_x'(U_n) A U_n + \epsilon_n O(1)\,|V_{xx}(\theta_n)| + O(\epsilon_n).$$

Since the projection is norm decreasing (norm $V(\cdot)$), we have $V(U_{n+1}) \le V(\tilde{U}_{n+1})$. By Theorem 4.2, $V_x'(U_n)AU_n \le -c|U_n|, c > 0$. By this and the linear bounds on $V(\cdot)$ in Theorem 4.2, there are positive q_i such that $V_x'(U_n)AU_n \le -q_1 V(U_n) + q_2$. Putting all of this together, we see that there are $\alpha_i > 0$ such that

$$E_n V(U_{n+1}) \le (1 - \alpha_1\epsilon_n)\, V(U_n) + \alpha_2\epsilon_n. \tag{4.3}$$

Now, consider $|U_n| \le K_1$. Under Assumption 2.3 and the global Lipschitz condition on $V(\cdot)$, there is real K_0 such that $E_n V(U_{n+1}) - V(U_n) \le K_0\sqrt{\epsilon_n}$. Thus, there is real K_2 such that if $|U_n| \le K_1$, then $E_n V(U_{n+1}) \le K_2$. Thus,

$$E_n V(U_{n+1}) \le \max\left\{(1 - \alpha_1\epsilon_n)V(U_n) + \alpha_2\epsilon_n, K_2\right\}.$$

Define $\hat{V}(U) = V(U) - K_2$. Then

$$E_n \hat{V}(U_{n+1}) \le \max\left\{(1 - \alpha_1\epsilon_n)\hat{V}(U_n) + O(\epsilon_n), 0\right\}. \tag{4.4}$$

Equation (4.4) implies that $V(U_n) = O(1)$, which yields the theorem. □

Forcing terms to the boundary. Now, let us consider the general case in which there are forcing terms to the boundary. The procedure is similar to that used in Theorem 3.1. Suppose that $g^i(\bar{\theta}) < 0$ for $i = k+1, \ldots, r$. Partition the coordinates such that $U_n = (U_{a,n}, U_{b,n})$, where $U_{a,n}$ (resp., $U_{b,n}$) consists of the first k (resp., last $r-k$) components of U_n. Partition $\theta, \bar{\theta}$, and $g(\cdot)$ analogously.

THEOREM 4.4. *Assume that $g^i(\bar{\theta}) = 0, i \le k$, and $g^i(\bar{\theta}) < 0, i > k$. Let $d_i = n_i$, an orthogonal reflection. Let A denote the matrix whose ith row is the gradient of $g_a(\cdot)$ with respect to θ_a at $\bar{\theta}$. Assume that Assumptions* 2.1–2.4 *and* 4.3 *hold. Then $\{U_n, n < \infty\}$ is tight.*

Proof. Again, only the case (2.5a) will be dealt with. Since manipulations with Liapunov functions can be tedious, we will present only the main steps. Let $V(\cdot)$ be the Liapunov function of Theorem 4.2, but which is applied to the k dimensional system $dx = Axdt + dz, x \in L$. We will use the expansion

$$g(\theta) - g(\bar{\theta}) = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{pmatrix} \theta_a - \bar{\theta}_a \\ \theta_b - \bar{\theta}_b \end{pmatrix} + \text{higher order terms},$$

where B, C, D are matrices and the higher order terms are

$$\int_0^1 \left[g_\theta(\bar{\theta} + s(\theta - \bar{\theta}) - g_\theta(\bar{\theta})\right](\theta - \bar{\theta})\, ds,$$

as in (3.3).

For some large C_0, which will be determined later, define the Liapunov function $\bar{V}(U) = V^2(U_a)/2 + C_0|U_b|^2/2$. We can write

$$\begin{aligned} E_n\bar{V}(\tilde{U}_{n+1}) - \bar{V}(U_n) \le{}& V(U_{a,n})V_x'(U_{a,n})\left[\epsilon_n A U_{a,n} + \epsilon_n B U_{b,n} + o(\epsilon_n)|U_n|\right] \\ &+ o(\epsilon_n)|U_n| + O(\epsilon_n)|V_x(U_{a,n})|^2 + O(\epsilon_n) + O(\epsilon_n)|V_{xx}(U_{a,n})||V(U_{a,n})| \\ &+ C_0 U_{b,n}'\left[\sqrt{\epsilon_n} g_b(\bar{\theta}) + \epsilon_n C U_{a,n} + \epsilon_n D U_{b,n}\right] + C_0 O(\epsilon_n). \end{aligned} \tag{4.5}$$

First, suppose that $|U_{a,n}| \geq K_1$, large. The value of C_0 is not important at this step. Then bound (see Theorem 3.1) $\sqrt{\epsilon_n} g_b(\bar{\theta}) \leq -\epsilon_n \bar{K} U_{b,n}$, where $\bar{K}$ is a diagonal matrix whose diagonal components are $K > 0$, which (using the localization argument) is as large as desired. Thus we can bound

$$\sqrt{\epsilon_n} C_0 U'_{b,n} g_b(\bar{\theta}) \leq -\epsilon_n C_0 K \left|U_{b,n}\right|^2. \tag{4.6}$$

Following the idea in Theorem 3.1, we trade off by selecting K large enough (and using properties P5 and P7 at the end of the appendix) so that, for some $\alpha > 0$,

$$\begin{aligned} -C_0 K \left|U_{b,n}\right|^2 + C_0 U'_{b,n} C U_{a,n} + C_0 U'_{b,n} D U_{b,n} + V(U_{a,n}) V'_x(U_{a,n}) [A U_{a,n} + B U_{b,n}] \\ \leq -\frac{C_0 K}{2} |U_{b,n}|^2 - \alpha |U_{a,n}|^2 + O(1). \end{aligned} \tag{4.7}$$

As $|U_{a,n}| \to \infty$, $|V_{xx}(U_{a,n})| \to 0$, as $1/|U_{a,n}|$ and $|V_{xx}(U_{a,n})||V(U_{a,n})| = O(1)$.

Using the above bounds, property P6 at the end of the appendix, and the fact that $\bar{V}(U_{n+1}) \leq \bar{V}(\tilde{U}_{n+1})$, we have for some $\alpha_2 > 0$

$$E_n \bar{V}(U_{n+1}) - \bar{V}(U_n) \leq -\epsilon_n \alpha_2 \bar{V}(U_n) + O(\epsilon_n). \tag{4.8}$$

This estimate is for $|U_{a,n}| \geq K_1$, some large number.

Now, consider the case where $|U_{a,n}| \leq K_1$ but $|U_{b,n}| \geq K_1$. As in Theorem 4.3,

$$E_n V^2(U_{a,n+1}) - V^2(U_{a,n}) \leq O(\epsilon_n). \tag{4.9}$$

The difference $|U_{b,n+1}|^2 - |U_{b,n}|^2$ is expanded as in (4.5), and we use the bound in (4.6). Choose C_0 large enough so that $-\epsilon_n C_0 K |U_{b,n}|^2/4 \leq -\epsilon_n |U_{a,n}|^2$. Then, for some $\alpha_3 > 0$ we can write

$$E_n \bar{V}(U_{n+1}) - \bar{V}(U_n) \leq -\epsilon_n \alpha_3 C_0 |U_{b,n}|^2 - \epsilon_n \alpha_3 |U_{a,n}|^2 + O(\epsilon_n). \tag{4.10}$$

Thus, unless both $|U_{a,n}|$ and $|U_{b,n}|$ are less than K_1, (4.10) holds for some $\alpha_3 > 0$. When both are less than K_1, the estimate is obtained as in Theorem 4.3, and then the proof is completed as there. □

5. Weak convergence of the $U^n(\cdot)$ to a reflected diffusion. Let $U^{i,n}(\cdot)$ denote the ith component of $U^n(\cdot)$. The weak convergence uses the Skorohod topology on the path space $D(\mathbb{R}^r; 0, \infty)$, the space of $\mathbb{R}^r$-valued functions on $[0, \infty)$ which are right continuous and have limits from the left. See [3, 13] for the general theory of weak convergence, and [21] for the details of the theory as it applies to the SA problem and for further references.

Now that tightness of the set $\{U_n, n < \infty\}$ has been shown under the various conditions used in section 4, we need to prove the weak convergence of the continuous time interpolations $U^n(\cdot)$ and characterize the weak sense limit process. The localization argument noted at the beginning of Theorem 3.1 will be used where needed, usually without explicit mention. The tightness proof concerned the behavior of the iterates U_n for all time. Given this, to deal with the weak convergence of the $U^n(\cdot)$ we need only work with these processes on an arbitrary finite time interval. The possible unboundedness of U_n is a complication in proving tightness of the set of processes $\{U^n(\cdot), n < \infty\}$ (as opposed to the tightness of the set of random variables $\{U_n, n < \infty\}$). The most convenient approach for proving tightness and characterizing the weak sense limits uses a truncation device, which is discussed in detail in

[17]. It is designed to avoid the problem of unbounded dynamical terms. The idea is to truncate the processes $U^n(\cdot)$, prove the weak convergence of the truncated forms, and then use the properties of the associated weak sense limits of these to show that (asymptotically) the truncation is actually unnecessary.

The truncated processes. To facilitate working with the shifted processes $U^n(\cdot)$ which start at iterate n, for $j \geq 0$ define $\epsilon_j^n = \epsilon_{n+j}, \theta_j^n = \theta_{n+j}, \xi_j^n = \xi_{n+j}$, $Z_j^n = Z_{n+j}$, and $y_j^n = y_{n+j}$. For each integer M, let $q_M(\cdot)$ be a continuous real-valued function on $\mathbb{R}^r$ satisfying $0 \leq q_M(x) \leq 1$, $q_M(x) = 1$ for $|x| \leq M$, and $q_M(x) = 0$ for $|x| \geq M+1$. Let $\bar{A}$ denote the matrix whose ith row is the gradient of $g^i(\cdot)$ with respect to θ at $\bar{\theta}$. Define the truncated iterates $U_j^{n,M}$ by $U_0^{M,n} = U_n$, and for $j \geq 0$ set

$$\begin{aligned} U_{j+1}^{M,n} &= U_j^{M,n} + \sqrt{\epsilon_j^n} g(\bar{\theta}) + \epsilon_j^n \bar{A} U_j^{M,n} q_M(U_j^{M,n}) + \epsilon_j^n y_j^{M,n} U_j^{M,n} q_M(U_j^{M,n}) \\ &\quad + \sqrt{\epsilon_j^n}\xi_j^n + \sqrt{\epsilon_j^n} Z_j^{M,n} + o(\epsilon_j^n)\left[g(\theta_j^n) + U_j^{M,n} q_M(U_j^{M,n}) + \xi_j^n + Z_j^{M,n}\right]. \end{aligned} \tag{5.1}$$

By comparing (3.4) and (5.1), we see that $U_j^{M,n} = U_{n+j}$, until the first index $j > 0$ at which U_{n+j} exceeds M in norm, and that the terms with coefficient $\sqrt{\epsilon_j^n}$ in (5.1) are not truncated. The $Z_j^{M,n}$ are the reflection terms which serve to keep the iterate in H. Define the interpolation $U^{M,n}(\cdot)$ analogously to the way in which the interpolation $U^n(\cdot)$ was defined but using the truncated iterates $U_j^{M,n}$.

No forcing terms to the boundary. We will first work with the case where there is no forcing term to the boundary. Hence $k = r$ and $g(\bar{\theta}) = 0$ in Theorems 4.1 and 4.3. Define the state space $H \subset \mathbb{R}^r$ to be the set of x such that $x^i \in [0, b^i]$ for some subset of coordinates, with the others unconstrained. To get L, center about $\bar{\theta}$ and use "local" coordinates. Thus $L \subset \mathbb{R}^r$ is the set of points x for which $x^i \geq 0$ for all coordinates i that are constrained and such that $\bar{\theta}^i = 0$.

THEOREM 5.1. *Assume that Assumptions* 2.1–2.4 *and* 4.3 *hold. Suppose that there are no forcing terms to the boundary and define L and H as above. Then $U^n(\cdot)$ is tight, and the weak limit of any weakly convergent subsequence satisfies* (1.9a) *or* (1.9b)*, according to the case, where σ is defined below Assumption* 2.2 *and $U(t) \in L$. Thus, the weak sense limits differ only in the initial condition. The processes $(U(\cdot), z(\cdot))$ are nonanticipative with respect to the Wiener process.*

Proof. Again, we work with (2.5a) only. Tightness of the set of initial conditions $U_n = U^n(0)$ was proved in Theorems 4.1 and 4.3. Given this tightness, one first proves tightness of the set of truncated random processes $\{U^{M,n}(\cdot), n < \infty\}$ and then characterizes the limit of any weakly convergent subsequence. Abusing terminology, we suppose that the chosen subsequence is indexed by n also. The result will not depend on the chosen subsequence.

It will be shown that the weak sense limit $U^M(\cdot)$ of the truncated processes $U^{M,n}(\cdot)$ satisfies

$$dU^M = AU^M q_M(U^M)dt + \sigma dw + dz^M, \tag{5.2}$$

where $A = \bar{A}$ and $w(\cdot)$ is a standard vector-valued Wiener process. By the tightness of $\{U_n, n < \infty\}$, the set of all possible $U^M(0)$ (over all convergent subsequences and all M) is tight. Then, the process (5.2) has the property that, for any $T > 0$, and where the limit (as $C \to \infty$) is taken on uniformly in the chosen subsequence,

$$\lim_{C\to\infty} \sup_M P\left\{\sup_{t\leq T} \left|U^M(t)\right| \geq C\right\} = 0. \tag{5.3}$$

(Equation (5.3) is obtained using the Lipschitz condition on the map taking the driving process $w(\cdot)$ to the reflection process for the Skorohod problem, which was proved in [7].) Then (5.2), together with the weak convergence of $U^{M,n}(\cdot)$ to $U^M(\cdot)$, implies that the original sequence $U^n(\cdot)$ satisfies

$$\lim_{C\to\infty} \sup_n P\left\{\sup_{t\le T}|U^n(t)| \ge C\right\} = 0. \tag{5.4}$$

This boundedness in probability, in turn, together with the tightness of $\{U^{M,n}(\cdot), n < \infty\}$ for each M, implies that the untruncated sequence is also tight and satisfies (1.9a). The weak sense limits can differ only in the initial condition.

Now, return to the proof of (5.2). Fix M until further notice. Using the localization method, we will suppose without loss of generality that $|\theta_n - \bar\theta|$ is as small as desired. Hence the $Z_j^{i,M,n}$ (the ith component of $Z_j^{M,n}$) can be supposed to equal zero for those i for which $\bar\theta^i \ne 0$. Define the process

$$W^n(t) = \sum_{i=n}^{m(t_n+t)-1} \sqrt{\epsilon_i}\xi_i. \tag{5.5}$$

Then, $W^n(\cdot)$ converges weakly to a Wiener process with covariance matrix Σ [21, Theorem 10.2.1]. Now, keep in mind that $|Uq_M(U)|$ is bounded by M. Hence, the processes defined by the interpolation of the fourth term on the right-hand side of (5.1) clearly converge weakly to the "zero" process. The processes defined by the interpolations of the first two terms in the bracket on the right-hand side of (5.1) also converge weakly to the "zero" process. Recall that the state space of concern is $L = \{x : x^j \ge 0, \text{ for } j \text{ such that } \bar\theta^j = 0 \text{ and coordinate } j \text{ is constrained}\}$.

It remains to treat the effects of $Z_j^{i,n,M}$ for those coordinates i which are constrained and for which $\bar\theta^i = 0$. We next outline the proof that $Z^{M,n}(\cdot)$ is "asymptotically continuous." Adapting a method of [19], it will be shown that for each $T > 0$ and $\nu > 0$

$$\lim_{\delta\to 0} \limsup_n P\left\{\sup_{t\le T}\sup_{s\le\delta}\left|Z^{n,M}(t+s) - Z^{n,M}(t)\right| \ge \nu\right\} = 0, \tag{5.6}$$

where $Z^{n,M}(\cdot)$ is the continuous time interpolation of the $Z_i^{n,M}$. The limit equality (5.6) implies the tightness and that each weak sense limit process has continuous paths with probability one. We refer to the property (5.6) as "asymptotic continuity."

It is clear that (5.6) holds for the processes defined by the interpolations of any of the terms on the right-hand side of (5.1), except possibly that due to the Z-term. Suppose that (5.6) does not hold for some $T > 0$ and $\nu > 0$. Then there are $0 < \delta_n \to 0$ and $\eta > 0$ such that

$$\limsup_n P\left\{\sup_{t\le T}\left|Z^{M,n}(t+\delta_n) - Z^{M,n}(t)\right| \ge \nu\right\} \ge \eta. \tag{5.7}$$

By the truncation and the convergence of the process defined by the sum of the noise terms to a Wiener process, for any $T > 0$, the maximum value of the (non-Z) terms on the right-hand side of (5.1) goes to zero as $n \to \infty$, where the maximum is taken over $i \le m(t_n + T) - n$.

Consider only coordinates i which are constrained and are such that $\bar\theta^i = 0$. The $Z_j^{i,M,n}$ can be nonzero only for those j such that the unprojected iterate $\tilde U_{j+1}^{i,M,n}$ is

negative, and then $\epsilon_j Z_j^{i,M,n}$ is the minimum value which keeps $U_{j+1}^{i,M,n} \geq 0$. By the comments in the last paragraph,

$$\sup_{i \leq m(t_n+T)-n} \sqrt{\epsilon_i^n} Z_j^{i,M,n} \to 0$$

in probability as $n \to \infty$, and (with an arbitrarily high probability) $Z_j^{i,M,n}$ can be nonzero only when $U_j^{i,M,n}$ is arbitrarily close to zero. These facts imply that (5.7) is not possible. Hence (5.6) holds.

Now, take a weakly convergent subsequence of all of the interpolated processes (indexed also by n). By what has been said, the limit satisfies (5.2), where $z^M(\cdot)$ is the reflection term. The proof (except for the nonanticipativity) is now concluded as discussed in the second paragraph of the proof.

It only remains to show the nonanticipativity. This also follows a standard procedure. Let $\mathcal{F}_t$ denote the filtration engendered by $\{U(s), z(s), w(s), s \leq t\}$, $h(\cdot)$ be a bounded real-valued function of its arguments, and $f(\cdot)$ be a twice continuously differentiable real-valued function on $\mathbb{R}^r$ with compact support. For arbitrary $p, t, \tau, s_i \leq t < t+\tau, i \leq p$, suppose that

$$\begin{aligned} &Eh(U(s_i), z(s_i), w(s_i), i \leq p) \\ &\quad \times \left[f(w(t+\tau)) - f(w(t)) - \frac{1}{2} \int_t^{t+\tau} \sum_{i,j} f_{w^i w^j}(w(s)) ds \right] = 0. \end{aligned} \tag{5.8}$$

Then $w(\cdot)$ is an $\mathcal{F}_t$-martingale and an $\mathcal{F}_t$-Wiener process. To prove (5.8), one shows that

$$\begin{aligned} &Eh(U^n(s_i), Z^n(s_i), w^n(s_i), i \leq p) \\ &\quad \times \left[f(w^n(t+\tau)) - f(w^n(t)) - \frac{1}{2} \int_t^{t+\tau} \sum_{i,j} f_{w^i w^j}(w^n(s)) \Sigma_{ij} ds \right] \to 0 \end{aligned}$$

and uses the weak convergence. Details for many such computations can be found in [17] and are omitted. □

Forcing terms to the boundary. Define H as for Theorem 5.1. Since $U_b^n(\cdot)$ will be shown to converge weakly to the "zero" process, L is defined only for the remaining coordinates. It is the state space of the limit process $U_a(t)$.

THEOREM 5.2. *Suppose that there are forcing terms to the boundary, in that $k < r$ in Theorem* 4.1 *or* 4.4. *Redefine A to be the matrix whose ith row is the gradient of $g_a^i(\cdot)$ with respect to θ_a, at $\bar{\theta}$. Assume that Assumptions* 2.1–2.4 *and* 4.3 *hold. Then $U^n(\cdot)$ is tight. The sequence $U_b^n(\cdot)$ converges weakly to the "zero" process. The limit of any weakly convergent subsequence of $U_a^n(\cdot)$ satisfies*

$$dU_a = AU_a dt + \sigma_a dw_a + dz_a, \text{ under (2.5a)}, \tag{5.9a}$$

or

$$dU_a = \left[A + \frac{I}{2} \right] U_a dt + \sigma_a dw_a + dz_a \text{ when } \epsilon_n \to 0, \text{ under (2.5b)}. \tag{5.9b}$$

In (5.9), $U_a(t) \in L$, σ_a *is the square root of the upper left hand $k \times k$ submatrix of Σ, $w_a(\cdot)$ is a standard $\mathbb{R}^k$-valued Wiener process, and $z_a(\cdot)$ is the reflection term.*

Thus the weak sense limits differ only in the initial condition. The $(U_a(\cdot), z_a(\cdot))$ *is nonanticipative with respect to* $w_a(\cdot)$.

Proof. The proof is similar to that of Theorem 5.1. In view of the proof of Theorem 5.1 and the tightness of $\{U_n, n < \infty\}$, we need only show that $U_b^n(\cdot)$ converges weakly to the "zero" process. The truncation method will be used again. Thus we need only show that $U_b^{M,n}(\cdot)$ converges weakly to the "zero" process for each M.

Return to (5.1), and using the fact of the truncation write the b-component as

$$U_{b,j+1}^{M,n} = U_{b,j}^{M,n} + \sqrt{\epsilon_j^n} g_b(\bar{\theta}) + \epsilon_j^n O(1) + \sqrt{\epsilon_j^n}\xi_{b,j}^n + \sqrt{\epsilon_j^n} Z_{b,j}^{M,n} + o(\epsilon_j^n)\left[g_b(\theta_j^n) + O(1) + \xi_{b,j}^n + Z_{b,j}^{M,n}\right]. \tag{5.10}$$

First note that the continuous time interpolation of the components of $\sqrt{\epsilon_j^n} g_b(\bar{\theta})$ converges to $-\infty$. It is then apparent from the weak convergence of the interpolation of the noise terms to a Wiener process and the boundedness of the coefficients of ϵ_n and of $o(\epsilon_n)$ that $U_b^{n,M}(\cdot)$ converges weakly to the zero process. Given this, use the proof in Theorem 5.1 for $U_a^{n,M}(\cdot)$. The $U_b^{n,M}(\cdot)$ does not appear in the limit. □

6. Stationarity of the limit solution.

6.1. Convergence to invariant measures: General results. The proof of the stationarity of the limit processes (1.9) or (5.9) rests on the following ergodic results.

ASSUMPTION 6.1. *Let* $L \subset \mathbb{R}^k$ *be the intersection of a finite number of half-planes and the closure of its interior. Let* $x(\cdot)$ *be a Markov–Feller process with a stationary transition function and with paths in* $D(L; 0, \infty)$. *Let* $P(x, t, \cdot)$ *denote the measure of* $x(t)$, *given that* $x(0) = x$, *and let it be mutually absolutely continuous with respect to Lebesgue measure for each* $t > 0$ *and each* $x \in L$.

Recall that the process $x(\cdot)$ with values in L is a *Feller* process if for each bounded continuous real-valued function $f(\cdot)$ on L and each $t > 0$,

$$\int f(y) P(x, t, dy) \text{ is continuous in } x. \tag{6.1}$$

Also, a Markov–Feller process is a strong Markov process. The process is said to be *strong Feller* if the continuity holds for any bounded and measurable $f(\cdot)$.

Comment. The crucial result (6.2) is implied by Theorem 4 of [6] (with modified notation). The rest of the proof is straightforward and the details are omitted. More detail is in [5].

THEOREM 6.1. *Assume that Assumption* 6.1 *holds. Let* $\mu(\cdot)$ *be an invariant probability measure for* $x(\cdot)$, *which we suppose exists. Then,* $\mu(\cdot)$ *is mutually absolutely continuous with respect to Lebesgue measure. Also,*

$$P(x, t, E) \to \mu(E) \text{ for each } x \in L \text{ and Borel set } E \subset L. \tag{6.2}$$

Hence $\mu(\cdot)$ *is the unique invariant probability measure. Furthermore,* (i) $x(\cdot)$ *is a strong Feller process;* (ii) *for any nonempty compact set* $C \subset L$ *and* $t > 0$, $\{P(x, t, \cdot), x \in C\}$ *is uniformly absolutely continuous with respect to Lebesgue measure* $l(\cdot)$ *on* L, *in that for each* $\epsilon > 0$ *there is* $\delta > 0$ *such that* $l(E) \le \delta$ *implies that* $P(x, t, E) \le \epsilon, x \in C$. *Assume, in addition, that for each* $\epsilon > 0$ *and compact set* $C_1 \subset L$

$$\lim_{t \to 0} \sup_{x \in C_1} P_x \{|x(t) - x| \ge \epsilon\} = 0. \tag{6.3}$$

Then (iii) *the uniformity in the absolute continuity holds for* $(x,t) \in C \times [t_0, t_1]$ *for any* t_0, t_1 *satisfying* $0 < t_0 < t_1 < \infty$.

Now, assume (6.1) *in addition. Let* $\mu(\cdot)$ *be an invariant measure for* $x(\cdot)$. *For* α *in some index set, let* $\{q_\alpha(\cdot)\}$ *be a tight family of probability measures on* L. *Then for each Borel set* $E \subset L$

$$\int q_\alpha(dx) P(x,t,E) \to \mu(E) \tag{6.4}$$

uniformly in α *as* $t \to \infty$.

6.2. Ergodic results for the Skorohod problem. We will be concerned with the solution to the Skorohod problem in a state space L satisfying Assumption 4.2 and the following conditions. The condition is more general than needed, but it is the natural formulation.

ASSUMPTION 6.2. *There are constants* $a_i > 0$ *such that for all* i

$$a_i \langle n_i, d_i \rangle > \sum_{j \neq i} a_j \left| \langle n_i, d_j \rangle \right|.$$

The condition needs to hold only for the faces adjoining each individual edge or corner of L *separately.*

ASSUMPTION 6.3. $\sigma(\cdot)$ *is an* $r \times r$ *matrix, bounded and Hölder continuous, with* $a(x) = \sigma(x)\sigma'(x)$ *positive definite, uniformly in* x. *Let* $b(\cdot)$ *be Hölder continuous and* $|b(x)| \leq K(1 + |x|)$ *for some* $K < \infty$.

These conditions hold for the case of interest for the SA in this paper. Indeed, for our cases, the reflection directions are normal to the boundary ($d_i = n_i$), and L is the set used in section 4 where some components are unconstrained and others are constrained to be nonnegative. Also for our case, $b(x) = Ax$, and $\sigma(x) = \sigma$ is constant.

Let $r(x)$ denote the set of reflection directions at $x \in L$. The Skorohod problem of interest is defined by

$$dx = b(x)dt + \sigma(x)dw + dz,\ x(t) \in L \text{ for all } t, \tag{6.5}$$

where $z(\cdot)$ is the reflection term and $w(\cdot)$ is a standard vector-valued Wiener process. Let $|z|(t)$ denote the variation of $z(\cdot)$ on $[0,t]$. Suppose (as holds for our case) that $b(\cdot)$ and $\sigma(\cdot)$ are Lipschitz continuous. Then [7] $|z|(t) < \infty$ with probability one for each t, and

$$|z|(t) = \int_0^t I_{\{x(s) \in \partial L\}} d|z|(s),$$

$$z(t) = \int_0^t \gamma(s) d|z|(s),$$

where $\gamma(\cdot)$ is a measurable function such that, for almost all (ω, t), $\gamma(t) \in r(x(t))$. It is a consequence of the results in [7] that a strong sense solution to (6.5) exists on $[0, \infty)$ and is unique in the strong sense for each initial condition $x(0) = x \in L$. (Thus $x(\cdot)$ is a well-defined Markov–Feller process with a transition function $P(x,t,\cdot)$.) Also, by Assumption 4.2 the variation of $z(\cdot)$ on any finite time interval is bounded in probability, uniformly in the initial condition in any bounded set [19, Theorem 11.1.1.]. Consider the Skorohod problem $x(t) = \phi(t) + z(t)$, where $\phi(\cdot) \in$

$D(\mathbb{R}^k; 0, \infty)$ and the reflection directions and state space satisfy Assumption 4.2. Let (x_1, ϕ_1, z_1) and (x_2, ϕ_2, z_2) be any solutions of the corresponding Skorohod problem. Then Assumption 6.2 implies that [7] there is a real K_1 such that for all t

$$\sup_{s \le t} [|z_1(s) - z_2(s)| + |x_1(s) - x_2(s)|] \le K_1 \sup_{s \le t} |\phi_1(s) - \phi_2(s)|.$$

This Lipschitz condition can be used to prove strong sense existence and uniqueness of the solution to (6.5) under a Lipschitz condition on $b(\cdot), \sigma(\cdot)$. This is its only purpose here.

We now restate and slightly extend (with essentially the same proof) the material from Harrison and Williams [15, section 7]. The proof of the next theorem is just that of [15, Lemma 7.2], nearly word for word, and the details are omitted. Theorems 6.3 and 6.4 together imply uniform mutual absolute continuity of the transition function and Lebesgue measure which is needed to apply Theorem 6.1.

THEOREM 6.2. *Assume that Assumptions* 4.2 *and* 6.2–6.3 *hold. Suppose that the nonanticipative solution to* (6.5) *exists and is unique in the weak sense for each initial condition. Then*

$$E_x \int_0^\infty I_{\partial L}(x(s)) ds = 0 \tag{6.6}$$

for each initial value $x \in L$. Furthermore,

$$P(x, t, \partial L) = 0, \ t > 0, \ \text{and all } x \in L. \tag{6.7}$$

More strongly, for any compact set $C \subset L$ and $t > 0$,

$$\lim_{\delta \to 0} \sup_{x \in C} P(x, t, N_\delta(\partial L)) = 0, \tag{6.8}$$

where $N_\delta(\partial L)$ is a δ-neighborhood of the boundary.

The next theorem essentially follows from Theorems 6.1 and 6.2, the properties of the unconstrained process from [12], and the argument in [15, Lemma 7.9]; the details are omitted. Theorem 6.4 is a consequence of Theorems 6.1 and 6.3. See [5] for more detail.

THEOREM 6.3. *Assume that Assumptions* 4.2 *and* 6.2–6.3 *hold. Suppose that the nonanticipative solution to* (6.5) *exists and is unique in the weak sense for each initial condition. Then for $t > 0$ the transition probability $P(x, t, \cdot)$ is mutually absolutely continuous with respect to Lebesgue measure for each $x \in L$ and is absolutely continuous with respect to Lebesgue measure uniformly for (x, t) in any compact set of the form $C \times [t_0, t_1]$, where $C \subset L$ is compact and $t_1 > t_0 > 0$.*

THEOREM 6.4. *Assume that Assumptions* 4.2 *and* 6.2–6.3 *hold. Suppose that the nonanticipative solution to* (6.5) *exists and is unique in the weak sense for each initial condition. Then $x(\cdot)$ is a strong Feller process. Let the sets E below be contained in a compact set $C_1 \subset L$, and suppose that $C \subset C_1^0$ is compact. Let $\epsilon > 0$ be arbitrary. Then there is $\delta > 0$ such that, for $t_1 > t_0 > 0$,*

$$\inf_{t_1 \ge t \ge t_0} \inf_{x \in C} \inf_{\{E : l(E) \ge \epsilon\}} P(x, t, E) \ge \delta. \tag{6.9}$$

6.3. Stationarity of the limit $U(\cdot)$ process.

THEOREM 6.5. *Under the conditions of Theorem* 5.1*, any weak sense limit* $U(\cdot)$ *of* $U^n(\cdot)$ *is stationary and the stationary process is unique. The analogous result holds for the weak sense limit* $U_a(\cdot)$ *under the conditions of Theorem* 5.2.

Outline of proof. We apply Theorem 6.1 and work with the case of Theorem 5.1 only (the case of Theorem 5.2 is treated identically). Owing to the tightness of $\{U_n, n < \infty\}$, the set

$$\{U(t), t < \infty, \text{all possible weak sense limits } U(\cdot)\}$$

is tight. This and the Markov–Feller property of $U(\cdot)$ imply that there is an invariant measure. Theorem 6.3 ensures the absolute continuity property (with respect to Lebesgue measure) needed in Theorem 6.1. Let $\mu(\cdot)$ denote the unique invariant measure, and $P(u, t, \cdot)$ the transition function for the process $U(\cdot)$. Let $\{\pi_\alpha(\cdot)\}$ denote a tight set of probability measures. The next step is to show that

$$\int \pi_\alpha(du) P(u, t, A) \to \mu(A)$$

as $t \to \infty$, uniformly in α. However, this is implied by Theorem 6.1.

To complete the proof, we use a "time shift" argument analogous to what was done in [21, Chapter 10]. Define $U_n = U_0$ for $n \leq 0$. We start the $U^n(\cdot)$ processes "earlier." Let $s > 0$ be an integer. Define the process $U^{s,n}(\cdot)$ by $U^{s,n}(0) = U_{m(t_n - s)}$ and, for $t > 0$, $U^{s,n}(t) = U_{m(t_n + t - s)}$. Thus $U^n(0) = U^{s,n}(s)$. The set $\{U^n(\cdot), U^{s,n}(\cdot)\}$ is tight for each fixed s. Take a subsequence n_i such that $(U^{n_i}(\cdot), U^{s,n_i}(\cdot))$ converges weakly to weak sense limits denoted by $(U(\cdot), U^s(\cdot))$. The processes $U(\cdot)$ and $U^s(\cdot)$ satisfy (1.9a) or (1.9b), according to the case. Also, $U(0) = U^s(s)$. As $s \to \infty$, take further weakly convergent subsequences. Owing to the tightness of $\{U^s(0), s > 0\}$ and (6.4), we see that $U(0)$ must be the stationary initial condition. □

7. More general noise. The tightness results of sections 3 and 4 depended on the assumption that the ξ_n were martingale differences. The proof of weak convergence in section 5 did not need the martingale difference condition. It required only that the process defined by (5.5) converge weakly to a Wiener process. (This, in itself, will not get the nonanticipativity, but the required addition is minor and is dealt with in [21].) Still, the key issue is the tightness of $\{U_n, n < \infty\}$. For a large class of correlated noise processes, this can be proved by use of the perturbed Liapunov function method [4, 17, 21]. We will make a few comments concerning the method for the case of Theorem 4.3. The cases of Theorems 4.1 and 4.4 are similar.

The main problem with correlated noise is that we no longer have $E_n V_x'(U_n)\xi_n = 0$ and $E_n|\xi_n|^2$ uniformly bounded when $\theta_n - \bar{\theta}$ is small. To simplify the development, work with (2.5a) and suppose that ξ_n is bounded. Then, the term $[\sqrt{\epsilon_n} + o(\epsilon_n)]\xi_n = \epsilon_n \xi_n / \sqrt{\epsilon_{n+1}}$ in (3.4) leads to the additional term $\sqrt{\epsilon_n} V_x'(U_n) E_n \xi_n$ in (4.1). Following the perturbed Liapunov function method used in [17, 21], define the perturbation

$$\delta F^n = \sum_{j=n}^{\infty} E_n \xi_j. \tag{7.1}$$

Assume that it is well defined and that the conditional expectations go to zero fast enough so that it is bounded by some constant B, uniformly in all variables. Now, define the Liapunov function perturbation $\delta V^n = [\epsilon_n / \sqrt{\epsilon_{n+1}}] V_x'(U_n) \delta F_n$. We will

use the perturbed Liapunov function $V^n = V(U_n) + \delta V^n$. Modifying the method in Theorem 4.3 by following the procedures for the perturbed Liapunov functions in [17, 21], one can prove the tightness of $\{U_n, n < \infty\}$.

8. An application: Constrained stochastic approximation via a Lagrangian method. Consider the problem of minimizing the real-valued function $f(x)$ on $\mathbb{R}^r$ subject to constraints $q^i(x) \leq 0, i \leq p$. Both $f(x)$ and the $q^i(x)$ are unknown but are observed with additive noise. A "Lagrangian" SA method for dealing with this problem was introduced in [18, 20], where convergence (with probability one) was proved.

Suppose that $f(\cdot)$ is strictly convex and twice continuously differentiable, and that the $q^i(\cdot)$ are twice continuously differentiable and convex. To simplify the discussion, suppose that we observe the derivatives plus (martingale difference) noise. Otherwise, the Kiefer–Wolfowitz procedure is used, and the finite difference bias error needs to be accounted for. Define the Lagrangian

$$L(x, \lambda) = f(x) + \sum_i \lambda^i q^i(x), \qquad \lambda^i \geq 0.$$

There is a unique saddle point $(\bar{X}, \bar{\lambda})$ and $\bar{\lambda}^i \geq 0$. Suppose that finite a^i, b^i, c^i are known for which $\bar{\lambda}^i < a^i$ and $b^i < \bar{X}^i < c^i$. The SA algorithm is [18, 20]

$$X_{n+1} = X_n - \epsilon_n L_x(X_n, \lambda_n) + \epsilon_n \xi_{x,n} + \epsilon_n Z_{x,n},$$

$$\lambda_{n+1} = \lambda_n + \epsilon_n L_\lambda(X_n, \lambda_n) + \epsilon_n \xi_{\lambda,n} + \epsilon_n Z_{\lambda,n},$$

where the $Z_{x,n}, Z_{\lambda,n}$ are reflection terms, which keep the iterate within the hard boundary, $0 \leq \lambda^i \leq a^i$, $b^i \leq x^i \leq c^i$. The $\xi_{x,n}, \xi_{\lambda,n}$ are observation noises. Define $\theta = (X, \lambda)$, and suppose that $\xi_n = (\xi_{x,n}, \xi_{\lambda,n})$ satisfies Assumptions 2.2 and 2.3. For specificity, suppose that (2.5a) holds. Then [18] θ_n converges weakly to its unique limit $\bar{\theta} = (\bar{X}, \bar{\lambda})$. Under stronger conditions on the $\epsilon_n \xi_n$, there is probability one convergence [18]. Let us suppose probability one convergence. Under our assumptions, $\bar{X}$ is inside its hard constraint box and $\bar{\lambda}^i < a^i$. Then, by the localization hypothesis, we can suppose that $Z_{x,n} = 0$. Also, we can suppose that $Z^i_{\lambda,n} = 0$ for all i such that $\bar{\lambda}^i > 0$.

Define

$$U_{x,n} = \frac{X_n - \bar{X}}{\sqrt{\epsilon_n}}, \; U_{\lambda,n} = \frac{\lambda_n - \bar{\lambda}}{\sqrt{\epsilon_n}}.$$

If $0 < \bar{\lambda}^i$ for all i, then $\bar{\theta}$ is interior to its hard constraint set, and the classical rate of convergence theory can be used. Thus, suppose that $\bar{\lambda}^i = 0$ for some i. Then the classical theory cannot be used. If $\bar{\lambda}^i = 0$, then we could have either $q^i(\bar{X}) < 0$ or $q^i(\bar{X}) = 0$.

Let us next deal with the latter case, where $q^i(\bar{X}) = 0$ for all i. Define the matrix

$$A = \begin{bmatrix} -L_{xx}(\bar{X}, \bar{\lambda}) & -q'_x(\bar{X}) \\ q_x(\bar{X}) & 0 \end{bmatrix}.$$

Note that $L_{xx}(\bar{X}, \bar{\lambda})$ is positive definite. Suppose that the vectors $q_{i,x}(\bar{X}), i \leq p$, are linearly independent. Consider the deterministic Skorohod problem

$$\begin{pmatrix} \dot{x} \\ \dot{\lambda} \end{pmatrix} = A \begin{pmatrix} x \\ \lambda \end{pmatrix} + \begin{pmatrix} 0 \\ \dot{z}_\lambda \end{pmatrix}.$$

The solution corresponding to any initial condition goes to zero as $t \to \infty$. This can be seen from the following Liapunov function argument. Use $V(\theta) = |x|^2/2 + |\lambda|^2/2$. Then, using the fact that $\lambda' \dot{z}_\lambda = 0$, we have

$$\dot{V}(\theta) = -x' L_{xx}(x) x.$$

This implies that $V(\theta(t))$ is bounded for each initial condition and that the solution goes to the set where $x = 0$. But this, in turn (using the linear independence of the $q_x^i(\bar{X}), i \le p$), implies that the solution $\theta(t)$ tends to zero.

With this result in hand, Theorem 4.3 and the results of sections 5 and 6 yield that $U^n(\cdot)$ converges weakly to the stationary solution to

$$\begin{pmatrix} dU_x \\ dU_\lambda \end{pmatrix} = A \begin{pmatrix} U_x \\ U_\lambda \end{pmatrix} dt + \sigma dw(t) + \begin{pmatrix} 0 \\ dz_\lambda \end{pmatrix},$$

where $\dot{z}_\lambda^i(t) = 0$ unless $\bar{\lambda}^i = 0$.

The case where $q^i(\bar{X}) < 0$ for some i corresponds to a forcing term to the boundary for the component λ_n^i, and the process defined by $U_{\lambda,n}^i$ will converge to zero. Then such $U_{\lambda,n}^i$ can be dropped, and the rate of convergence equation can be developed for the remaining components.

9. Nonorthogonal reflection directions. The convergence results for constrained algorithms in [21] are for reflections which return the iterate to the closest point in the hyperrectangular constraint set. However, the proofs can be modified to allow constraint sets which are just the closures of their interiors and have piecewise linear boundaries. Suppose that the constraint set H and the reflection directions simply satisfy Assumption 4.2 and that Assumption 4.3 holds. Then Theorem 4.2 holds, so one can construct the Liapunov function $V(\cdot)$.

Suppose that there are no forcing terms to the boundary. Then one can repeat all of the steps in Theorem 4.3. With the tightness of $\{U_n, n < \infty\}$ given, the only new problem is the proof of the tightness of the reflection processes $Z^{M,n}(\cdot)$ in Theorem 5.1. This can be done by essentially the method in Theorem 5.1, since the condition on the reflection directions in Assumption 4.2 can be used to show asymptotic continuity of these processes. While there is no analytical problem when there are no forcing terms to the boundary, the interaction of the oblique reflection directions and the limit linearized dynamics $\dot{x} = Ax + \dot{z}$ can be very complicated. When there are forcing terms to the boundary, as under the assumptions of Theorem 3.1, the individual components of $U^n(\cdot)$ can be separated into two subsets. One converges to the "zero" process, and the other to the limit reflected diffusion.

10. Discussion: Comparisons with the unconstrained case. For the unconstrained case, the stationary distribution of either form of (1.5) is normal with zero mean, and the covariance can be computed analytically. Consider the constrained problem. Suppose that the components of θ can be divided into two classes, where the first corresponds to forcing terms to the boundary and, for the latter, $\bar{\theta}^i$ is interior to the constraint interval. Then the steady state distribution is simple: The normalized mean square errors of the components with forcing terms to the boundary are zero. The others are unconstrained in the limit, and their distribution is normal with zero mean and can be computed analytically. The variances can be much smaller than those for the original unconstrained problem, since the dimension is smaller. It is common for constrained problems to have some components with forcing terms to the boundary.

If, however, some components without forcing terms to the boundary have $\bar{\theta}^i$ at the end of the constraint interval, then we have to deal with the general reflected diffusion of one of the forms in (1.9). The form of the stationary distribution of (1.9) is not known, and some sort of simulation seems to be called for to evaluate it. The stationary means and covariances were evaluated for several two dimensional systems. For those examples, where there were no forcing terms to the boundary, the constraints did not increase the mean square values. Generally, the mean square values of the limit variances for the constrained problem were close to the variances for the unconstrained problem, perhaps a little smaller. However, the mean of the limit distribution was not zero. Thus, the limit variances were smaller than those for the unconstrained problem, suggesting that the constrained problem has less asymptotic variability, even when the limit point is the same as that for the unconstrained problem. If one of the components did have a forcing term to the boundary, then even if the other component of the limit was on the boundary, its stationary variance was smaller (again, perhaps much smaller), since the effective dimension is reduced. A method for analytically evaluating or approximating the first two moments for the constrained problem is needed.

Appendix. Construction of the Liapunov function $V(\cdot)$. Theorem 4.2 will be proved in this section. Consider an unconstrained system $\dot{x} = b(x)$. Uniform asymptotic stability of this ODE implies the existence of a Liapunov function. This Liapunov function can be smoothed to make it twice continuously differentiable and then used to prove the recurrence of the process defined by the unconstrained SDE $dx = b(x)dt + \sigma(x)dw$ if $\sigma(\cdot)$ is bounded. The aim is to do this for the reflected SDE of concern, where the deterministic (fluid) model is (4.1a) ((4.1b) is obviously similar) and the reflected SDE is (1.9a). If the result is to be adapted for use on the discrete parameter SA algorithm, then we need to allow reflection from a neighborhood "a little outside" of L, the state space of the limit process (1.9). A serious problem, which is not present in the unconstrained case, is that there are two vector fields to be dealt with, the drift in L and the reflection on the boundary and just outside of the boundary.

This problem was solved in [11] for the case where the drift vector $b(x)$ was simply a constant vector $\bar{b}$. The general idea of the proof there can be extended to cover our case, where $b(x) = Ax$, but a number of alterations need to be made. Since the proof in [11] is complicated, we will provide a detailed guide to the necessary changes. Their state space was the orthant $\{x : x^i \geq 0, i \leq k\}$. In our case, some components are unconstrained, and we use the L defined in Assumption 4.2, where components $1, \ldots, \nu$ are constrained and components $\nu + 1, \ldots, k$ are unconstrained. This alteration is insignificant and, in itself, requires only a minor notational change in [11]. The radial homogeneity of the dynamical terms, drift, and reflection (i.e., they have the same value at all points on each ray from the origin) was heavily used in [11]. This is the reason for the use of the form $dx = [Ax/|x|]\,dt + dz$ below. The paths of this normalized model, when plotted in phase space, are the same as those of the original fluid model (1.5a), and actual paths differ only by a time scaling. Thus, we work with

$$dx = \frac{Ax}{|x|}dt + dz, \qquad x \in L, \tag{A.1}$$

and assume that Assumption 4.2 holds.

The method of analysis uses an extension of the dynamics to all of $\mathbb{R}^k - \{0\}$. The

major difficulty concerns the need to get a smooth Liapunov function in $\mathbb{R}^k - \{0\}$, where the dynamical term is discontinuous on ∂L, where it switches from the drift to the reflection.

The main steps of the proof in [11], adjusted to our form of L but where $dx = \bar{b}dt + dz$, are the following.

1. First define the extended system $\dot{x} = v(x)$: For $x \in L - \{0\}$, set $v(x) = \bar{b}$, where $\bar{b}$ is their constant drift vector. For $x \in \mathbb{R}^k - L - \{0\}$, set $v(x) = r_0(x)$, where $r_0(x)$ is any unit vector in cone$\{d_i : i \in \text{In}(x)\}$, where $\text{In}(x) = \{i : x^i = 0\}$. The definitions at the origin are arbitrary.
2. Smooth $v(\cdot)$: Let $\rho(\cdot)$ be a smoothing kernel and for small $a > 0$ define

$$v^a(x) = c(a|x|) \int \rho\left(\frac{x-y}{a|x|}\right) v(y)dy,$$

where $c(a)$ is the normalizing constant which ensures that the integral is unity when $v(\cdot) \equiv 1$.
3. Form a convex combination of $v^a(\cdot)$ and the original dynamics as follows. Let $d(x, M) = \inf_{y \in M} |x - y|$ denote the distance between x and the set M. Define the real-valued, infinitely differentiable, and nonincreasing function on $[0, \infty)$ by $g(s) = 1, s \in [0, .5]$, $g(s) = 0, s \in [1, \infty)$. Define

$$v_i^a(x) = g\left(\frac{d(x, \partial L_i)}{a|x|}\right) d_i + \left[1 - g\left(\frac{d(x, \partial L_i)}{a|x|}\right)\right] v^a(x), \qquad 1 \le i \le \nu,$$

$$v_0^a(x) = g\left(\frac{d(x, L)}{a|x|}\right) \bar{b} + \left[1 - g\left(\frac{d(x, L)}{a|x|}\right)\right] v^a(x).$$

4. Define the set-valued function

$$K^a(x) = \text{conv}\left\{v_i^a(x), 0 \le i \le \nu\right\},$$

where conv denotes the set of all convex combinations. For each x, define $K(x)$ as the "limit" set containing points y satisfying the following condition: there exists $a_n \to 0, x_n \to x, y_n \to y$, where $y_n \in K^{a_n}(x_n)$. Define the ODE with the set-valued right-hand side

$$\dot{\phi} \in K^a(\phi), \ \phi(0) \in \text{ some compact set.} \tag{A.2}$$

5. Let $k(\cdot)$ be a $[0, 1]$-valued, infinitely differentiable, and nondecreasing function on $[0, \infty)$ such that $k(s) = 0$ for $s \in [0, 1]$ and $k(s) = 1$ for $s \in [2, \infty)$. Define the first "tentative" Liapunov function

$$V^a(x) = \sup_{\phi} \int_0^\infty k(|\phi(s)|)ds,$$

where the sup is over all solutions to (A.2) with $\phi(0) = x$.
6. For small $b > 0$, smooth $V^a(\cdot)$ as

$$V^{a,b}(x) = c(b) \int \rho\left(\frac{x-y}{b}\right) V^a(y)dy.$$

7. There is a star-shaped set $S \subset \mathbb{R}^k$ such that, for some $\bar{v} > 0$,

$$\partial S = \left\{x : V^{a,b}(x) = \bar{v}\right\}$$

is twice continuously differentiable. Now define the final Liapunov function $V(y), y \in \mathbb{R}^k - \{0\}$ by

$$V(\alpha x) = \alpha V^{a,b}(x),\ \alpha > 0, x \in \partial S, \quad \text{for } y = \alpha x.$$

8. $V(\cdot)$ is twice continuously differentiable in $\mathbb{R}^k - \{0\}$. For $\alpha > 0, x \neq 0$, $V(\alpha x) = \alpha V(x)$ and $\alpha V_{yy}(y)\big|_{y=\alpha x} = V_{xx}(x)$, and hence $V_{xx}(x)$ goes to zero as $O(1/|x|)$ as $x \to \infty$. There is $\epsilon > 0$ such that for $x \in L - N_\epsilon(0), V'_x(x)\bar{b} \leq -c,\ c > 0$, and for $x \in \mathbb{R}^k - L - N_\epsilon(0)$, $V'_x(x)d \leq -c$, $d \in r(x)$. The bounds in Theorem 4.2 hold.

In our case, $Ax/|x|$ replaces the constant vector $\bar{b}$ in steps 1 and 3, but all other steps and definitions are the same. The radial homogeneity which played such an important part in [11] holds. Suppose that the Liapunov function has the same properties for our case but with $Ax/|x|$ replacing $\bar{b}$. Then $V'_x(x)Ax/|x| \leq -c$ for $x \in L - N_\epsilon(0)$. Hence $V'_x(x)Ax \leq -c|x|$ there. For $x \in \mathbb{R}^k - L - N_\epsilon(0)$, the fact that $V'_x(x)d \leq -c$ for all $d \in r(x)$ implies that $V(x) \leq V(\tilde{x})$, where $\tilde{x} \in \mathbb{R}^k - L - N_\epsilon(0)$ and x is its "projection" onto L in any feasible reflection direction. This fact is used to get that $V(U_{n+1}) \leq V(\tilde{U}_{n+1})$ for large U_n. We next outline the required modifications of the proof in [11]. This will be done in a series of lemmas which go over the main steps in the reference but are adjusted for our case.

The properties of $K^a(\cdot)$ and $K(\cdot)$ given in [11] are easily seen to still hold. Specifically, we have the following:

1. For each $a > 0$, $K^a(\cdot)$ is the convex hull of $\nu + 1$ vector-valued functions that are locally Lipschitz continuous on $\mathbb{R}^k - \{0\}$. This holds since $Ax/|x|$ is locally Lipschitz at $x \neq 0$.
2. For each $a > 0$, $\alpha > 0$, and $x \in \mathbb{R}^k$, $K^a(\alpha x) = K^a(x)$.
3. Let $x \neq 0$. If $d(x, L) \leq a|x|/2$, then $Ax/|x| \in K^a(x)$. If $d(x, \partial L_i) \leq a|x|/2$, then $d_i \in K^a(x)$.
4. Let $\lambda(x) = \{i \leq \nu : x^i < 0\}$. If $d(x, L) > 0$, then $K(x)$ is contained in $\text{conv}\{d_i : i \in \lambda(x)\}$.
5. If $x \in L^0$, the interior of L, then $K(x) = \{Ax/|x|\}$, and if $x \in \partial L$, then

$$K(x) = \text{conv}\left(\{Ax/|x|\} \cup \{d_i : i \in \text{In}(x)\}\right).$$

6. $K(x)$ is an upper-semicontinuous function of $x \in \mathbb{R}^k - \{0\}$, in the sense that $x_n \to x$, $v_n \to v$, and $v_n \in K(x_n)$ implies that $v \in K(x)$.

We say that a solution $\phi(\cdot)$ to a reflected ODE is *attracted to the origin* if for any $\epsilon > 0$ there exists $T < \infty$ such that $t \geq T$ implies that $|\phi(t)| \leq \epsilon$. This is the same as saying that all solutions converge to the origin.

LEMMA A.1 (see Proposition 3.3 in [11]). *Let $\Gamma(a)$ denote the set of solutions to*

$$\dot{\phi}(t) \in K^a\left(\phi(t)\right), \quad \phi(0) = x^a,$$

where $\{x^a,\ a \in (0,\ 1]\}$ is any bounded set in $\mathbb{R}^k$. We then have the following conclusions:

(i) *The set $\{\Gamma(a),\ a \in (0,\ 1]\}$ is precompact.*

(ii) *Suppose that* $a_n \to 0$. *Suppose that* $\phi^{a_n}(\cdot) \to \phi(\cdot)$ *uniformly on each bounded time interval, where* $\phi^{a_n}(\cdot) \in \Gamma(a_n)$, *and* $x^{a_n} = \phi^{a_n}(0) \to x$. *Then we have the following:* (a) *If* $x \in L$, *then modulo a rescaling of time,* $\phi(\cdot), \phi(0) = x$ *solves* (A.1). (b) *If* $x \notin L$, *then* $\phi(\tau) \in L$ *for some* $\tau < \infty$. *Furthermore, modulo a rescaling of time,* $\phi(\cdot + \tau)$ *solves* (A.1).

Comment on the proof. Since $K^a(x)$ is uniformly bounded over all x and a, Ascoli's theorem can easily be applied to prove (i) exactly the same as in [11].

For part (ii), it is first shown in [11] that if $\phi^a(\cdot) \to \phi(\cdot)$, as $a \to 0$, where $\phi^a(\cdot) \in \Gamma(a)$, then $\phi(\cdot)$ satisfies

$$\dot{\phi}(t) \in K(\phi(t)), \text{ almost all } t \in [0, \infty). \tag{A.3}$$

The proof of this relies on the uniform boundedness of $K^a(u)$, which we still have, and the proof is the same as in the reference, except for one point. The proof is slightly different by the fact that a Lipschitz continuous function $\phi(\cdot)$, which satisfies $\dot{\phi}(t) \in K(\phi(t))$ for almost all t, can be written as the convex combination

$$\begin{aligned} \dot{\phi}(t) &= q_0(t)\frac{A\phi(t)}{|\phi(t)|} + \sum_{i \in \mathrm{In}(\phi(t))} q_i(t) d_i, \\ q_0(t) &+ \sum_{i \in \mathrm{In}(\phi(t))} q_i(t) = 1, \end{aligned}$$

where $q_i(t)$ are nonnegative measurable functions. Following the general approach in the reference, Michael's selection theorem [1] is used to obtain the functions $q_i(\cdot)$, $i = 0, \ldots, \nu$. The details are omitted here and can be found in [5].

LEMMA A.2 (see Proposition 3.4 in [11]). *Assume that all solutions of* (A.1) *are attracting to the origin. Then the following conclusions hold:* (i) *Given* $\alpha > 0$, *there exist* $r > 0$ *and* $a_0 > 0$ *such that for all* $a \in (0, a_0)$

$$\dot{\phi} \in K^a(\phi) \quad \textit{and} \quad |\phi(0)| \le r$$

implies that

$$|\phi(t)| \le \alpha \quad \textit{for all } t \ge 0.$$

(ii) *Given* $r > 0$ *and* $R < \infty$, *there exist* $T < \infty$ *and* $a_0 > 0$ *such that, for all* $a \in (0, a_0)$,

$$\dot{\phi} \in K^a(\phi) \quad \textit{and} \quad |\phi(0)| \le R$$

implies that

$$|\phi(t)| \le r \quad \textit{for some } t \le T.$$

Comment on the proof. The proof is the same as in [11]. If part (ii) did not hold, then the assumption that the solutions to (A.1) are attracting to the origin would be violated in view of Lemma A.1. To prove part (i), choose $R = 1$ and $r = 1/2$ in part (ii) of this lemma and define

$$\kappa = \sup_{a \in (0, a_0)} \sup_{\phi(\cdot): |\phi(0)| \le 1, \dot{\phi}(t) \in K^a(\phi(t))} \sup_{0 \le t \le \tau_\phi} |\phi(t)|, \tag{A.4}$$

where $\tau_\phi = \inf\{t : |\phi(t)| \le \frac{1}{2}\}$. We have $\kappa < \infty$ for $a_0 > 0$ sufficiently small; otherwise the assumption of the lemma is violated.

So far we have shown part (i) for the case in which $\alpha = \kappa$, and $r = 1$ can be used. We will now show that given an *arbitrary* $\alpha > 0$, we may choose $r = \alpha/\kappa$. Given $\phi_\alpha(0)$ satisfying $|\phi_\alpha(0)| \le \alpha/\kappa$, define $\phi(0)$ such that $\phi_\alpha(0) = \alpha\phi(0)/\kappa$. For any path $\phi(\cdot)$ with such an initial condition and $\dot\phi \in K^a(\phi)$, define $\tau_{\phi,\alpha,\kappa} = \inf\{t : |\phi(t)| \le \alpha/(2\kappa)\}$. Then from the radial homogeneity of the vector field and (A.4),

$$\alpha = \sup_{a\in(0,a_0)} \sup_{\phi_\kappa(\cdot):\dot\phi_\kappa(t)\in K^a(\phi_\kappa(t)),|\phi_\kappa(0)|\le\alpha/\kappa} \sup_{0\le t\le\tau_{\phi,\alpha,\kappa}} |\phi(t)|,$$

This proves part (i). □

The following lemmas give some properties of $V^a(u)$.

LEMMA A.3 (see Proposition 3.5 in [11]). *Recall the definition of $k(\cdot)$ given below* (A.2). *There exist $a_0 > 0$ and $r_0 > 0$ such that, for all $a \in (0, a_0)$, $V^a(x) = 0$ for $|x| \le r_0$, $V^a(\cdot)$ is finite and locally Lipschitz continuous on $\mathbb{R}^k$, and*

$$[V_x^a(x)]'y \le -k(x) \tag{A.5}$$

for almost all $x \in \mathbb{R}^k - \{0\}$ and every $y \in K^a(x)$.

Comment on the proof. The proof is the same as in [11], but we will make a few comments. The fact that $V^a(x) = 0$ for $a \in (0, a_0)$ and $|x| \le r_0$ follows easily from Lemma A.2. Specifically we have, for $r = r_0$ small enough, that for any $a \in (0, a_0)$, $|\phi(t)| \le 1\ \forall t \ge 0$, where $\dot\phi \in K^a(\phi)$. Thus $k(\phi(t)) = 0\ \forall t \ge 0$, and the result follows from the definition of $V^a(\cdot)$.

That $V^a(\cdot)$ is finite also follows easily from Lemma A.2. By part (ii), for given $r > 0$ and $|x| < R$, there is an a_0 such that for any $a \in (0, a_0)$, $|\phi(t)| \le r$ for some $t < T$. Choosing r small enough and using the analysis in the previous paragraph gives that $0 \le V^a(x) < \infty$ for $|x| < R$. Finally, the radial homogeneity of $K^a(\cdot)$ gives that $0 \le V^a(x) < \infty$ for all x, and where $a > 0$ is sufficiently small.

The proof that $V^a(\cdot)$ is locally Lipschitz follows from the same method given in [11, pp. 693–694]. In particular, recall that $K^a(\cdot)$ is the convex hull of vector value functions that are locally Lipschitz continuous in $\mathbb{R}^k - \{0\}$. The proof that $V^a(\cdot)$ is locally Lipschitz follows from this property and the local Lipschitz continuity of $k(\cdot)$.

We now show (A.5). Let $V^a(\cdot)$ be differentiable at x. From the definition of $V^a(\cdot)$ we know that, for any $\gamma > 0$,

$$V^a(x) \ge \int_0^\gamma k(\phi(s))ds + \int_\gamma^\infty k(\phi(s))ds, \tag{A.6}$$

where $\phi(0) = x$ and $\dot\phi \in K^a(\phi)$. First we will specify a path $\phi(\cdot)$ in the time intervals $[0, \gamma]$ and $[\gamma, \infty)$, where γ is small and $\gamma > 0$. Then we examine (A.6) using this $\phi(\cdot)$.

For $s \in [0, \gamma]$ choose $\phi(\cdot)$ such that $\dot\phi(s) \in K^a(\phi(s))$, but $\dot\phi(s) \to \dot\phi(0)$ as $s \to 0$. Write $w = \phi(\gamma)$. Let $\epsilon > 0$ be arbitrary. For $s \in [\gamma, \infty)$, choose $\phi(s) = \phi^{\epsilon,w}(s - \gamma)$, where $\phi^{\epsilon,w}(\cdot)$ is such that

$$\phi^{\epsilon,w}(0) = w, \quad \dot\phi^{\epsilon,w}(s) \in K^a\left(\phi^{\epsilon,w}(s)\right),$$

and

$$V^a(w) \le \int_0^\infty k(\phi^{\epsilon,w}(s))ds + \epsilon. \tag{A.7}$$

By the continuity of $k(\cdot)$, the first integral in (A.6) can be written as $\gamma k(x)+o(\gamma)$. The second integral can be written as

$$\int_0^\infty k(\phi(s+\gamma))ds = \int_0^\infty k\left(\phi^{\epsilon,w}(s)\right) ds.$$

Thus

$$V^a(x) \geq \gamma k(x) + o(\gamma) + V^a(w) - \epsilon. \tag{A.8}$$

By the fact that $\dot{\phi}(s) \to \dot{\phi}(0)$, $w = x + \gamma\dot{\phi}(0) + o(\gamma)$. Let $\epsilon \to 0$. Then we can write

$$V^a(x) \geq \gamma k(x) + o(\gamma) + V^a(x + \gamma\dot{\phi}(0) + o(\gamma)).$$

Sending $\gamma \to 0$ and using the differentiability of $V^a(\cdot)$ at x gives the result. □

LEMMA A.4 (see Proposition 3.6 in [11]). *There exist constants $d_1 > 0$ and d_2 such that*

$$V^a(x) \geq d_1|x| - d_2. \tag{A.9}$$

Comment on the proof. The proof is the same as in [11]. Since $K^a(\cdot)$ is uniformly bounded, $\dot{\phi}(t) \in K^a(\phi(t))$ is uniformly bounded. Thus there is a lower bound on the time it takes for a solution to the differential inclusion to reach a small neighborhood of the origin. This leads to (A.9). □

LEMMA A.5 (see Proposition 3.7 in [11]). *Let $x \neq 0$ be a point at which $V^a(\cdot)$ is differentiable. Then*

$$[V_x^a(x)]' x/|x| \geq V^a(x)/|x|.$$

Comment on the proof. The proof is the same as in [11]. Let $\alpha > 0$. By radial homogeneity, the evolution of a path $\phi(t)$ satisfying the differential inclusion (A.2) and $\phi(0) = x$ can be used to construct a path $\theta^\alpha(t)$ satisfying the same differential inclusion and $\theta^\alpha(0) = (1+\alpha)x$. That is, we let $\theta^\alpha(t) = (1+\alpha)\phi(t/(1+\alpha))$. From this we can get $V^a((1+\alpha)x) - V^a(x) \geq \alpha V^a(x)$, and the conclusion of the lemma follows from this inequality. □

The construction and properties of $V(\cdot)$. Since $V^a(\cdot)$ is only Lipschitz continuous and our final Liapunov function $V(\cdot)$ needs to be twice continuously differentiable, $V^a(\cdot)$ was smoothed to create $V^{a,b}(\cdot)$:

$$V^{a,b}(x) = c(b) \int \rho\left(\frac{x-y}{b}\right) V^a(y)dy.$$

From (A.5), for large enough $M < \infty$ and $b \in (0,1]$ we have

$$\left[V_x^{a,b}(x)\right]' Ax/|x| \leq -1, \quad |x| \geq M,$$

and

$$\left[V_x^{a,b}(x)\right]' d_i \leq -1, \quad i \in \mathrm{In}(x),\ |x| \geq M. \tag{A.10}$$

Let

$$E = \sup_{x:|x|\leq M_1} V^a(x),$$

where $M \leq M_1 < \infty$ and is such that $[V_x^a(x)]' x/|x| \geq C_0$ for some $C_0 > 0$ when $|x| \geq M_1 - 1$.

The existence of such M and M_1 follows from Lemmas A.4 and A.5. We have the following properties for $V^a(\cdot)$:

(i) $\{x : V^a(x) = E+2\} \subset \{x : M_1 < |x| < M_2\}$, where $M < M_1 < M_2 < \infty$ and M_2 is such that $|x| \geq M_2$ implies $V^a(x) \geq E+3$.
(ii) $[V_x^a(x)]' x/|x| \geq C_0$, for $|x| \geq M_1 - 1$ and some $C_0 > 0$.
(iii) $V^a(x) \leq E$ for $|x| \leq M_1$.

By the local Lipschitz continuity of $V^a(\cdot)$ given in Lemma A.3 and the definition of $V^{a,b}(\cdot)$ it is easy to see that, for small enough b, properties (i)–(iii) hold for $V^{a,b}(\cdot)$ if for part (ii) we substitute $C_0/2$ for C_0 and for part (iii) we substitute $E+1$ for E. Define the set

$$S = \left\{x : V^{a,\,b}(x) \leq E+2\right\}.$$

These properties imply that S is star-shaped; i.e., a segment containing the origin and a point in S is contained in the interior of S. Then, as in [11], the final Liapunov function $V(\cdot)$ is defined by its level sets as

$$\{x : V(x) \leq l\} = \{lx : x \in S\}.$$

The star-shaped property implies that $V(\cdot)$ is well defined; i.e., for each $u \in \mathbb{R}^r$ there exists a unique l such that $u \in \partial(lS)$. These details are in [11, pp. 697–698].

There are several easily seen properties of $V(\cdot)$ which are useful for the tightness proofs of the normalized iterates in section 4. We list these properties and briefly comment on some of them.

P1. $V(\cdot)$ is twice continuously differentiable on $\mathbb{R}^k - \{0\}$. The proof uses the implicit function theorem, and the details are in [11, pp. 697–698].
P2. $V(\alpha x) = \alpha V(x)$. This is easily seen from the construction of $V(\cdot)$.
P3. $V_y(y)\big|_{y=\alpha x} = V_x(x)$.
P4. $\alpha V_{yy}(y)\big|_{y=\alpha x} = V_{xx}(x)$. Hence $|V_{xx}(x)| \to 0$ as $1/|x|$.
P5. For some real $c_1 > 0$, $c_2 > 0$, and $d_1 > 0$, $c_1|x| + c_2 \geq V(x) \geq d_1|x|$. This is a consequence of the fact that the gradient in the radial direction is constant.
P6. For $x \in \mathbb{R}^k - \{0\}$, $0 < V(x)/|x| \leq C$ for some $C > 0$. To see this, from P5 we have

$$c_1 + \frac{c_2}{|x|} \geq \frac{V(x)}{|x|} \geq d_1 > 0.$$

Since $V(\alpha x)/|\alpha x| = V(x)/|x|$, substituting αx for x in the above inequality, where α is arbitrarily large, we have that for some $C > c_1$

$$0 < \frac{V(x)}{|x|} \leq C \quad \text{for } x \in \mathbb{R}^K - \{0\}.$$

P7. There exists a $c > 0$ such that
$V_x'(x)Ax/|x| \leq -c$ and
$V_x'(x)d_i \leq -c$, $i \in \mathrm{In}(x)$.
This follows from the analogous inequalities for $V^{a,b}(\cdot)$.
P8. With the definition $V(0) = 0$, $V(\cdot)$ is continuous on $\mathbb{R}^k$ and globally Lipschitz continuous. The continuity follows from the construction. The global Lipschitz continuity follows from property P6.
P9. $V(x) \leq V(\tilde{x})$, where $\tilde{x} \in \mathbb{R}^k - L - N_\epsilon(0)$ and its projection onto L is x. This follows from P7.

REFERENCES

[1] J. P. Aubin and A. Cellina, *Differential Inclusions*, Springer-Verlag, Berlin, New York, 1980.

[2] A. Benveniste, M. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximation*, Springer-Verlag, Berlin, New York, 1990.

[3] P. Billingsley, *Convergence of Probability Measures*, John Wiley, New York, 1968.

[4] G. Blankenship and G. C. Papanicolaou, *Stability and control of stochastic systems with wide-band noise disturbances* I, SIAM J. Appl. Math., 34 (1978), pp. 437–476.

[5] R. Buche, *Stochastic Approximation: Rate of Convergence for Constrained Algorithms; Asynchronous Algorithms and Analysis of a Competitive Resource Sharing System*, Ph.D. thesis, Applied Mathematics Department, Brown University, Providence, RI, 2000.

[6] J. L. Doob, *Asymptotic properties of Markov transition probabilities*, Trans. Amer. Math. Soc., 63 (1948), pp. 393–421.

[7] P. Dupuis and H. Ishii, *On Lipschitz continuity of the solution mapping to the Skorokhod problem, with applications*, Stochastics Stochastics Rep., 35 (1991), pp. 31–62.

[8] P. Dupuis and H. J. Kushner, *Stochastic approximations via large deviations: Asymptotic properties*, SIAM J. Control Optim., 23 (1985), pp. 675–696.

[9] P. Dupuis and H. J. Kushner, *Asymptotic behavior of constrained stochastic approximations via the theory of large deviations*, Probab. Theory Related Fields, 75 (1987), pp. 223–244.

[10] P. Dupuis and H. J. Kushner, *Stochastic approximation and large deviations: Upper bounds and w.p.*1 *convergence*, SIAM J. Control Optim., 27 (1989) pp. 1108–1135.

[11] P. Dupuis and R. J. Williams, *Lyapunov functions for semimartingale reflecting Brownian motions*, Ann. Probab., 22 (1994), pp. 680–702.

[12] E. B. Dynkin, *Markov Processes*, Springer-Verlag, Berlin, New York, 1965.

[13] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*, John Wiley, New York, 1986.

[14] L. Gerencér, *Rate of convergence of recursive estimators*, SIAM J. Control Optim., 30 (1992), pp. 1200–1227.

[15] J. M. Harrison and R. J. Williams, *Brownian models of open queueing networks with homogeneous customer populations*, Stochastics Stochastics Rep., 22 (1987), pp. 77–115.

[16] A. P. Korostelev, *Stochastic Recurrent Processes*, Nauka, Moscow, 1984.

[17] H. J. Kushner, *Approximation and Weak Convergence Methods for Random Processes with Applications to Stochastic Systems Theory*, MIT Press, Cambridge, MA, 1984.

[18] H. J. Kushner and D. S. Clark, *Stochastic Approximation for Constrained and Unconstrained Systems*, Springer-Verlag, Berlin, New York, 1978.

[19] H. J. Kushner and P. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time*, Springer-Verlag, Berlin, New York, 1992.

[20] H. J. Kushner and E. Sanvicente, *Stochastic approximation for constrained systems with observation noise on the system and constraint*, Automatica J. IFAC, 11 (1975), pp. 375–380.

[21] H. J. Kushner and G. Yin, *Stochastic Approximation Algorithms and Applications*, Springer-Verlag, Berlin, New York, 1997.

[22] M. T. Wasan, *Stochastic Approximation*, Cambridge University Press, Cambridge, UK, 1969.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 4, pp. 1042–1071

NONLINEAR TRACKING OVER COMPACT SETS WITH LINEAR DYNAMICALLY VARYING H^∞ CONTROL*

STEPHAN BOHACEK† AND EDMUND JONCKHEERE†

Abstract. Linear dynamically varying H^∞ controllers are developed for tracking natural trajectories of a broad class of nonlinear systems defined over compact sets. It is shown that the existence of a suboptimal H^∞ controller is related to the existence of a bounded solution to a functional algebraic Riccati equation. Even though nonlinear systems running over compact sets could exhibit sensitive dependence on initial conditions, the Riccati solution is continuous in the suboptimal case, but it may be discontinuous in the optimal case.

Key words. nonlinear systems, trajectory tracking, linear parametrically and dynamically varying systems, state space H^∞ robust control, functional algebraic Riccati equation

AMS subject classifications. 49K40, 49L20, 37A99, 37B25, 37C05, 37C75, 37F15

PII. S0363012999350584

1. Introduction.

Nonlinear tracking has been thoroughly investigated. A popular approach is to linearize the system around an operating point, generate a linear controller for each operating point, and "schedule" the controllers in such a way that the closed-loop system remains stable as the operating point changes. In this approach, the nonlinear tracking error system is modeled, approximately, as a linear system with parameters that vary as the operating point varies. Such systems have been extensively studied [3], [4], [5], [6], [25], [28], [33] and are known as *linear parametrically varying* (LPV) systems.

For the purpose of comparing the various LPV concepts, it is convenient to introduce *linear set-valued dynamically varying* (LSVDV) systems [11]:

$$\text{(1)}\qquad \begin{bmatrix} x(k+1) \\ z(k) \end{bmatrix} = \begin{bmatrix} A_{\theta(k)} & B_{1_{\theta(k)}} & B_{2_{\theta(k)}} \\ C_{\theta(k)} & D_{1_{\theta(k)}} & D_{2_{\theta(k)}} \end{bmatrix} \begin{bmatrix} x(k) \\ w(k) \\ u(k) \end{bmatrix},$$
$$\theta(k+1) \in \mathcal{F}(\theta(k)) \subseteq \Theta,$$
$$\text{with } \theta(0) = \theta_o \text{ and } x(0) = x_o.$$

Here the parameter vector θ varies according to a set-valued dynamical system, continuous for the Hausdorff metric; w is the disturbance input, u the control, and z the controlled output.

In the most traditional LPV approach [4], [6], [5], [23], [25], all that is known about the parameter dynamics is that $\mathcal{F}(\theta) = \Theta$. The advantage of this model is that if Θ is a convex polytope, then there are many computationally efficient controller synthesis methods [14]. Most of these approaches generate a suboptimal solution via a *linear matrix inequality* (LMI). However, these approaches can be conservative.

A slight refinement of the above LPV method consists of putting bounds on the rate at which the system parameters vary, i.e., $\mathcal{F}(\theta(k)) = \mathcal{B}_{\theta(k)}(\Delta)$, the ball with

*Received by the editors January 13, 1999; accepted for publication (in revised form) January 31, 2001; published electronically November 28, 2001. This research was supported in part by NSF grant ECS-98-12594 and in part by AFOSR grant F49620-93-I-0505.
http://www.siam.org/journals/sicon/40-4/35058.html
†Department of Electrical Engineering—Systems, University of Southern California, 3740 McClintock Ave., Room 306, Los Angeles, CA 90089-2563 (bohacek@math.usc.edu, jonckhee@eudoxus.usc.edu).

radius Δ and with its center at $\theta(k)$. There are efficient methods based on functional LMIs for designing controllers for these modified LPV systems [16], [33], [34], [35]. However, these design methods could fail when the parameters vary drastically, for example, when the controller needs to account for failures which lead to sudden changes in the system parameters [21]. Also, typically, these methods are conservative. Nonconservative LPV approaches are pursued in [11] and [29].

Another popular type of LPV system is the *jump linear* (JL) system [15], [19]. Here $\Theta = \{\Theta_1, \Theta_2, \dots\}$ is discrete, and $\mathcal{F}(\theta(k))$ is equipped with a probability measure depending on $\theta(k)$ only, so that the transition among the Θ_i's is a Markov chain. The jump linear method for designing a controller for a such system is optimal (hence, nonconservative). The controller is provided by the solution to a system of coupled Riccati equations. Furthermore, there are efficient methods to compute the optimal controller [1], [2], [12].

A *linear dynamically varying* (LDV) system is a LSVDV system in which the parameter dynamics are completely known, that is, $\mathcal{F}(\theta(k))$ is reduced to a point $f(\theta(k))$. In [8] it was shown that a linear-quadratic (LQ) controller for such a system (with $w = 0$) can be found by solving a *functional* algebraic Riccati equation (FARE). It should be noted that this functional algebraic Riccati *equation* is the LDV substitute for the functional linear matrix *inequality* of most other LPV approaches. Furthermore, the FARE of LDV design provides the *optimal* solution, while the functional LMI only provides a *suboptimal* solution. The mathematical difficulty with the LDV approach is proving that the solution to the FARE is continuous, in which case the feedback gain matrix is a continuous function of the parameters. The LPV approaches described above avoid this continuity question by assuming a priori that the solution to the relevant functional LMI is continuous [33], polynomial [35], affine [16], [34], or even constant [4], [6], [5], [23], [25]. Since an arbitrary accuracy approximation of a discontinuous function has to duplicate the *exact* behavior at the discontinuity points, which are potentially uncountable in number, a discontinuous solution is numerically intractable, so that the continuity assumption is justifiable. However, it is important to know how constraining this continuity assumption is.

Tracking trajectories of the important class of hyperbolic nonlinear systems on compact sets can be accomplished by modeling the dynamics as a Markov chain [22] and resorting to JL methods. However, the resulting closed-loop system is only *stochastically* stable, and it is not possible to show directly that the system is robustly stable. For this reason, the typical JL approach is not appropriate for the nonlinear tracking problem. The connection between JL and LDV control systems designs is examined in [10].

While in [8] LDV systems were stabilized using LQ methods, here the same systems are stabilized by means of H^∞ methods. This paper shows that, if the parameter dynamics are completely known, then the existence of a suboptimal H^∞ controller is equivalent to the existence of a continuous solution to the FARE. Of particular interest are LDV systems that arise as linearized versions of nonlinear tracking error dynamics. In this case, it can be shown that the linearization error is a bounded feedback around the linearized system, so that the H^∞ formulation is well-suited to minimize the effect of the error due to linearization and amplify the domain of attraction.

The paper proceeds as follows. The next section formalizes the tracking control problem of interest and shows how the tracking error dynamics can be approximated as an LDV system. Section 3 formally develops LDV systems. Section 4 develops the

suboptimal H^∞ controller for this class of systems. Section 5 provides the proofs of the main technical results. Section 6 shows that these linear controllers are suitable for stabilization of nonlinear dynamical systems.

Notation: $|x(k)| := (x'(k)\,x(k))^{1/2}$, $\|x\|_{[k,j]} := (\sum_{i=k}^{j} x'(i)\,x(i))^{1/2}$, and $\|x\|_{l_2} := \|x\|_{[0,\infty)}$. If $x \in \mathbb{R}^n$, then $\|x\|_\infty := \max_{i\le n} |x_i|$ and, if $x \in \mathbb{R}^{n\times\mathbb{Z}}$, then $\|x\|_{l_\infty} := \sup_{k\in\mathbb{Z}} |x(k)|$. If A is a matrix, then $\|A\| := \sup_{|x|=1} |Ax|$, whereas, if $T : l_2 \to l_2$, then $\|T\| := \sup_{\|x\|_{l_2}=1} \|Tx\|_{l_2}$; the context in which these norms are used will resolve potential confusion. If $f : \Theta \times \mathbb{R}^m \to \Theta$ with $\Theta \subset \mathbb{R}^n$, then $\frac{\partial f}{\partial \theta}(\theta, u)$ denotes the Jacobian matrix of f where the derivatives are taken with respect to θ and are evaluated at $(\theta, u) \in \Theta \times \mathbb{R}^m$. Define $\frac{\partial f}{\partial u}(\theta, u)$ similarly. With reference to system (1), let $z_{\theta_o}(u, w, x_o)$ denote the output signal z due to initial conditions $\theta(0) = \theta_o$ and $x(0) = x_o$, and input signals u and w. Let $z_{\theta_o}(u, w, x_o; k)$ denote this output at time k. Let $z_{\theta_o}(F, w, x_o)$ and $z_{\theta_o}(F, w, x_o; k)$ be defined similarly, except that the control u is replaced by the control law defined by F. For succinctness, we often write $f(\theta) := f(\theta, 0)$.

2. Problem statement. A dynamical system $\theta(k+1) = f(\theta(k))$, where $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$, gives rise to a string of nested invariant subsets $P(f) \subseteq \overline{P(f)} \subseteq \overline{R(f)} \subseteq NW(f)$, where $P(f)$ is the periodic set, $\overline{P(f)}$ its closure, $\overline{R(f)}$ the closure of the recurrent set, and $NW(f)$ the nonwandering set [22]. We specifically consider systems where $NW(f)$ is bounded, in which case $\overline{P(f)}$, $\overline{R(f)}$, and $NW(f)$ are compact, and we choose the domain Θ to be any of those compact invariant sets. More generally, Θ could be taken to be any compact invariant subset. In particular, if f is an Axiom A diffeomorphism satisfying the strong transversality condition, then $NW(f)$ is a disjoint union of attractors [27], which by definition are compact and invariant and hence could be taken to be Θ. If the uniform hyperbolic conditions fails, f could still have an attractor, which could be taken to be Θ.

We take the control u to be a small perturbation of the parameters of the nominal dynamics f. More specifically, the nominal and perturbed dynamics are, respectively,

$$\theta(k+1) = f(\theta(k), 0) + v_1(k), \quad \text{with } \theta(0) = \theta_o, \tag{2}$$

$$\varphi(k+1) = f(\varphi(k), u(k)) + v_2(k), \quad \text{with } \varphi(0) = \varphi_o, \tag{3}$$

where

1.
$$f \in C^1(\mathbb{R}^n \times \mathbb{R}^m, \mathbb{R}^n); \tag{4}$$

2. $f(\Theta, 0) \subset \Theta$, i.e., Θ is f-invariant, and $f(\cdot, 0) : \Theta \to \Theta$;
3. Θ is a compact subset of $\mathbb{R}^n$.

Here $\{\theta(k) : k \ge 0\}$ is the desired trajectory, and $\varphi(k)$ is the state of the system under control. The exogenous inputs $v_1(k)$ and $v_2(k)$ are typically small with $\theta(k+1) \in \Theta$. The purpose of v_1 is to allow the desired trajectory to occasionally jump from a point on one orbit to a nearby point on another orbit [9]. On the other hand, v_2 is to allow for some modeling inaccuracies. At time k, it is assumed that both $\theta(k)$ and $\varphi(k)$ are known. The basic objective is to find a control u such that, when $v_1 = v_2 = 0$ and for $|\theta(0) - \varphi(0)|$ small enough, we have $\lim_{k\to\infty} |\varphi(k) - \theta(k)| = 0$.

A distinguishing feature of the present approach is that the tracking controller takes the form of a *spatially varying* gain $F : \Theta \to \mathbb{R}^{m\times n}$, guaranteed to be continuous under suitable conditions. As the first and most generic application, given

an *arbitrary* desired trajectory $\{\theta(k) : k = 0, \dots\}$, evaluating the controller F along the trajectory $\{\theta(k) : k = 0, \dots\}$ yields the time-varying controller $F_{\theta(k)}$ that makes the nonlinear system $\varphi(k+1) = f(\varphi(k), F_{\theta(k)}(\varphi(k) - \theta(k)))$ asymptotically track $\theta(k+1) = f(\theta(k))$. More importantly, the *globally* defined controller F becomes fully motivated in those specialized applications where there is a need to quickly adapt the tracking controller to a new reference trajectory without recomputing a new time-varying controller along the new trajectory [9], [13], [20], [21].

If $v_1 = v_2 = 0$, then stability of the closed-loop system, which implies asymptotic tracking, is guaranteed if $|\varphi(0) - \theta(0)| < R_{Capture}$, where $R_{Capture} > 0$. If $v_1 \neq 0$ and/or $v_2 \neq 0$, then asymptotic tracking can still be guaranteed if $\|v_1 - v_2\|_{l_\infty}$ and $|\varphi(0) - \theta(0)|$ are small enough and $v_1(k) - v_2(k)$ is intermittent enough. If $v_1 - v_2$ is persistent, then one cannot expect asymptotic tracking; however, under suitable conditions, the gain $\frac{\|\theta - \varphi\|_{\ell_\infty}}{\|v_1 - v_2\|_{\ell_\infty}}$ can easily be shown to be bounded [7]. Besides, the effect of the model uncertainty v_2 can be minimized using standard H^∞ methods. Therefore we shall not pursue investigation of the effects of v_1, v_2 any further.

The tracking controller design relies on linearizing the tracking error dynamics as follows. Define the tracking error

$$x(k) := \varphi(k) - \theta(k).$$

Then

$$x(k+1) = f(\varphi(k), u(k)) - f(\theta(k), 0).$$

The first-degree Taylor approximation of $f(\varphi(k), u(k))$ around $\varphi(k) = \theta(k)$ and $u(k) = 0$ yields

$$\begin{aligned} f(\varphi(k), u(k)) = {} & f(\theta(k), 0) + A_{\theta(k)}(\varphi(k) - \theta(k)) \\ & + B_{2_{\theta(k)}} u(k) + \eta(x(k), u(k), \theta(k)), \end{aligned}$$

where

$$A_\theta := \frac{\partial f}{\partial \theta}(\theta, 0), \quad B_{2_\theta} := \frac{\partial f}{\partial u}(\theta, 0), \tag{5}$$

and $\eta(x(k), u(k), \theta(k))$ accounts for nonlinear terms. Thus

$$x(k+1) = A_{\theta(k)} x(k) + B_{2_{\theta(k)}} u(k) + \eta(x(k), u(k), \theta(k)). \tag{6}$$

Since $f \in C^1$, η can be decomposed as

$$\eta(x(k), u(k), \theta(k)) = \eta_x(x(k), u(k), \theta(k))\, x(k) + \eta_u(x(k), u(k), \theta(k))\, u(k), \tag{7}$$

where

$$\eta_x(x, u, \theta)_{i,j} = \int_0^1 \left(\frac{\partial f_i}{\partial x_j}(tx + \theta, tu) - \frac{\partial f_i}{\partial x_j}(\theta, 0) \right) dt \tag{8}$$

and

$$\eta_u(x, u, \theta)_{i,j} = \int_0^1 \left(\frac{\partial f_i}{\partial u_j}(tx + \theta, tu) - \frac{\partial f_i}{\partial u_j}(\theta, 0) \right) dt. \tag{9}$$

Since $f \in C^1$ and Θ is compact, if x and u are bounded, then $\frac{\partial f_i}{\partial x_j}(tx+\theta, tu) - \frac{\partial f_i}{\partial x_j}(\theta, 0)$ is uniformly continuous. In particular, for any $\varepsilon > 0$ there is a $\delta > 0$ such that, if $|x|, |u| < \delta$, then $|\frac{\partial f_i}{\partial x_j}(tx+\theta, tu) - \frac{\partial f_i}{\partial x_j}(\theta, 0)| < \varepsilon$. Therefore,

$$\lim_{\bar{x}\to 0, \bar{u}\to 0} \sup\{\|\eta_x(x,u,\theta)\| : |x| < \bar{x},\ |u| < \bar{u},\ \theta \in \Theta\} = 0 \tag{10}$$

and

$$\lim_{\bar{x}\to 0, \bar{u}\to 0} \sup\{\|\eta_u(x,u,\theta)\| : |x| < \bar{x},\ |u| < \bar{u},\ \theta \in \Theta\} = 0. \tag{11}$$

If u and x are small, we can approximate the error dynamics as

$$\begin{aligned} x(k+1) &= A_{\theta(k)} x(k) + B_{2_{\theta(k)}} u(k), \\ \theta(k+1) &= f(\theta(k), 0). \end{aligned} \tag{12}$$

This system is linear in the tracking error x, but the coefficient matrices A and B vary (generally in a nonlinear way) as θ varies. Since $\theta(k)$ varies according to (2), the system described in (12) is an *LDV* system. Before controllers can be developed for such systems, linear systems with dynamically varying parameters must be formalized.

3. Linear dynamically varying systems and LQ control. Motivated by the preceding considerations, a general LDV system is defined as

$$\begin{bmatrix} x(k+1) \\ z(k) \end{bmatrix} = \begin{bmatrix} A_{\theta(k)} & B_{1_{\theta(k)}} & B_{2_{\theta(k)}} \\ C_{\theta(k)} & D_{1_{\theta(k)}} & D_{2_{\theta(k)}} \end{bmatrix} \begin{bmatrix} x(k) \\ w(k) \\ u(k) \end{bmatrix}, \tag{13}$$

$$\begin{aligned} \theta(k+1) &= f(\theta(k)), \\ \text{with } \theta(0) &= \theta_o \text{ and } x(0) = x_o, \end{aligned} \tag{14}$$

subject to the following general conditions:

1. $\Theta \subset \mathbb{R}^n$ is compact and $f : \Theta \to \Theta$ is a continuous function.
2. $A : \Theta \to \mathbb{R}^{n\times n}$, $B_1 : \Theta \to \mathbb{R}^{n\times l}$, $B_2 : \Theta \to \mathbb{R}^{n\times m}$, $C : \Theta \to \mathbb{R}^{p\times n}$, $D_1 : \Theta \to \mathbb{R}^{p\times l}$, and $D_2 : \Theta \to \mathbb{R}^{p\times m}$ are functions that need not be continuous.

In the above, $\theta(k) \in \Theta$ is the state of the dynamic system, $x(k) \in \mathbb{R}^n$ is the state of the linear system, $u(k) \in \mathbb{R}^m$ is the control input, $w(k) \in \mathbb{R}^l$ is the disturbance input, and $z(k) \in \mathbb{R}^p$ is the output to be controlled.

It is often assumed that the system coefficient matrices A, B_1, B_2, C, D_1, and D_2 are continuous. We will refer to such systems as *continuous* LDV systems. In section 2 it was assumed that $f \in C^1$, and since A and B are matrices of partial derivatives of f, A and B are indeed continuous. Thus the tracking error system associated with (2) and (3) can be approximated by a *continuous* LDV system. However, if a feedback $F : \Theta \to \mathbb{R}^{m\times n}$ is used to stabilize a continuous LDV system, then the resulting closed-loop system is a continuous LDV system if and only if F is continuous. Although this paper will focus on stabilizing continuous LDV systems, we cannot assume a priori that the feedback is continuous. Therefore the definition of an LDV system must allow for possibly discontinuous coefficient matrices.

Since an LDV system is an uncountable collection of linear time-varying systems indexed by $\theta(0)$, the concept of stability is slightly more complex in the dynamically varying case than it is in the time-varying case.

DEFINITION 3.1. *The LDV system* (13) *is uniformly exponentially stable if, for* $u(k) = 0$ *and* $w(k) = 0$, *there exist an* $\alpha \in [0,1)$ *and a* $\beta < \infty$ *such that, for all* $\theta(0) \in \Theta$,

$$|x(k)| \leq \beta \alpha^k |x(0)|.$$

System (13) *is exponentially stable if, for* $u(k) = 0$, $w(k) = 0$, *and for each* $\theta(0) \in \Theta$, *there exist an* $\alpha_{\theta(0)} \in [0,1)$ *and a* $\beta_{\theta(0)} < \infty$ *such that, for all* $x(j)$ *and* $j \leq k$,

$$|x(k)| \leq \beta_{\theta(0)} \alpha_{\theta(0)}^{k-j} |x(j)|.$$

System (13) *is asymptotically stable if, for* $u(k) = 0$, $w(k) = 0$, *any* $|x(0)| < \infty$, *and any* $\theta(0) \in \Theta$,

$$|x(k)| \to 0 \quad \text{as} \quad k \to \infty.$$

Note that an exponentially stable system is stable uniformly in time k, but not necessarily uniformly in the initial condition $\theta(0)$. That is, along any given positive trajectory $\left\{f^k(\theta(0)) : k \geq 0\right\}$, an exponentially stable system is (uniformly in time) exponentially stable; however, if $\{\theta(0)_i : i \geq 0\}$ is a convergent sequence, with $\theta(0) = \lim_{i\to\infty} \theta(0)_i$, it is possible that $\alpha_{\theta(0)_i} \to 1$ while $\alpha_{\theta(0)} < 1$, in which case the system is exponentially stable, but not $\theta(0)$-uniformly exponentially stable. To emphasize the difference between exponential and uniformly exponential stability, exponential stability will occasionally be referred to as uniform *in time* exponential stability.

In the case of continuous LDV systems, asymptotic, exponential, and uniform exponential stability are equivalent (Proposition 2 in [8]). Since uniformly exponentially stable systems are inherently more robust than exponentially stable systems, it is preferable to remain within the confines of continuous LDV systems. Thus, when synthesizing a feedback for controlling a continuous LDV system, it is important to ensure that the feedback is not only asymptotically stabilizing but also continuous. However, to maintain generality, an LDV system is considered stabilizable if there exists an exponentially stabilizing feedback, that is, the following holds.

DEFINITION 3.2. *System* (13) *is stabilizable if there exists a (not necessarily continuous) function* $F : \mathbb{Z} \times \Theta \to \mathbb{R}^{m\times n}$ *with bound* $\overline{F}_{\theta(0)} < \infty$ *such that, for all* $\theta(0) \in \Theta$ *and for all* $k \geq 0$, *we have* $\left\|F_{\theta(0)}(k)\right\| \leq \overline{F}_{\theta(0)}$, *and the system*

$$\begin{aligned} x(k+1) &= \left(A_{\theta(k)} + B_{2_{\theta(k)}} F_{\theta(0)}(k)\right) x(k), \\ \theta(k) &= f^k(\theta(0)) \end{aligned}$$

is exponentially stable. That is, there exist $\alpha_{\theta(0)} \in [0,1)$ *and* $\beta_{\theta(0)} < \infty$ *such that, for any* $\theta(0) \in \Theta$, *there exists a time-varying, bounded feedback* $F_{\theta(0)}(k)$, *which may depend on* $\theta(0)$, *such that*

$$\left\| \prod_{i=j}^{k-1} \left(A_{f^i(\theta(0))} + B_{2_{f^i(\theta(0))}} F_{\theta(0)}(i) \right) \right\| \leq \beta_{\theta(0)} \alpha_{\theta(0)}^{k-j},$$

where the factors of the matrix product are taken in the proper order.

Therefore, along every trajectory $\left\{f^k(\theta(0)) : k \geq 0\right\}$, the time-varying system is (uniformly in time) exponentially stabilizable by means of a function F which, as defined in Definition 3.2, depends on the initial condition $\theta(0)$. In this sense, the

control is not quite "closed-loop," and more importantly there are no assumptions about the global properties of the feedback F. In particular, the feedback may not be a continuous nor even a uniformly bounded function of $\theta(0)$. However, in the case of continuous LDV systems, it was shown in [8] that a stabilizable system has a continuous, uniformly exponentially stabilizing feedback $F : \Theta \to \mathbb{R}^{m\times n}$. In this case, the feedback gain takes the form $F_{\theta(k)}$ and does not depend on the initial condition $\theta(0)$, but only the current state $\theta(k)$. Hence the controller is "closed-loop."

The dual concept of detectability has two versions. The first one is *uniform* detectability.

DEFINITION 3.3. *System* (13) *is uniformly detectable if there exists a (not necessarily continuous) function* $H : \Theta \to \mathbb{R}^{n\times p}$ *with uniform bound* $\bar{H} < \infty$ *such that, for all* $\theta \in \Theta$, *we have* $\|H_\theta\| \le \bar{H}$, *and the system*

$$\begin{aligned} x(k+1) &= \left(A_{\theta(k)} + H_{\theta(k)}C_{\theta(k)}\right)x(k), \\ \theta(k) &= f^k(\theta(0)) \end{aligned}$$

is uniformly exponentially stable. That is, there exist an $\alpha_d \in [0,1)$ *and a* $\beta_d < \infty$ *such that, for all* $\theta(0) \in \Theta$,

$$\|x(k)\| \le \beta_d \alpha_d^k \|x(0)\|.$$

DEFINITION 3.4. *System* (13) *is detectable if there exists a (not necessarily continuous), function* $H : \mathbb{Z} \times \Theta \to \mathbb{R}^{n\times p}$ *with bound* $\bar{H}_{\theta(0)} < \infty$ *such that, for all* $\theta(0) \in \Theta$ *and all* k, *we have* $\left\|H_{\theta(0)}(k)\right\| \le \overline{H}_{\theta(0)} < \infty$, *and the system*

$$\begin{aligned} x(k+1) &= \left(A_{\theta(k)} + H_{\theta(0)}(k)\, C_{\theta(k)}\right)x(k), \\ \theta(k) &= f^k(\theta(0)) \end{aligned}$$

is exponentially stable.

If f is invertible, the LDV system has an adjoint system running backwards in time. If a continuous LDV is detectable and f is invertible, then the adjoint system is stabilizable. It is easily shown that this implies that the adjoint LDV is in fact *uniformly* stabilizable and therefore that the LDV system is *uniformly* detectable. Thus, if f is invertible and the LDV system is continuous, then uniform detectability and detectability are equivalent. Although stabilizability and uniform detectability are slightly asymmetric, to avoid putting extra assumptions on f, stabilizable and uniformly detectable continuous LDV systems will be considered.

Since stabilizability only depends on A, B_2, and f, we will say that the triple (A, B_2, f) is stabilizable to mean that system (13) is stabilizable. Similarly, we say that the triple (A, C, f) is uniformly detectable to mean that system (13) is uniformly detectable.

Since an LDV system is a collection of time-varying systems, the following time-varying Lyapunov stability theorem is useful.

THEOREM 3.5. *Assume that system* (13) *is uniformly detectable and* $w \equiv 0$. *Then there exist an* $\alpha_{\theta_o} \in [0,1)$ *and a* $\beta_{\theta_o} < \infty$ *such that, for* $\theta(0) = \theta_o$ *and any* $x(j) \in \mathbb{R}^n$,

$$|x(k)| \le \beta_{\theta_o} \alpha_{\theta_o}^{k-j} |x(j)|$$

if and only if there exists a sequence $\left\{X_{f^k(\theta_o)} : k \ge 0\right\}$ *with bound* $\bar{X} : \Theta \to \mathbb{R}$ *such that* $\left\|X_{f^k(\theta_o)}\right\| \le \bar{X}_{\theta_o} < \infty$, $X_{f^k(\theta_o)} = X'_{f^k(\theta_o)} \ge 0$, *and*

$$A'_{f^k(\theta_o)} X_{f^{k+1}(\theta_o)} A_{f^k(\theta_o)} - X_{f^k(\theta_o)} \le -C'_{f^k(\theta_o)} C_{f^k(\theta_o)}. \tag{15}$$

Furthermore, if (15) *is satisfied, then* α_{θ_o} *and* β_{θ_o} *can be taken to depend only on the bound* $\bar{X}_{\theta_o}$ *and on* α_d *and* β_d *in the definition of detectability.*

Proof. For $\theta(0)$ fixed, the system is a time-varying system. Thus the theorem is simply a statement about the stability of linear time-varying systems and can be found on page 41 in [18]. □

COROLLARY 3.6. *Assume that system* (13) *is uniformly detectable and* $w \equiv 0$. *Then there exist an* $\alpha \in [0,1)$ *and a* $\beta < \infty$ *such that*

$$|x(k)| \leq \beta \alpha^k |x(0)|$$

if and only if there exists a uniformly bounded function $X : \Theta \to \mathbb{R}^{n \times n}$ *with* $X_\theta = X_\theta' \geq 0$ *such that, for all* $\theta_o \in \Theta$,

$$A'_{f^k(\theta_o)} X_{f^{k+1}(\theta_o)} A_{f^k(\theta_o)} - X_{f^k(\theta_o)} \leq -C'_{f^k(\theta_o)} C_{f^k(\theta_o)}. \tag{16}$$

Proof. Since X_θ is uniformly bounded and the system is uniformly detectable, Theorem 3.5 can be applied at each $\theta_o \in \Theta$. □

The main result of [8] is the following.

THEOREM 3.7. *Suppose that these conditions hold.*

1. $f : \Theta \to \Theta$ *is continuous and* Θ *is compact.*
2. *The functions* A, B_2, C, D_2 *are continuous.*
3. $D'_{2_\theta} D_{2_\theta} > 0$.
4. $C'_\theta D_{2_\theta} = 0$ *for all* $\theta \in \Theta$, *and* (A, C, f) *is uniformly detectable.*

Then the triple (A, B_2, f) *is stabilizable if and only if there exists a unique, uniformly bounded solution* $X_2 : \Theta \to \mathbb{R}^{n \times n}$ *such that*

1. X_2 *satisfies the FARE*

$$\begin{aligned} X_{2_\theta} &= A'_\theta X_{2_{f(\theta)}} A_\theta \\ &\quad - A'_\theta X_{2_{f(\theta)}} B_{2_\theta} \left(D'_{2_\theta} D_{2_\theta} + B'_{2_\theta} X_{2_{f(\theta)}} B_{2_\theta} \right)^{-1} B'_{2_\theta} X_{2_{f(\theta)}} A_\theta + C'_\theta C_\theta; \end{aligned} \tag{17}$$

2. $X_{2_\theta} \geq 0$.

In this case, the closed-loop control

$$u_{LQ}(k) := -\left(D'_{2_{\theta(k)}} D_{2_{\theta(k)}} + B'_{2_{\theta(k)}} X_{2_{f(\theta(k))}} B_{2_{\theta(k)}} \right)^{-1} B'_{2_{\theta(k)}} X_{2_{f(\theta(k))}} A_{\theta(k)} x(k) \tag{18}$$

uniformly exponentially stabilizes system (13). *Moreover, for* $|x(0)| < \infty$ *and* $w \equiv 0$,

$$x'(0) X_{2_{\theta(0)}} x(0) = \inf \left\{ \sum_{k=0}^{\infty} |z(k)|^2 : u \in l_2 \right\}, \tag{19}$$

where the infimum is attained for $u = u_{LQ}$. *Furthermore,* X_2 *is a uniformly continuous function. Finally, if* $X_{2_\theta}(k, N+1)$ *solves the finite horizon Riccati equation, i.e.,*

$$\begin{aligned} X_{2_\theta}(k, N+1) &= A'_{f^k(\theta)} X_{2_\theta}(k+1, N+1) A_{f^k(\theta)} + C'_{f^k(\theta)} C_{f^k(\theta)} \\ &\quad - A'_{f^k(\theta)} X_{2_\theta}(k+1, N+1) B_{2_{f^k(\theta)}} \\ &\quad \times \left(D'_{2_{f^k(\theta)}} D_{2_{f^k(\theta)}} + B'_{2_{f^k(\theta)}} X_{2_\theta}(k+1, N+1) B_{2_{f^k(\theta)}} \right)^{-1} \\ &\quad \times B'_{2_{f^k(\theta)}} X_{2_\theta}(k+1, N+1) A_{f^k(\theta)} \end{aligned} \tag{20}$$

with

$$X_{2_\theta}(N+1,N+1) = C'_{f^{N+1}(\theta)} C_{f^{N+1}(\theta)},$$

then $X_{2_\theta}(0,N+1) \to X_{2_\theta}$ *uniformly in* θ.

4. Linear dynamically varying H^∞ control. In the following, the H^∞ control problem for LDV systems of the general form (13) will be formulated and the solution will be provided. There are two related problems.

The first is the finite horizon problem. For all $\theta \in \Theta$, let the terminal weighting $X_\theta(N+1,N+1) \geq 0$ be given. The objective in this problem is to find a controller F_u such that, if

$$u(k) = F_{u_{\theta_o}}(k,N+1) \begin{bmatrix} x(k) \\ w(k) \end{bmatrix} \quad \text{for } k \leq N,$$

then we have the following.

Objective A. For $x(0) = 0$ there exists an $\varepsilon > 0$ such that, for $w \in l_2[0,N]$ and $\theta_o \in \Theta$,

$$\|z\|^2_{[0,N]} - \gamma^2 \|w\|^2_{[0,N]} + x'(N+1) X_{\theta_o}(N+1,N+1) x(N+1) \leq -\varepsilon \|w\|^2_{[0,N]}.$$

The second problem is the infinite horizon problem, where the objective is to find a uniformly exponentially stabilizing controller F_u such that, if

$$u(k) = F_{u_{\theta_o}}(k) \begin{bmatrix} x(k) \\ w(k) \end{bmatrix},$$

then our objective becomes the following.

Objective B. For $x(0) = 0$ there exists an $\varepsilon > 0$ such that, for $w \in l_2$ and $\theta_o \in \Theta$,

$$\|z\|^2_{l_2} - \gamma^2 \|w\|^2_{l_2} \leq -\varepsilon \|w\|^2_{l_2},$$

and if $w = 0$ and $x(0) \neq 0$, then $x(k) \to 0$.

If Objective B is achieved, then

$$\frac{\|z\|_{l_2}}{\|w\|_{l_2}} < \gamma.$$

It will be shown that the solution to Objective B is the limit as $N \to \infty$ of solutions to Objective A.

4.1. Finite horizon full information controller. For notational simplicity, we define

$$\begin{bmatrix} A_\theta & \bar{B}_\theta \\ \bar{C}_\theta & \bar{D}_\theta \end{bmatrix} := \begin{bmatrix} A_\theta & B_{1_\theta} & B_{2_\theta} \\ C_\theta & D_{1_\theta} & D_{2_\theta} \\ 0 & I_l & 0 \end{bmatrix} \quad \textit{and} \quad J =: \begin{bmatrix} I_p & 0 \\ 0 & -\gamma^2 I_l \end{bmatrix}.$$

Let $X_{\theta_o}(N+1,N+1) \geq 0$ be given. In a recursive manner, define

$$\begin{aligned} X_{\theta_o}(k,N+1) = {} & A'_{f^k(\theta_o)} X_{\theta_o}(k+1,N+1) A_{f^k(\theta_o)} + C'_{f^k(\theta_o)} C_{f^k(\theta_o)} \\ & - L_{\theta_o}(k,N+1)' R^{-1}_{\theta_o}(k,N+1) L_{\theta_o}(k,N+1), \end{aligned} \tag{21}$$

where

$$R_{\theta_o}(k,N+1) := \bar{D}'_{f^k(\theta_o)} J \bar{D}_{f^k(\theta_o)} + \bar{B}'_{f^k(\theta_o)} X_{\theta_o}(k+1,N+1)\, \bar{B}_{f^k(\theta_o)}, \tag{22}$$

$$L_{\theta_o}(k,N+1) := \bar{D}'_{f^k(\theta_o)} J \bar{C}_{f^k(\theta_o)} + \bar{B}'_{f^k(\theta_o)} X_{\theta_o}(k+1,N+1)\, A_{f^k(\theta_o)}. \tag{23}$$

We partition $R = \left[\begin{smallmatrix} R_1 & R_2' \\ R_2 & R_3 \end{smallmatrix}\right]$ and $L = \left[\begin{smallmatrix} L_1 \\ L_2 \end{smallmatrix}\right]$ such that $R_3 \in \mathbb{R}^{m\times m}$ and $L_2 \in \mathbb{R}^{m\times n}$. With the assumption that $R_{\theta_o}(k,N+1)$ is nonsingular, the Schur decomposition yields

$$\begin{aligned} &R_{\theta_o}(k,N+1) \\ &= \begin{bmatrix} I & R'_{2_{\theta_o}}(k,N+1,\theta_o)\, R^{-1}_{3_{\theta_o}}(k,N+1) \\ 0 & I \end{bmatrix} \begin{bmatrix} \nabla_{\theta_o}(k,N+1) & 0 \\ 0 & R_{3_{\theta_o}}(k,N+1) \end{bmatrix} \\ &\times \begin{bmatrix} I & 0 \\ R^{-1}_{3_{\theta_o}}(k,N+1)\, R_{2_{\theta_o}}(k,N+1) & I \end{bmatrix}, \end{aligned}$$

where

$$\nabla_{\theta_o}(k,N+1) := R_{1_{\theta_o}}(k,N+1) - R'_{2_{\theta_o}}(k,N+1)\, R^{-1}_{3_{\theta_o}}(k,N+1)\, R_{2_{\theta_o}}(k,N+1). \tag{24}$$

Note that since $R_{3_{\theta_o}}(k,N+1) = D'_{2_{f^k(\theta_o)}} D_{2_{f^k(\theta_o)}} + B'_{2_{f^k(\theta_o)}} X_{\theta_o}(k,N+1)\, B_{2_{f^k(\theta_o)}}$ and $D'_{2_{f^k(\theta_o)}} D_{2_{f^k(\theta_o)}} > 0$, we have

$$R_{3_{\theta_o}}(k,N+1) > 0 \tag{25}$$

whenever $X_{\theta_o}(k,N+1) \geq 0$. Hence, if

$$X_{\theta_o}(k,N+1) \geq 0, \tag{26}$$

$$\nabla_{\theta_o}(k,N+1) \leq -\varrho I, \tag{27}$$

then $R_{\theta_o}(k,N+1)$ is nonsingular.

For X, R, and ∇ defined as above, it is possible to show by completion of squares (see [17, p. 485]) that for all $x(k)$ and all $u, w \in l_2[0,N]$ we have

$$\begin{aligned} &\|z\|^2_{[k,N]} - \gamma^2 \|w\|^2_{[k,N]} + x'(N+1)\, X_{\theta_o}(N+1,N+1)\, x(N+1) \\ &= x'(k)\, X_{\theta_o}(k,N+1)\, x(k) \\ &+ \sum_{j=k}^{N} (u(j) - u_N(j))'\, R_{3_{\theta_o}}(j,N+1)\, (u(j) - u_N(j)) \\ &+ \sum_{j=k}^{N} (w(j) - w_N(j))'\, \nabla_{\theta_o}(j,N+1)\, (w(j) - w_N(j)), \end{aligned} \tag{28}$$

with

$$\begin{aligned} w_N(k) &:= -\nabla^{-1}_{\theta_o}(k,N+1)\, L_{\nabla_{\theta_o}}(k,N+1)\, x(k), \\ u_N(k) &:= -R^{-1}_{3_{\theta_o}}(k,N+1) \begin{bmatrix} L_{2_{\theta_o}}(k,N+1) & R_{2_{\theta_o}}(k,N+1) \end{bmatrix} \begin{bmatrix} x(k) \\ w(k) \end{bmatrix}, \end{aligned} \tag{29}$$

and where

(30)
$$L_{\nabla_{\theta_o}}(k,N+1) := L_{1_{\theta_o}}(k,N+1) - R'_{2_{\theta_o}}(k,N+1)\, R^{-1}_{3_{\theta_o}}(k,N+1)\, L_{2_{\theta_o}}(k,N+1).$$

From (25), (27), and (28), it is clear that, for $\theta(0) = \theta_o$,

(31)
$$x'_o X_{\theta_o}(0,N+1)\, x_o = \sup_{w\in l_2[0,N]} \inf_{u\in l_2[0,N]} \left\{ \|z\|^2_{[0,N]} - \gamma^2 \|w\|^2_{[0,N]} + x'(N+1)\, X_{\theta_o}(N+1,N+1)\, x(N+1) \right\}.$$

The above is summarized by the following theorem, which is a straightforward extension of [17, p. 484].

THEOREM 4.1. *Let us suppose that* $D'_{2_{f^k(\theta_o)}} D_{2_{f^k(\theta_o)}} > 0$ *for all* $k \le N$ *and* $X_{\theta_o}(N+1,N+1) \ge 0$. *In this case, there exists a causal full information control* $u(k) = F_{u_{\theta_o}}(k,N+1) \left[\begin{smallmatrix} x(k) \\ w(k) \end{smallmatrix} \right]$ *that satisfies Objective A if and only if, for* $0 \le k \le N+1$, *the following conditions hold:*

1. $X_{\theta_o}(k,N+1)$ *satisfies the time-varying Riccati recursion* (21).
2. *For some* $\varrho > 0$, (26) *and* (27) *hold.*

In this case, the control given by (29) *achieves Objective* A.

4.2. Infinite horizon full information controller. The second problem is the infinite horizon problem where the objective is to find a (uniformly in time) exponentially stabilizing controller F such that, if

$$u(k) = F_{\theta_o}(k) \begin{bmatrix} x(k) \\ w(k) \end{bmatrix},$$

then Objective B can be achieved. The following assumptions on system (13) are needed:

1. $f : \Theta \to \Theta$ is continuous and Θ is compact.
2. The system parameters A, B_1, B_2, C, D_1, and D_2 are matrix-valued continuous functions of θ.
3. $D'_{2_\theta} D_{2_\theta} > 0$ for all $\theta \in \Theta$.
4. For all $\theta \in \Theta$, we have $D'_{2_\theta} \begin{bmatrix} C_\theta & D_{1_\theta} \end{bmatrix} = 0$, and the triple (A, C, f) is uniformly detectable.
5. The triple (A, B_2, f) is stabilizable.

Assumption 4 is equivalent to the following.

4′ The triple $(A - B_2 (D'_2 D_2)^{-1} D'_2 C, (I - D_2 (D'_2 D_2)^{-1} D'_2)C, f\ )$ is uniformly detectable.

Indeed, if Assumption 4′ holds, then the feedback

$$u(k) = -\left(D'_{2_{f^k(\theta)}} D_{2_{f^k(\theta)}}\right)^{-1} D'_{2_{f^k(\theta)}} C_{f^k(\theta)} x(k) - \left(D'_{2_{f^k(\theta)}} D_{2_{f^k(\theta)}}\right)^{-1} D'_{2_{f^k(\theta)}} D_{1_{f^k(\theta)}} w(k) + r(k)$$

converts it to Assumption 4. Perhaps these assumptions could be weakened (for example, see [30]), but they are common.

The main result of the paper is the following.

THEOREM 4.2. *Suppose Assumptions* 1–5 *hold. There exists a (uniformly in time) exponentially stabilizing controller* $u(k) = F_{u_{\theta_o}}(k) \left[\begin{smallmatrix} x(k) \\ w(k) \end{smallmatrix}\right]$ *such that Objective* B *can be achieved if and only if there exists a uniformly bounded map* $X_\infty : \Theta \to \mathbb{R}^{n \times n}$ *such that the following hold:*

1. X_∞ *satisfies the FARE*

$$X_{\infty_\theta} = C'_\theta C_\theta + A'_\theta X_{\infty_{f(\theta)}} A_\theta - L'_\theta R_\theta^{-1} L_\theta, \tag{32}$$

where

$$\begin{aligned} R_\theta &:= \overline{D}'_\theta J \overline{D}_\theta + \bar{B}'_\theta X_{\infty_{f(\theta)}} \bar{B}_\theta, \\ L_\theta &:= \overline{D}'_\theta J \bar{C}_\theta + \bar{B}'_\theta X_{\infty_{f(\theta)}} A_\theta. \end{aligned} \tag{33}$$

2. *For some* $\varrho > 0$ *and all* $\theta \in \Theta$,

$$\begin{aligned} X_{\infty_\theta} &\geq 0, \\ \nabla_\theta &:= R_{1_\theta} - R'_{2_\theta} R_{3_\theta}^{-1} R_{2_\theta} \leq -\varrho I. \end{aligned} \tag{34}$$

3. *The closed-loop system*

$$x(k+1) = \left(A_{\theta(k)} - \bar{B}_{\theta(k)} R_{\theta(k)}^{-1} L_{\theta(k)} \right) x(k) \tag{35}$$

is uniformly exponentially stable.

In this case, the control

$$u_\infty(k) := -R_{3_{\theta(k)}}^{-1} \left[\begin{smallmatrix} L_{2_{\theta(k)}} & R_{2_{\theta(k)}} \end{smallmatrix}\right] \left[\begin{smallmatrix} x(k) \\ w(k) \end{smallmatrix}\right] \tag{36}$$

achieves Objective B, X_∞ *is continuous, and the closed-loop system with control* (36), *that is,*

$$\begin{aligned} x(k+1) = &\left(A_{f^k(\theta_o)} - B_{2_{f^k(\theta_o)}} R_{3_{f^k(\theta_o)}}^{-1} L_{2_{f^k(\theta_o)}} \right) x(k) \\ &+ \left(B_{1_{f^k(\theta_o)}} - B_{2_{f^k(\theta_o)}} R_{3_{f^k(\theta_o)}}^{-1} L_{2_{f^k(\theta_o)}} \right) w(k), \end{aligned}$$

is a uniformly (in θ*) exponentially stable system.*

The proof of this theorem is withheld until section 5. The proof entails the major difficulty of proving continuity relative to $\theta(0)$, an issue that does not exist in the traditional time-varying case of [18] and [26]. Even though our approach is inspired by [17], [30], and [31], the continuity issue of the LDV case makes it of interest in its own right.

The control $u(k)$ produced by Theorem 4.2 depends on $w(k)$. Since $w(k)$ is meant to model the linearization error (see section 6), it will likely depend on $u(k)$. Thus $u(k)$ and $w(k)$ are linked by some algebraic relationship, which may not be easily solved. The following shows how to find a control $u(k)$ that depends on the information $x(k)$ only. This type of control is referred to as *strictly causal*.

COROLLARY 4.3. *Suppose the assumptions of Theorem* 4.2 *hold and there exists a controller as described. Suppose also that* $R_{1_\theta} \leq -\varrho I$. *Then the above control can be taken to be strictly proper. In particular, the control*

$$\begin{aligned} u_*(k) := &- \left(R_{3_{\theta(k)}} - R_{2_{\theta(k)}} R_{1_{\theta(k)}}^{-1} R'_{2_{\theta(k)}} \right)^{-1} \\ &\times \left(L_{2_{\theta(k)}} - R_{2_{\theta(k)}} R_{1_{\theta(k)}}^{-1} L_{1_{\theta(k)}} \right) x(k) \\ = &-\Delta_{\theta(k)}^{-1} L_{\Delta_{\theta(k)}} x(k), \end{aligned}$$

where

$$\Delta_\theta := R_{3_\theta} - R_{2_\theta} R_{1_\theta}^{-1} R_{2_\theta}'$$

and

$$L_{\Delta_\theta} := L_{2_\theta} - R_{2_\theta} R_{1_\theta}^{-1} L_{1_\theta},$$

achieves Objective B.

Proof. This corollary follows as a minor variation of the proof of Theorem 4.2. □

Remark 1. The above results show the importance of the FARE (32). Solving a functional equation may be computationally difficult. However, in [8], [10] several methods for solving the FARE associated with a LQ objective were developed. These methods can easily be extended to solving the FARE (32). Furthermore, the stability of (35) can be checked via their respective FAREs.

Remark 2. The continuity of the solution to the FARE is crucial when numerically computing it. For example, suppose that $\Theta = [0, 1]$, that there exists a jump discontinuity at some point $0 \leq \rho \leq 1$, and that

$$X_\theta := \begin{cases} 0 \text{ if } \theta \leq \rho, \\ \delta \text{ otherwise.} \end{cases}$$

Consider the construction of $\hat{X}$, an approximation of X, with error $\varepsilon < \delta$, i.e., $\|X_\theta - \hat{X}_\theta\| < \varepsilon$ for all $\theta \in \Theta$. In general, the point ρ would be estimated via some search method. However, unless ρ is known *exactly* (which entails an infinite search), $\|X_{\theta^*} - \hat{X}_{\theta^*}\| > \varepsilon$ for some $\theta^* \in \Theta$. If θ^* is a fixed point of f, then $\|X_{f^k(\theta^*)} - \hat{X}_{f^k(\theta^*)}\| > \varepsilon$ for all k, and a similar problem occurs if θ^* is a recurrent point of f. In general, if $X : \Theta \to \mathbb{R}^{n \times n}$ is continuous and Θ is compact, then X can be estimated by its value at a finite number of points. If X is not continuous, such an estimate is not possible in general. It is this continuity issue, and hence the ability to numerically evaluate the Riccati solution, that is the main distinction between an LDV controller and a family of infinite horizon, time-varying controllers.

Remark 3. Another difference between an LDV controller and a family of infinite horizon, time-varying controllers is that the LDV controller guarantees that the closed-loop system is uniformly exponentially stable, whereas the family of time-varying controllers only guarantees stability along every trajectory $\{\theta(k) : k \geq 0\}$. One situation in which this distinction is important is noise rejection. For example, suppose that the signal w in system (13) is bounded as $\|w\|_{l_\infty} \leq \bar{w}$. Such a situation arises when the f in (2) is different from the f in (3). Then it follows from section 6 that the maximum allowable $\bar{w}$ depends on the parameters α_{θ_o} and β_{θ_o} in the definition of stability. Hence we write $\bar{w}_{\theta_o}$. Now suppose that the system is not *uniformly* exponentially stable, i.e., there exists a sequence $\{\theta(0)_i : i \geq 0\}$ such that either $\lim_{i \to \infty} \alpha_{\theta(0)_i} = 1$ or $\lim_{i \to \infty} \beta_{\theta(0)_i} = \infty$. In this case, even though $\bar{w}_{\theta_o} > 0$ for each θ_o, we have $\lim_{i \to \infty} \bar{w}_{\theta(0)_i} = 0$; that is, there is no positive bound on the noise that results in a stable system for all initial conditions θ_o.

5. Proof of main theorem.

5.1. Necessity.

In the subsequent discussion, we assume the following.

Assumption A. Assumptions 1–5 of Theorem 4.2 hold, and there exists a stabilizing controller that achieves Objective B.

Since (A, C, f) is uniformly detectable, $D_{2_\theta}' D_{2_\theta} > 0$, and (A, B_2, f) is stabilizable, the optimal stabilizing LDV linear-quadratic controller exists (Theorem 3.7). That is, there exists a unique, continuous, bounded function $X_2 : \Theta \to \mathbb{R}^{n \times n}$ such that $X_{2_\theta}' = X_{2_\theta} \geq 0$ solves (17). Furthermore, for $w \equiv 0$,

$$\inf_{u \in l_2} \|z\|_{l_2}^2 = x_o' X_{2_\theta} x_o, \tag{37}$$

and this infimum is attained for u given by (18).

Define $X_{\theta_o}(k, N+1)$ as in (21) with terminal cost $X_{2_{f^{N+1}(\theta_o)}}$. It will be shown (in Lemma 5.10) that

$$X_{\infty_\theta} = \lim_{N \to \infty} X_\theta(0, N+1) \tag{38}$$

provides a solution to (32) (see Lemma 5.6) such that system (35) is uniformly exponentially stable (see Lemma 5.9) and inequalities (34) are satisfied (see inequalities (59) and (60)). Furthermore, the convergence in (38) is uniform in θ, and hence X_{∞_θ} is a continuous function in θ (see Lemma 5.10) and the control given by (36) satisfies Objective B (see Lemma 5.4).

The proof of the following lemma follows from an easy adaptation of the arguments of Section B.2.3 of [17] and from [31].

LEMMA 5.1. *If Assumption* A *holds,* $X_{\theta_o}(k, N+1)$ *is given by* (21)*, and* $\nabla_{\theta_o}(k, N+1)$ *is given by* (24)*, then*

1. *for* $\theta_o = \theta(0) \in \Theta$*, all* $k \leq N+1$*, and* $N \geq 0$*, we have* $\nabla_{\theta_o}(k, N+1) \leq -\varrho I$ *and* $X_{\theta_o}(k, N+1) \geq 0$*;*
2. *for all* $\theta_o \in \Theta$*, there exists a* $\bar{X}_{\infty_{\theta_o}} < \infty$ *such that* $\|X_{\theta_o}(k, N+1)\| \leq \bar{X}_{\infty_{\theta_o}}$ *for all* $k \leq N+1$ *and all* $N \geq 0$*;*
3. $X_{\theta_o}(k, N+1)$ *is monotone increasing in* N*.*

The bound $\bar{X}_{\infty_{\theta_o}}$ depends on θ_o, so we cannot say that there exists a single bound on $X_{\theta_o}(k, N+1)$ for all $\theta_o \in \Theta$. Since Θ is compact, if $\bar{X}_\infty$ is continuous, then $\bar{X}_\infty$ is bounded. However, we have not yet shown that $\bar{X}_\infty$ is continuous.

For fixed θ_o, $X_{\theta_o}(k, N+1)$ exists, is bounded, and is nondecreasing in N. Thus

$$X_{\theta_o}(k) := \lim_{N \to \infty} X_{\theta_o}(k, N+1)$$

exists for $k < \infty$. Furthermore, $X_{\theta_o}(k)$ solves

$$X_{\theta_o}(k) = A_{f^k(\theta_o)}' X_{\theta_o}(k+1) A_{f^k(\theta_o)} + C_{f^k(\theta_o)}' C_{f^k(\theta_o)} - L_{\theta_o}'(k) R_{\theta_o}^{-1}(k) L_{\theta_o}(k), \tag{39}$$

where

$$R_{\theta_o}(k) := \overline{D}_{f^k(\theta_o)}' J \overline{D}_{f^k(\theta_o)} + \bar{B}_{f^k(\theta_o)}' X_{\theta_o}(k+1) \bar{B}_{f^k(\theta_o)},$$
$$L_{\theta_o}(k) := \overline{D}_{f^k(\theta_o)}' J \bar{C}_{f^k(\theta_o)} + \bar{B}_{f^k(\theta_o)}' X_{\theta_o}(k+1) A_{f^k(\theta_o)}.$$

This is simply the Riccati equation associated with the infinite horizon, time-varying H^∞ control problem. Note that, since $X_{\theta_o}(k, N+1) \geq 0$,

$$X_{\theta_o}(k) \geq 0. \tag{40}$$

Next, since $X_{\theta_o}(k, N+1)$ converges, $\nabla_{\theta_o}(k) := \lim_{N \to \infty} \nabla_{\theta_o}(k, N+1)$ exists. Furthermore, $\nabla_{\theta_o}(k, N+1) \leq -\varrho I$ (from the finite horizon problem) implies that $\nabla_{\theta_o}(k)$

$\leq -\varrho I$. Furthermore, since $X_{\theta_o}(k)$ is bounded and $D'_{2_\theta} D_{2_\theta} > 0$ for all $\theta \in \Theta$, it is clear from (24) that $\nabla_{\theta_o}(k)$ is bounded from below. Hence

$$-\infty < \nabla_{\theta_o}(k) \leq -\varrho I. \tag{41}$$

Similarly, define $L_{\nabla_{\theta_o}}(k)$ as the limit of (30).

Next define

$$u_\infty(k) := F_{u_{\theta_o}}(k) \begin{bmatrix} x(k) \\ w(k) \end{bmatrix} := -R_{3_{\theta_o}}^{-1}(k) \begin{bmatrix} L_{2_{\theta_o}}(k) & R_{2_{\theta_o}}(k) \end{bmatrix} \begin{bmatrix} x(k) \\ w(k) \end{bmatrix} \tag{42}$$

and

$$w_\infty(k) := F_{w_{\theta_o}}(k)\, x(k) := -\nabla_{\theta_o}^{-1}(k)\, L_{\nabla_{\theta_o}}(k)\, x(k). \tag{43}$$

It will be shown that, with $\theta(0) = \theta_o$, (42) is the best control and (43) is the worst disturbance (in the sense of Objective B).

LEMMA 5.2. *For $w = 0$, the control $u(k) = u_\infty(k)$ given by (42) makes the closed-loop system $x(k+1) = A_{u_{\theta_o}}(k)x(k)$, where*

$$A_{u_{\theta_o}}(k) := A_{\theta(k)} - B_{2_{\theta(k)}} R_{3_{\theta_o}}^{-1}(k)\, L_{2_{\theta_o}}(k), \tag{44}$$

exponentially stable.

Proof. Since $u_\infty(k) = -R_{3_{\theta_o}}^{-1}(k)\, L_{2_{\theta_o}}(k)\, x(k) - R_{3_{\theta_o}}^{-1}(k)\, R_{2_{\theta_o}}(k)\, w(k)$, the closed-loop system with $w = 0$ and $u = u_\infty$ is

$$\begin{aligned} x(k+1) &= \left(A_{\theta(k)} - B_{2_{\theta(k)}} R_{3_{\theta_o}}^{-1}(k)\, L_{2_{\theta_o}}(k)\right) x(k) \\ &= A_{u_{\theta_o}}(k)\, x(k). \end{aligned}$$

Set $\Gamma_{\theta_o}(k) = X_{\theta_o}(k) - X_{2_{f^k(\theta_o)}}$. By Lemma 5.1, $X_{\theta_o}(k, N+1) \geq X_{\theta_o}(k,k) = X_{2_{f^k(\theta_o)}}$. Thus $X_{\theta_o}(k) = \lim_{N\to\infty} X_{\theta_o}(k, N+1) \geq X_{2_{f^k(\theta_o)}}$ and $\Gamma_{\theta_o}(k) \geq 0$. It is possible to show (see [17, equation (B.2.39)]) that

$$\begin{aligned} \Gamma_{\theta_o}(k) \geq{}& A'_{u_{\theta_o}}(k)\, \Gamma_{\theta_o}(k+1)\, A_{u_{\theta_o}}(k) \\ &+ A'_{u_{\theta_o}}(k)\, \Gamma_{\theta_o}(k+1)\, B_{2_{\theta(k)}} \\ &\times \left(R_{3_{\theta_o}}(k) - B'_{2_{\theta(k)}} \Gamma_{\theta_o}(k+1)\, B_{2_{\theta(k)}}\right)^{-1} B'_{2_{\theta(k)}} \Gamma_{\theta_o}(k+1)\, A_{u_{\theta_o}}(k). \end{aligned} \tag{45}$$

However,

$$A_\theta - B_{2_\theta} \left(D'_{2_\theta} D_{2_\theta} + B'_{2_\theta} X_{2_{f(\theta)}} B_{2_\theta}\right)^{-1} B'_{2_\theta} X_{2_{f(\theta)}} A_\theta$$

is the closed-loop system if the LQ feedback is used, that is, if $w = 0$ and $u = u_{LQ}$, where u_{LQ} is given by (18). After some manipulation, we find

$$\begin{aligned} &A_{f^k(\theta_o)} - B_{2_{f^k(\theta_o)}} \left(D'_{2_{f^k(\theta_o)}} D_{2_{f^k(\theta_o)}} + B'_{2_{f^k(\theta_o)}} X_{2_{f^{k+1}(\theta_o)}} B_{2_{f^k(\theta_o)}}\right)^{-1} \\ &\quad \times B'_{2_{f^k(\theta_o)}} X_{2_{f^{k+1}(\theta_o)}} A_{f^k(\theta_o)} \\ &= A_{u_{\theta_o}}(k) \\ &+ B_{2_{f^k(\theta_o)}} \left(R_{3_{\theta_o}}(k) - B'_{2_{f^k(\theta_o)}} \Gamma_{\theta_o}(k+1)\, B_{2_{f^k(\theta_o)}}\right)^{-1} B'_{2_{f^k(\theta_o)}} \Gamma_{\theta_o}(k+1)\, A_{u_{\theta_o}}(k) \end{aligned}$$

and

$$B_{2_{f^k(\theta)}}\left(R_{3_{\theta_o}}(k)-B'_{2_{f^k(\theta)}}\Gamma_{\theta_o}(k+1)B_{2_{f^k(\theta)}}\right)^{-1}$$
$$=B_{2_{f^k(\theta)}}\left(D'_{2_{f^k(\theta)}}D_{2_{f^k(\theta)}}+B'_{2_{f^k(\theta)}}X_{2_{f^{k+1}(\theta_o)}}B_{2_{f^k(\theta)}}\right)^{-1},$$

which is bounded since $D'_{2_\theta}D_{2_\theta}>0$ and $X_{2_\theta}\geq 0$ for all $\theta\in\Theta$. Therefore

$$\xi(k+1)=A_{u_{\theta_o}}(k)\xi(k),$$
$$\upsilon(k)=B'_{2_{\theta(k)}}\Gamma_{\theta_o}(k+1)A_{u_{\theta_o}}(k)\xi(k)$$

is a uniformly detectable system. Moreover,

$$\left(R_{3_{\theta_o}}(k)-B'_{2_{f^k(\theta_o)}}\Gamma_{\theta_o}(k+1)B_{2_{f^k(\theta_o)}}\right)^{-1}>0$$

and $\|\Gamma_{\theta_o}(k)\|=\|X_{\theta_o}(k)-X_{2_{f^k(\theta_o)}}\|\leq\|X_{\theta_o}(k)\|\leq\bar{X}_{\infty_{\theta_o}}<\infty$. Therefore (45) is a Lyapunov equation, and Theorem 3.5 implies that

$$x(k+1)=A_{u_{\theta_o}}(k)x(k)$$

is an exponentially stable system. □

We cannot as yet claim that A_u as defined by (44) is uniformly exponentially stable in the sense of Definition 3.1. To conclude uniformly exponential stability, $\Gamma_{\theta_o}(k)$, the solution to the Lyapunov equation (45), must be uniformly bounded for all $\theta_o\in\Theta$ and all k. $\Gamma_{\theta_o}(k)$ is uniformly bounded only if $X_{\theta_o}(k)$, the solution to (39), is uniformly bounded. Lemma 5.7 will show that $X_{\theta_o}(k)$ is uniformly bounded.

LEMMA 5.3. *Let* $u_\infty(k)=F_{u_{\theta_o}}(k)\left[\begin{smallmatrix}x(k)\\ w(k)\end{smallmatrix}\right]$, *where* F_u *is given by* (42). *Let* $w_\infty(k)=F_{w_{\theta_o}}(k)x(k)$ *be defined as in* (43). *Then, for* $w\in l_2$,

$$\|z_{\theta_o}(F_u,w,x_o)\|_{l_2}^2-\gamma^2\|w\|_{l_2}^2$$
$$=x'_oX_{\theta_o}(k)x_o+\sum_{k=0}^{\infty}\left(w(k)-w_\infty(k)\right)'\nabla_{\theta_o}(k)\left(w(k)-w_\infty(k)\right). \tag{46}$$

(See the end of section 1 for the definition of the notation $z_{\theta_o}(F_u,w,x_o)$.)

Proof. By the previous lemma, $u_\infty(k)=F_{u_{\theta_o}}(k)\left[\begin{smallmatrix}x(k)\\ w(k)\end{smallmatrix}\right]$ is exponentially stabilizing. Hence if $w\in l_2$, then $x\in l_2$ and $x(k)\to 0$. Furthermore, $X_{\theta_o}(k)$ is bounded and, by (41), $-\infty<\nabla_{\theta_o}(k)\leq-\varrho I$. Thus $F_{w_{\theta_o}}(k)$ is bounded, where F_w is defined by (43). Hence $w_\infty\in l_2$. Thus letting $N\to\infty$ in (28) yields (46). □

Now it is shown that the control u_∞ achieves Objective B.

LEMMA 5.4. *If* $x(0)=0$ *and* $u(k)=u_\infty(k)=F_{u_{\theta_o}}(k)\left[\begin{smallmatrix}x(k)\\ w(k)\end{smallmatrix}\right]$, *then, for all* $w\in l_2$, $\|z\|_{l_2}^2-\gamma^2\|w\|_{l_2}^2\leq-\varepsilon\|w\|_{l_2}^2$.

Proof. By Assumption A, there exists an exponentially stabilizing control u_* that satisfies Objective B, that is, if $u=u_*$, $x(0)=0$, and $w\in l_2$, then $x\in l_2$ and $\|z_{\theta_o}(u_*,w,0)\|_{l_2}^2-\gamma^2\|w\|_{l_2}^2\leq-\varepsilon\|w\|_{l_2}^2$. Since $X_{\theta_o}(k)$ is bounded and $\nabla_{\theta_o}(k)\leq-\varrho I$, $F_{u_{\theta_o}}(k)$ and $F_{w_{\theta_o}}(k)$ are bounded. Therefore $u_\infty(k)=F_{u_{\theta_o}}(k)\left[\begin{smallmatrix}x_{\theta_o}(u_*,w,0;k)\\ w(k)\end{smallmatrix}\right]\in l_2$, and $w_\infty(k)=F_{w_{\theta_o}}(k)x_{\theta_o}(u_*,w,0;k)\in l_2$. Thus we can take the limit of (28) as

$N \to \infty$, which yields

$$
\begin{aligned}
-\varepsilon \|w\|_{l_2}^2 &\geq \|z_{\theta_o}(u_*, w, 0)\|_{l_2}^2 - \gamma^2 \|w\|_{l_2}^2 \\
&= \sum_{k=0}^{\infty} (u_*(k) - u_\infty(k))' R_{3_{\theta_o}}(k) (u_*(k) - u_\infty(k)) \\
&\qquad + \sum_{k=0}^{\infty} (w(k) - w_\infty(k))' \nabla_{\theta_o}(k) (w(k) - w_\infty(k)) \\
&\geq \sum_{k=0}^{\infty} (w(k) - w_\infty(k))' \nabla_{\theta_o}(k) (w(k) - w_\infty(k)) \\
&= \|z_{\theta_o}(F_u, w, 0)\|_{l_2}^2 - \gamma^2 \|w\|_{l_2}^2,
\end{aligned} \tag{47}
$$

where the last equality follows from Lemma 5.3. □

From (47), it is clear that u_∞ is the best control and w_∞ is the worst disturbance in the sense of Objective B.

LEMMA 5.5. $\sup_{w \in l_2} \inf_{u \in l_2} \|z_{\theta_o}(u, w, x_o)\|_{l_2}^2 - \gamma^2 \|w\|_{l_2}^2 = \|z_{\theta_o}(F_u, w_\infty, x_o)\|_{l_2}^2 - \gamma^2 \|w_\infty\|_{l_2}^2 = x_o' X_{\theta_o}(0) x_o$.

Proof. Since $\nabla_{\theta_o}(k) \leq -\varrho I$, if $u(k) = u_\infty(k) = F_{u_{\theta_o}}(k) \left[\begin{smallmatrix} x(k) \\ w(k) \end{smallmatrix} \right]$, then (46) implies that

$$
\begin{aligned}
&\sup_{w \in l_2} \|z_{\theta_o}(F_u, w, x_o)\|_{l_2}^2 - \gamma^2 \|w\|_{l_2}^2 \\
&= x_o' X_{\theta_o}(0) x_o + \sup_{w \in l_2} \sum_{k=0}^{\infty} (w(k) - w_\infty(k))' \nabla_{\theta_o}(k) (w(k) - w_\infty(k)) \\
&= x_o' X_{\theta_o}(k) x_o,
\end{aligned}
$$

where $w_\infty(k) = F_{w_{\theta_o}}(k) x(k)$. Therefore

$$
\begin{aligned}
\sup_{w \in l_2} \inf_{u \in l_2} \|z_{\theta_o}(u, w, x_o)\|_{l_2}^2 - \gamma^2 \|w\|_{l_2}^2 &\leq \sup_{w \in l_2} \|z_{\theta_o}(F_u, w, x_o)\|_{l_2}^2 - \gamma^2 \|w\|_{l_2}^2 \\
&= x_o' X_{\theta_o}(k) x_o.
\end{aligned} \tag{48}
$$

Similarly, if $w = w_\infty$, then $\inf_{u \in l_2} \|z_{\theta_o}(u, w_\infty, x_o)\|_{l_2}^2 - \gamma^2 \|w_\infty\|_{l_2}^2 = \|z_{\theta_o}(F_u, w_\infty, x_o)\|_{l_2}^2 - \gamma^2 \|w_\infty\|_{l_2}^2 = x_o' X_{\theta_o}(k) x_o$. Therefore

$$
\begin{aligned}
\sup_{w \in l_2} \inf_{u \in l_2} \|z_{\theta_o}(u, w, x_o)\|_{l_2}^2 - \gamma^2 \|w\|_{l_2}^2 &\geq \inf_{u \in l_2} \|z_{\theta_o}(u, w_\infty, x_o)\|_{l_2}^2 - \gamma^2 \|w_\infty\|_{l_2}^2 \\
&= x_o' X_{\theta_o}(k) x_o.
\end{aligned} \tag{49}
$$

Combining inequalities (48) and (49) yields the desired result. □

Up to this point, $\theta(0) = \theta_o \in \Theta$ has been fixed. $X_{\theta_o}(k)$ is nothing more than the stabilizing solution of the time-varying Riccati equation associated with the time-varying system

$$
\begin{aligned}
x(k+1) &= A_{f^k(\theta_o)} x(k) + B_{1_{f^k(\theta_o)}} w(k) + B_{2_{f^k(\theta_o)}} u(k), \\
z(k) &= C_{f^k(\theta_o)} x(k) + D_{1_{f^k(\theta_o)}} w(k) + D_{2_{f^k(\theta_o)}} u(k).
\end{aligned}
$$

Since θ_o is arbitrary, for all $\theta \in \Theta$, define $X_\infty : \Theta \to \mathbb{R}^{n\times n}$ by

$$X_{\infty_\theta} := \lim_{N\to\infty} X_\theta(0, N+1). \tag{50}$$

LEMMA 5.6. *The function*

$$X_\infty : \Theta \to \mathbb{R}^{n\times n},$$
$$\theta \mapsto X_{\infty_\theta}$$

satisfies the FARE (32)–(33)*, viz.,*

$$\begin{aligned} X_{\infty_\theta} &= A'_\theta X_{\infty_{f(\theta)}} A_\theta + C'_\theta C_\theta \\ &\quad - \left(\bar{D}'_\theta J \bar{C}_\theta + \bar{B}'_\theta X_{\infty_{f(\theta)}} A_\theta\right)' \left(\bar{D}'_\theta J \bar{D}_\theta + \bar{B}'_\theta X_{\infty_{f(\theta)}} \bar{B}_\theta\right)^{-1} \left(\bar{D}'_\theta J \bar{C}_\theta + \bar{B}'_\theta X_{\infty_{f(\theta)}} A_\theta\right). \end{aligned} \tag{51}$$

Proof. Let $f(\theta_1) = \theta_2$. Clearly,

$$X_{\theta_1}(N+1, N+1) = X_{2_{f^{N+1}(\theta_1)}} = X_{2_{f^N(\theta_2)}} = X_{\theta_2}(N, N). \tag{52}$$

Next, by (21),

$$\begin{aligned} &X_{\theta_1}(N, N+1) \\ &= A'_{f^N(\theta_1)} X_{\theta_1}(N+1, N+1) A_{f^N(\theta_1)} + C'_{f^N(\theta_1)} C_{f^N(\theta_1)} \\ &- \left(\bar{D}'_{f^N(\theta_1)} J \bar{C}_{f^N(\theta_1)} + \bar{B}'_{f^N(\theta_1)} X_{\theta_1}(N+1, N+1) A_{f(\theta_1)}\right)' \\ &\times \left(\bar{D}'_{f^N(\theta_1)} J \bar{D}_{f^N(\theta_1)} + \bar{B}'_{f^N(\theta_1)} X_{\theta_1}(N+1, N+1) \bar{B}_{f^N(\theta_1)}\right)^{-1} \\ &\times \left(\bar{D}'_{f^N(\theta_1)} J \bar{C}_{f^N(\theta_1)} + \bar{B}'_{f^N(\theta_1)} X_{\theta_1}(N+1, N+1) A_{f^N(\theta_1)}\right) \\ &= A'_{f^{N-1}(\theta_2)} X_{\theta_2}(N, N) A_{f^{N-1}(\theta_2)} + C'_{f^{N-1}(\theta_2)} C_{f^{N-1}(\theta_2)} \\ &- \left(\bar{D}'_{f^{N-1}(\theta_2)} J \bar{C}_{f^{N-1}(\theta_2)} + \bar{B}'_{f^{N-1}(\theta_2)} X_{\theta_2}(N, N) A_{f^{N-1}(\theta_2)}\right)' \\ &\times \left(\bar{D}'_{f^{N-1}(\theta_2)} J \bar{D}_{f^{N-1}(\theta_2)} + \bar{B}'_{f^{N-1}(\theta_2)} X_{\theta_2}(N, N) \bar{B}_{f^{N-1}(\theta_2)}\right)^{-1} \\ &\times \left(\bar{D}'_{f^{N-1}(\theta_2)} J \bar{C}_{f^{N-1}(\theta_2)} + \bar{B}'_{f^{N-1}(\theta_2)} X_{\theta_2}(N, N) A_{f^{N-1}(\theta_2)}\right) \\ &= X_{\theta_2}(N-1, N). \end{aligned} \tag{53}$$

Repeating the above, we reach the result:

$$X_{\theta_1}(k, N+1) = X_{\theta_2}(k-1, N). \tag{54}$$

Setting $k = 0$ and $\theta = \theta_1$ in (21) and substituting $X_{\theta_2}(0, N)$ for $X_{\theta_1}(1, N+1)$ into the right-hand side yields

$$\begin{aligned} X_{\theta_1}(0, N+1) &= A'_{\theta_1} X_{\theta_2}(0, N) A_{\theta_1} + C'_{\theta_1} C_{\theta_1} \\ &\quad - \left(\bar{D}'_{\theta_1} J \bar{C}_{\theta_1} + \bar{B}'_{\theta_1} X_{\theta_2}(0, N) A_{\theta_1}\right)' \left(\bar{D}'_{\theta_1} J \bar{D}_{\theta_1} + \bar{B}'_{\theta_1} X_{\theta_2}(0, N) \bar{B}_{\theta_1}\right)^{-1} \\ &\quad \times \left(\bar{D}'_{\theta_1} J \bar{C}_{\theta_1} + \bar{B}'_{\theta_1} X_{\theta_2}(0, N) A_{\theta_1}\right). \end{aligned} \tag{55}$$

Next we take the limit as $N \to \infty$. In order to take this limit, we must ensure that the right-hand side is continuous in $X_{\theta_2}(0, N)$. Since $R_{3_{\theta_2}}(0, N) \geq D'_{2_{\theta_2}} D_{2_{\theta_2}} > 0$ and

$\nabla_{\theta_2}(0,N) \leq -\varrho I$ for all N, and since R_3 and ∇ are continuous functions of $X_{\theta_2}(0,N)$, we have $\lim_{N\to\infty} R_{3_{\theta_2}}(0,N) \geq D'_{2_{\theta_2}} D_{2_{\theta_2}} > 0$ and $\lim_{N\to\infty} \nabla_{\theta_2}(0,N) \leq -\varrho I$. Thus $\left(\bar{D}'_{\theta_1} J \bar{D}_{\theta_1} + B'_{\theta_1} Y B_{\theta_1}\right)^{-1}$ exists for Y in a neighborhood of $X_{\infty_{\theta_2}}$. Therefore

$$\lim_{N\to\infty} \left(\bar{D}'_{\theta_1} J \bar{D}_{\theta_1} + \bar{B}'_{\theta_1} X_{\theta_2}(0,N)\, \bar{B}_{\theta_1}\right)^{-1} = \left(\bar{D}'_{\theta_1} J \bar{D}_{\theta_1} + \bar{B}'_{\theta_1} X_{\infty_{\theta_2}} \bar{B}_{\theta_1}\right)^{-1}. \tag{56}$$

Likewise,

$$\lim_{N\to\infty} \left(\bar{D}'_{\theta_1} J \bar{C}_{\theta_1} + \bar{B}'_{\theta_1} X_{\theta_2}(0,N)\, A_{\theta_1}\right) = \left(\bar{D}'_{\theta_1} J \bar{C}_{\theta_1} + \bar{B}'_{\theta_1} X_{\infty_{\theta_2}} A_{\theta_1}\right). \tag{57}$$

Thus

$$\begin{aligned}
X_{\infty_{\theta_1}} &= \lim_{N\to\infty} X_{\theta_1}(0,N+1) \\
&= \lim_{N\to\infty} \left(A'_{\theta_1} X_{\theta_2}(0,N)\, A_{\theta_1} + C'_{\theta_1} C_{\theta_1}\right. \\
&\quad - \left(\bar{D}'_{\theta_1} J \bar{C}_{\theta_1} + \bar{B}'_{\theta_1} X_{\theta_2}(0,N)\, A_{\theta_1}\right)' \left(\bar{D}'_{\theta_1} J \bar{D}_{\theta_1} + \bar{B}'_{\theta_1} X_{\theta_2}(0,N)\, \bar{B}_{\theta_1}\right)^{-1} \\
&\quad \left. \times \left(\bar{D}'_{\theta_1} J \bar{C}_{\theta_1} + \bar{B}'_{\theta_1} X_{\theta_2}(0,N)\, A_{\theta_1}\right)\right) \\
&= A'_{\theta_1} X_{\infty_{\theta_2}} A_{\theta_1} + C'_{\theta_1} C_{\theta_1} \\
&\quad - \left(\bar{D}'_{\theta_1} J \bar{C}_{\theta_1} + \bar{B}'_{\theta_1} X_{\infty_{\theta_2}} A_{\theta_1}\right)' \left(\bar{D}'_{\theta_1} J \bar{D}_{\theta_1} + \bar{B}'_{\theta_1} X_{\infty_{\theta_2}} \bar{B}_{\theta_1}\right) \\
&\quad \times \left(\bar{D}'_{\theta_1} J \bar{C}_{\theta_1} + \bar{B}'_{\theta_1} X_{\infty_{\theta_2}} A_{\theta_1}\right).
\end{aligned} \tag{58}$$

Since $f(\theta_1) = \theta_2$, equation (51) follows. □

Now we can drop the dependence on the initial condition θ_o in R, L, and ∇, that is, $R_{\theta(k)} := R_{\theta_o}(k)$, $L_{\theta(k)} := L_{\theta_o}(k)$, $\nabla_{\theta(k)} := \nabla_{\theta_o}(k)$. As a simple consequence of (40), we find

$$X_{\infty_\theta} \geq 0, \tag{59}$$

and by (41), we get

$$\nabla_\theta \leq -\varrho I. \tag{60}$$

Thus the best control u_∞ and worst disturbance w_∞ feedback matrices depend only on the current state $\theta(k)$. That is, (42) and (43) can be rewritten as

$$u_\infty(k) = F_{u_{\theta(k)}} \begin{bmatrix} x(k) \\ w(k) \end{bmatrix} := -R^{-1}_{3_{\theta(k)}} \begin{bmatrix} L_{2_{\theta(k)}} & R_{2_{\theta(k)}} \end{bmatrix} \begin{bmatrix} x(k) \\ w(k) \end{bmatrix} \tag{61}$$

and

$$w_\infty(k) = F_{w_{\theta(k)}} x(k) := -\nabla^{-1}_{\theta(k)} L_{\nabla_{\theta(k)}} x(k). \tag{62}$$

LEMMA 5.7. *The function X_∞ given by* (50) *is uniformly bounded. That is, there exists a $\bar{X}_\infty < \infty$ such that $\|X_{\infty_\theta}\| < \bar{X}_\infty$.*

Proof. Let $X_{2_\theta} = X'_{2_\theta} \geq 0$ be the solution to (17), and define

$$A_{LQ_\theta} := A_\theta - B_{2_\theta} \left(D'_{2_\theta} D_{2_\theta} + B'_{2_\theta} X_{2_{f(\theta)}} B_{2_\theta}\right)^{-1} B'_{2_\theta} X_{2_{f(\theta)}} A_\theta$$

to be the closed-loop state transition matrix with $w = 0$ and $u = u_{LQ}$ given by (18). Define

$$v(k) := \sum_{i=k}^{\infty} \left(\prod_{j=k}^{i} A_{LQ_{f^j(\theta_o)}}\right)' \left(X_{2_{f^{i+1}(\theta_o)}} B_{1_{f^i(\theta_o)}} w(i) + C'_{f^{i+1}(\theta_o)} D_{1_{f^{i+1}(\theta_o)}} w(i+1)\right)$$

and

$$G_{\theta_o}(w,x_o;k) := \left(D'_{2_{f^k(\theta_o)}} D_{2_{f^k(\theta_o)}} + B'_{2_{f^k(\theta_o)}} X_{2_{f^{k+1}(\theta_o)}} B_{2_{f^k(\theta_o)}}\right)^{-1} B'_{2_{f^k(\theta_o)}} \times \left(X_{2_{f^{k+1}(\theta_o)}} A_{f^k(\theta_o)} x(k) - v(k)\right). \tag{63}$$

It is possible to show (see [26, Claim 3, p. 257] or [30, Lemma 9.6]) that if $w \in l_2$, then

$$G_{\theta_o}(w,x_o) = \arg\inf\left\{\|z_{\theta_o}(u,w,x_o)\|_{l_2} : u \in l_2\right\}, \tag{64}$$

where the notation $z_{\theta_o}(u,w,x_o)$ was introduced at the end of section 1. Note that $G_{\theta_o}(w,x_o)$ and therefore $z_{\theta_o}(G_{\theta_o}(w,x_o),w,x_o)$ are linear in (w,x_o).

By assumption, there exists a control satisfying Objective B. Thus $G_{\theta_o}(w,0)$ must also satisfy Objective B, that is,

$$\|z_{\theta_o}(G_{\theta_o}(w,0),w,0)\|_{l_2}^2 - \gamma^2\|w\|_{l_2}^2 \le -\varepsilon\|w\|_{l_2}^2 \tag{65}$$

and

$$\|z_{\theta_o}(G_{\theta_o}(w,0),w,0)\|_{l_2} < \gamma\|w\|_{l_2}. \tag{66}$$

If $w \equiv 0$, then $v = 0$, and by comparing (18) and (63) we see that $G_{\theta_o}(0,x_o;k) = u_{LQ}(k)$, that is, $G_{\theta_o}(0,x_o;k)$ is the optimal LQ control given by (18). Thus, if $w = 0$, then by (19) we obtain

$$\|z_{\theta_o}(G_{\theta_o}(0,x_o),0,x_o)\|_{l_2}^2 = \inf_{u\in l_2}\|z_{\theta_o}(u,0,x_o)\|_{l_2}^2 = x'_o X_{2_{\theta_o}} x_o, \tag{67}$$

where X_{2_θ} is the solution to the Riccati equation (17). It was shown in [8] that X_2 is uniformly bounded. Denote this bound by $\bar{X}_2$, that is, for all $\theta \in \Theta$, $\|X_{2_\theta}\| \le \bar{X}_2 < \infty$. Hence

$$\|z_{\theta_o}(G_{\theta_o}(0,x_o),0,x_o)\|_{l_2}^2 = x'_o X_{2_{\theta_o}} x_o \le \bar{X}_2|x_o| < \infty. \tag{68}$$

Combining (65), (66), and (68) yields

$$\begin{aligned}
&\|z_{\theta_o}(G_{\theta_o}(w,x_o),w,x_o)\|_{l_2}^2 - \gamma^2\|w\|_{l_2}^2 \\
&= \|z_{\theta_o}(G_{\theta_o}(0,x_o),0,x_o) + z_{\theta_o}(G_{\theta_o}(w,0),w,0)\|_{l_2}^2 - \gamma^2\|w\|_{l_2}^2 \\
&= \|z_{\theta_o}(G_{\theta_o}(0,x_o),0,x_o)\|_{l_2}^2 + 2\langle z_{\theta_o}(G_{\theta_o}(w,0),w,0), z_{\theta_o}(G_{\theta_o}(0,x_o),0,x_o)\rangle \\
&+ \|z_{\theta_o}(G_{\theta_o}(w,0),w,0)\|_{l_2}^2 - \gamma^2\|w\|_{l_2}^2 \\
&\le \|z_{\theta_o}(G_{\theta_o}(0,x_o),0,x_o)\|_{l_2}^2 \\
&+ 2\langle z_{\theta_o}(G_{\theta_o}(w,0),w,0), z_{\theta_o}(G_{\theta_o}(0,x_o),0,x_o)\rangle - \varepsilon\|w\|_{l_2}^2 \\
&\le x'_o X_{2_{\theta_o}} x_o + 2\gamma\|w\|_{l_2}\sqrt{x'_o x_o \bar{X}_2} - \varepsilon\|w\|_{l_2}^2 \\
&= x'_o X_{2_{\theta_o}} x_o + \|w\|_{l_2}\left(2\gamma|\sqrt{x'_o x_o \bar{X}_2} - \varepsilon\|w\|_{l_2}\right) \\
&\le x'_o X_{2_{\theta_o}} x_o + \max_{\|w\|\in\mathbb{R}}\left\{\|w\|_{l_2}\left(2\gamma|x_o|\sqrt{\bar{X}_2} - \varepsilon\|w\|_{l_2}\right)\right\} \\
&\le x'_o X_{2_{\theta_o}} x_o + \frac{\gamma^2\bar{X}_2|x_o|^2}{\varepsilon} \le |x_o|^2\left(\bar{X}_2 + \frac{\gamma^2\bar{X}_2}{\varepsilon}\right) < \infty.
\end{aligned} \tag{69}$$

Thus

$$\sup_{w\in l_2}\inf_{u\in l_2}\|z_{\theta_o}(u,w,x_o)\|_{l_2}^2-\gamma^2\|w\|_{l_2}^2$$
$$=\sup_{w\in l_2}\|z_{\theta_o}(G_{\theta_o}(w,x_o),w,x_o)\|_{l_2}^2-\gamma^2\|w\|_{l_2}^2\le|x_o|^2\left(\bar{X}_2+\frac{\gamma^2\bar{X}_2}{\varepsilon}\right)<\infty.$$

Lemma 5.5 implies that

(70)
$$x_o'X_{\infty_{\theta_o}}x_o=\sup_{w\in l_2}\inf_{u\in l_2}\|z_{\theta_o}(u,w,x_o)\|_{l_2}^2-\gamma^2\|w\|_{l_2}^2\le|x_o|^2\left(\bar{X}_2+\frac{\gamma^2\bar{X}_2}{\varepsilon}\right)<\infty,$$

which concludes the proof. □

Note that the worst disturbance has the property,

$$\|w_\infty\|_{l_2}^2\le P|x_o|^2, \tag{71}$$

where

$$P:=\frac{4\gamma^2\bar{X}_2}{\varepsilon^2}|x_o|^2. \tag{72}$$

To see this, observe that if $\|w_\infty\|_{l_2}^2>P|x_o|^2$, equation (69) implies that

$$\begin{aligned}&\|z_{\theta_o}(G_{\theta_o}(w_\infty,x_o),w_\infty,x_o)\|_{l_2}^2-\gamma^2\|w_\infty\|_{l_2}\\&\le x_o'X_{2_{\theta_o}}x_o+\|w\|_{l_2}\left(2\gamma|x_o|\sqrt{\bar{X}_2}-\varepsilon\|w\|_{l_2}\right)\\&<x_o'X_{2_{\theta_o}}x_o=\|z_{\theta_o}(G_{\theta_o}(0,x_o),0,x_o)\|_{l_2}^2.\end{aligned}$$

That is, the cost resulting from $w\equiv0$ is larger than the cost resulting from $w=w_\infty$, which contradicts the maximizing property of w_∞.

LEMMA 5.8. *For $w=0$, $u(k)=u_\infty(k)$ uniformly exponentially stabilizes the system.*

Proof. Since X_∞ is uniformly bounded, $\bar{X}_\infty(\theta)$ defined in Lemma 5.1 is uniformly bounded. The proof of Lemma 5.2 can be applied with no changes to conclude that A_u is uniformly exponentially stable. □

LEMMA 5.9. *The closed-loop system with $u(k)=u_\infty(k)$ and $w(k)=w_\infty(k)$ is uniformly exponentially stable. In other words, the system $x(k+1)=(A_{\theta(k)}-\bar{B}_{\theta(k)}R_{\theta(k)}^{-1}L_{\theta(k)})x(k)$ is uniformly exponentially stable.*

Proof. Let $|x_o|\le1$. Define w_∞ as

$$w_\infty(k)=F_{w_{\theta(k)}}x_{\theta_o}(F_u,w_\infty,x_o;k).$$

Then w_∞ is a linear function of x_o. By Lemmas 5.5 and 5.7,

$$\begin{aligned}\sup_{w\in l_2}\inf_{u\in l_2}\|z_{\theta_o}(u,w,x_o)\|_{l_2}^2-\gamma^2\|w\|_{l_2}^2&=\|z_{\theta_o}(F_u,w_\infty,x_o)\|_{l_2}^2-\gamma^2\|w_\infty\|_{l_2}^2\\&=x_o'X_{\infty_{\theta_o}}x_o\le\bar{X}_\infty|x_o|^2,\end{aligned}$$

and, by (71), $\|w_\infty\|_{l_2}^2\le P|x_o|^2$. Thus

$$\|z_{\theta_o}(F_u,w_\infty,x_o)\|_{l_2}^2=\gamma^2\|w_\infty\|_{l_2}^2+x_o'X_{\infty_{\theta_o}}x_o\le\left(\gamma^2P+\bar{X}_\infty\right)|x_o|^2. \tag{73}$$

Note that, if $w(k) = w_\infty(k)$ and $u(k) = F_{u_{\theta(k)}}\left[\begin{smallmatrix} x(k) \\ w_\infty(k) \end{smallmatrix}\right]$, then $x(k+1) = (A_{\theta(k)} - \bar{B}_{\theta(k)} R^{-1}_{\theta(k)} L_{\theta(k)}) x(k)$.

Define the system

$$x(k+1) = \left(A_{\theta(k)} - \bar{B}_{\theta(k)} R^{-1}_{\theta(k)} L_{\theta(k)}\right) x(k) + r(k), \tag{74}$$
$$v(k) = \begin{bmatrix} -\nabla^{-1}_{\theta(k)} L_{\nabla_{\theta(k)}} \\ \tilde{C}_{\theta(k)} \end{bmatrix} x(k),$$

where

$$\tilde{C} = C - D_1 \nabla^{-1} L_\nabla + D_2 \left(R_3^{-1} R_2 \nabla^{-1} L_\nabla - R_3^{-1} L_2\right).$$

Then $v = \left[\begin{smallmatrix} w_\infty \\ z_{\theta_o}(F_u, w_\infty, x_o) \end{smallmatrix}\right]$. Fix $j \geq 0$ and set $r(k) = r_o \delta(k-j)$, that is, $r(k) = 0$ for $k \neq j$ and $r(j) = r_o$. Then (73) implies that $\left\| z_{f^j(\theta_o)}(F_u, w_\infty, r_o) \right\|^2_{l_2} \leq \left(\gamma^2 P + \bar{X}_\infty\right) |r_o|^2$. Likewise, $\|w_\infty\|^2_{l_2} \leq P |r_o|^2$. Therefore

$$\|v\|^2_{l_2} = \left\| z_{f^j(\theta_o)}(F_u, w_\infty, r_o) \right\|^2_{l_2} + \|w\|^2_{l_2} \leq \left(\left(\gamma^2 P + \bar{X}_\infty\right) + P\right) |r_o|^2. \tag{75}$$

Note that there exists a matrix $\left[\begin{smallmatrix} H_1 & H_2 \end{smallmatrix}\right]$ such that

$$\left(A - \bar{B} R^{-1} L\right) + \begin{bmatrix} H_1 & H_2 \end{bmatrix} \begin{bmatrix} -\nabla^{-1} L_\nabla \\ \left(C - D_1 \nabla^{-1} L_\nabla + D_2 \left(R_3^{-1} R_2 \nabla^{-1} L_\nabla - R_3^{-1} L_2\right)\right) \end{bmatrix} \tag{76}$$

is uniformly exponentially stable. For example, set $H_1 = -B_1 - H_2 D_1$ and $H_2 = H_d C (C'C)^+ C' - B_2 (D_2' D_2)^{-1} D_2'$, where H_d is the feedback such that $A - H_d C$ is uniformly exponentially stable, the existence of which is guaranteed by the detectability assumption, and $(C'C)^+$ is the pseudoinverse of $C'C$. Since H_d is bounded, $D_2' D_2 > 0$, D_2, B_1, C, D_1 are uniformly continuous, Θ is compact, and $C (C'C)^+ C'$ is bounded,[1] there is an $\bar{H} < \infty$ such that $\left[\begin{smallmatrix} H_1 & H_2 \end{smallmatrix}\right]' \left[\begin{smallmatrix} H_1 & H_2 \end{smallmatrix}\right] \leq \bar{H}$.

Now, let

$$y(k+1) = \left(\left(A_{\theta(k)} - \bar{B}_{\theta(k)} R^{-1}_{\theta(k)} L_{\theta(k)}\right) + \begin{bmatrix} H_{1_{\theta(k)}} & H_{2_{\theta(k)}} \end{bmatrix} \begin{bmatrix} \tilde{C}_{\theta(k)} \\ \nabla^{-1}_{\theta(k)} L_{\nabla_{\theta(k)}} \end{bmatrix}\right) y(k) \tag{77}$$
$$- \begin{bmatrix} H_{1_{\theta(k)}} & H_{2_{\theta(k)}} \end{bmatrix} v(k) + r_o \delta(k-j),$$

that is, y is an estimate of x. Since system (77) is uniformly exponentially stable, there exists an $R < \infty$ such that

$$\begin{aligned} \|y\|_{l_2} &\leq R \left\| r - \begin{bmatrix} H_1 & H_2 \end{bmatrix} v \right\|_{l_2} \leq R \|r\|_{l_2} + \bar{H} R \|v\|_{l_2} \\ &\leq R |r_o| + \bar{H} R \sqrt{\left(\left(\gamma^2 P + \bar{X}_\infty\right) + P\right)} |r_o| \\ &\leq \left(R + R \bar{H} \sqrt{\left(\left(\gamma^2 P + \bar{X}_\infty\right) + P\right)}\right) |r_o|. \end{aligned}$$

[1] Although $C (C'C)^+ C$ is not continuous, $\|C (C'C)^+ C\| \leq 1$.

On the other hand, if the system is initially at rest, that is, $y(0) = x(0) = 0$, then $y(k) = x(k)$. Thus $\|y\|_{l_2} = \|x\|_{l_2}$, and therefore

$$\|x\|_{l_2} \leq \left(R + R\bar{H}\sqrt{\left(\left(\gamma^2 P + \bar{X}_\infty\right) + P\right)} \right) |r_o| .$$

Let $\Phi_{\theta_o}(k,j)$ be the state transition matrix of system (74) with initial conditions $\theta(0) = \theta_o$, $x(0) = 0$, and let $r(i) = r_o \delta(i-j)$. Then we have

$$\begin{aligned} \|x\|^2_{[j,\infty)} &= \sum_{i=j}^{\infty} x'(i)\, x(i) = \sum_{i=j}^{\infty} \|\Phi_{\theta_o}(i,j)\, r(j)\|^2 \\ &\leq \left(R + R\bar{H}\sqrt{\left(\left(\gamma^2 P + \bar{X}_\infty\right) + P\right)} \right)^2 |r_o|^2 . \end{aligned}$$

Furthermore, for $i \geq j$, $\|\Phi_{\theta_o}(i,j)\|^2 \leq (R + R\bar{H}\sqrt{\left(\left(\gamma^2 P + \bar{X}_\infty\right) + P\right)})^2$. Applying standard techniques, we find that

$$\begin{aligned} K\,\|\Phi_{\theta_o}(j+K,j)\|^2 &= \sum_{i=j}^{K} \|\Phi_{\theta_o}(j+K,j)\|^2 \\ &\leq \sum_{i=j}^{K} \|\Phi_{\theta_o}(j+K,i)\|^2 \|\Phi_{\theta_o}(i,j)\|^2 \\ &\leq \left(R + R\bar{H}\sqrt{\left(\left(\gamma^2 P + \bar{X}_\infty\right) + P\right)} \right)^2 \sum_{i=j}^{K} \|\Phi_{\theta_o}(i,j)\|^2 \\ &\leq \left(R + R\bar{H}\sqrt{\left(\left(\gamma^2 P + \bar{X}_\infty\right) + P\right)} \right)^4 . \end{aligned}$$

Choosing $K \in \mathbb{Z}$ such that $K \geq \sqrt{2}(R + R\bar{H}\sqrt{((\gamma^2 P + \bar{X}_\infty) + P)})^4$ yields $\|\Phi(j+K, j, \theta_o)\|^2 \leq 1/\sqrt{2}$. Since this is true for all j and all θ_o, setting $M \in \mathbb{Z}$ with $M \geq 0$ and $k - (j + MK) < K$, we conclude that

$$\begin{aligned} \|\Phi_{\theta_o}(k,j)\| &\leq \|\Phi_{\theta_o}(k, j+MK)\| \prod_{m=1}^{M} \|\Phi_{\theta_o}(j+mK, j+(m-1)K)\| \\ &\leq \left(R + R\bar{H}\sqrt{\left(\left(\gamma^2 P + \bar{X}_\infty\right) + P\right)} \right) \left(\frac{1}{2}\right)^M \\ &\leq \left(R + R\bar{H}\sqrt{\left(\left(\gamma^2 P + \bar{X}_\infty\right) + P\right)} \right) 2 \left(\frac{1}{2}\right)^{\frac{k-j}{K}} . \end{aligned}$$

That is, system (74) is uniformly exponentially stable. □

LEMMA 5.10. *As $N \to \infty$, $X_\theta(0, N+1) \to X_{\infty_\theta}$ uniformly in θ and X_∞ is a continuous function.*

Proof. Let $x(k+1) = (A_{\theta(k)} - \bar{B}_{\theta(k)} R^{-1}_{\theta(k)} L_{\theta(k)}) x(k)$. Then, by Lemma 5.9, $x(k) \to 0$ uniformly exponentially fast. Set $w_\infty(k) = F_{w_{\theta(k)}} x(k)$ as in (62) and define

$$\tilde{w}_N(k) := \begin{cases} w_\infty(k) & \text{for } k \leq N, \\ 0 & \text{otherwise.} \end{cases}$$

Since X_∞ is uniformly bounded and $\nabla_\theta \leq -\varrho I$, it follows that F_w is uniformly bounded. Since $x(k) \to 0$ uniformly exponentially fast and F_w is bounded, $w_\infty \to 0$ uniformly exponentially fast. Therefore $\lim_{N\to\infty} \|w_\infty - \tilde{w}_N\|_{l_2} = \lim_{N\to\infty} \|w_\infty\|_{[N+1,\infty)} = 0$, where the convergence is uniformly exponentially fast.

Recall the following: (31) states that

$$
\begin{aligned}
(78)\qquad & x_o' X_{\theta_o}(0, N+1)\, x_o \\
&= \sup_{w\in l_2[0,N]} \inf_{u\in l_2[0,N]} \Big\{ \|z\|^2_{[0,N]} - \gamma^2 \|w\|^2_{[0,N]} \\
&\qquad\qquad + x'(N+1)\, X_{\theta_o}(N+1, N+1)\, x(N+1) \Big\}.
\end{aligned}
$$

From (19) of Theorem 3.7, it follows that

$$
(79)\qquad x(N+1)' X_{2_{f^{N+1}(\theta_o)}} x(N+1) = \inf_{u\in l_2} \|z\|^2_{[N+1,\infty)}.
$$

From (64), we have

$$
(80)\qquad \inf_{u\in l_2} \|z_{\theta_o}(u, w, x_o)\|^2_{[0,\infty)} - \gamma^2 \|w\|^2_{[0,\infty)} = \|z_{\theta_o}(G_{\theta_o}(w, x_o), w, x_o)\|^2_{[0,\infty)} - \gamma^2 \|w\|^2_{[0,\infty)}.
$$

Combining (64) and Lemma 5.5 yields

$$
(81)\qquad \|z_{\theta_o}(G_{\theta_o}(w_\infty, x_o), w_\infty, x_o)\|^2_{[0,\infty)} - \gamma^2 \|w_\infty\|_{l_2} = x_o' X_{\infty_{\theta_o}} x_o.
$$

From (66), we have

$$
(82)\qquad \|z_{\theta_o}(G_{\theta_o}(w, 0), w, 0)\|^2_{[0,\infty)} < \gamma^2 \|w\|^2_{[0,\infty)},
$$

and, from inequality (71), we have

$$
(83)\qquad \|w_\infty\|^2_{l_2} \leq P |x_o|^2.
$$

Combining the preceding relations yields the following string:

$$
\begin{aligned}
& x_o' X_{\theta_o}(0, N+1)\, x_o \\
&= \sup_{w\in l_2[0,N]} \inf_{u\in l_2[0,N]} \Big\{ \|z_{\theta_o}(u, w, x_o)\|^2_{[0,N]} - \gamma^2 \|w\|^2_{[0,N]} \\
&\qquad\qquad + x'(N+1)\, X_{2_{f^{N+1}(\theta_o)}} x(N+1) \Big\} \\
&= \sup_{\{w\in l_2 : w(k)=0,\ k\geq N\}} \inf_{u\in l_2} \Big\{ \|z_{\theta_o}(u, w, x_o)\|^2_{[0,\infty)} - \gamma^2 \|w\|^2_{[0,\infty)} \Big\} \\
&= \sup_{\{w\in l_2 : w(k)=0,\ k\geq N\}} \Big\{ \|z_{\theta_o}(G_{\theta_o}(w, x_o), w, x_o)\|^2_{[0,\infty)} - \gamma^2 \|w\|^2_{[0,\infty)} \Big\} \\
&\geq \|z_{\theta_o}(G_{\theta_o}(\tilde{w}_N, x_o), \tilde{w}_N, x_o)\|^2_{[0,\infty)} - \gamma^2 \|\tilde{w}_N\|^2_{[0,\infty)} \\
&= \|z_{\theta_o}(G_{\theta_o}(w_\infty, x_o), w_\infty, x_o) + z_{\theta_o}(G_{\theta_o}(\tilde{w}_N - w_\infty, 0), \tilde{w}_N - w_\infty, 0)\|^2_{[0,\infty)} \\
&\quad - \gamma^2 \|\tilde{w}_N\|^2_{[0,\infty)}
\end{aligned}
$$

$$
\begin{aligned}
&= \|z_{\theta_o}\left(G_{\theta_o}\left(w_\infty, x_o\right), w_\infty, x_o\right)\|^2_{[0,\infty)} + \|z_{\theta_o}\left(G_{\theta_o}\left(\tilde{w}_N - w_\infty, 0\right), \tilde{w}_N - w_\infty, 0\right)\|^2_{[0,\infty)} \\
&\quad - \gamma \|\tilde{w}_N\|^2_{[0,\infty)} \\
&\quad + 2\left\langle z_{\theta_o}\left(G_{\theta_o}\left(w_\infty, x_o\right), w_\infty, x_o\right), z_{\theta_o}\left(G_{\theta_o}\left(\tilde{w}_N - w_\infty, 0\right), \tilde{w}_N - w_\infty, 0\right)\right\rangle \\
&\geq \|z_{\theta_o}\left(G_{\theta_o}\left(w_\infty, x_o\right), w_\infty, x_o\right)\|^2_{[0,\infty)} - \gamma^2 \|\tilde{w}_N\|^2_{[0,\infty)} \\
&\quad + 2\left\langle z_{\theta_o}\left(G_{\theta_o}\left(w_\infty, x_o\right), w_\infty, x_o\right), z_{\theta_o}\left(G_{\theta_o}\left(\tilde{w}_N - w_\infty, 0\right), \tilde{w}_N - w_\infty, 0\right)\right\rangle \\
&\geq \|z_{\theta_o}\left(G_{\theta_o}\left(w_\infty, x_o\right), w_\infty, x_o\right)\|^2_{[0,\infty)} - \gamma^2 \|w_\infty\|^2_{[0,\infty)} \\
&\quad - 2\|z_{\theta_o}\left(G_{\theta_o}\left(w_\infty, x_o\right), w_\infty, x_o\right)\|_{[0,\infty)} \|z_{\theta_o}\left(G_{\theta_o}\left(\tilde{w}_N - w_\infty, 0\right), \tilde{w}_N - w_\infty, 0\right)\|_{[0,\infty)} \\
&\geq x_o' X_{\infty_{\theta_o}} x_o - 2\left(\sqrt{\gamma^2 \|w_\infty\|^2 + x_o' X_{\infty_{\theta_o}} x_o}\right)\gamma \|w_\infty - \tilde{w}_N\|_{[0,\infty)} \\
&\geq x_o' X_{\infty_{\theta_o}} x_o - 2\left|x_o\right|\left(\sqrt{\gamma^2 P + \bar{X}_\infty}\right)\gamma \|w_\infty - \tilde{w}_N\|_{[0,\infty)}.
\end{aligned}
$$

Lemma 5.1 implies that $X_{\infty_{\theta_o}} - X_{\theta_o}(0, N+1) \geq 0$. Thus

$$
0 \leq x_o'\left(X_{\infty_{\theta_o}} - X_{\theta_o}(0, N+1)\right) x_o \leq 2\gamma\left|x_o\right|\left(\sqrt{\gamma^2 P + \bar{X}_\infty}\right)\|w_\infty - \tilde{w}_N\|_{[0,\infty)}.
$$

Since $\|w_\infty - \tilde{w}_N\|_{[0,\infty)} \to 0$ uniformly in θ and exponentially in N, and since $2\gamma(\sqrt{\gamma^2 P + \bar{X}_\infty})$ does not depend on θ_o, we have $X_{\theta_o}(0, N+1) \to X_{\infty_{\theta_o}}$ uniformly in θ and exponentially in N.

Since X_2 is continuous, $X_\theta(0, N+1)$ is continuous in θ for $N < \infty$. Since Θ is compact, and $X(0, N+1) \to X_\infty$ in the uniform metric, Theorem 7.1.4 in [24] implies that X_∞ is continuous. □

The time-invariant version of the first claim of this lemma can be found in [31].

5.2. Sufficiency. Suppose that the assumptions of the theorem hold and that (32), (33), and (34) hold. It will be shown that the control given by (36) is internally stabilizing and, if $u = u_\infty$ as defined by (42), then Objective B is satisfied. This proof is similar to the proof given in [17].

Under the above conditions, (32) can be written as

$$
X_{\infty_\theta} = C_\theta' C_\theta + A_\theta' X_{\infty_{f(\theta)}} A_\theta - L_{2_\theta}' R_{3_\theta}^{-1} L_{2_\theta} - L_{\nabla_\theta}' \nabla_\theta^{-1} L_{\nabla_\theta}.
$$

It follows that

$$
\begin{aligned}
&\begin{bmatrix} A_\theta' & C_\theta' \\ B_{2_\theta}' & D_{2_\theta}' \end{bmatrix}\begin{bmatrix} X_{\infty_{f(\theta)}} & 0 \\ 0 & I \end{bmatrix}\begin{bmatrix} A_\theta & B_{2_\theta} \\ C_\theta & D_{2_\theta} \end{bmatrix} \\
&= \begin{bmatrix} X_{\infty_\theta} + L_{\nabla_\theta}' \nabla_\theta^{-1} L_{\nabla_\theta} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} L_{2_\theta}' \\ R_{3_\theta} \end{bmatrix} R_{3_\theta}^{-1} \begin{bmatrix} L_{2_\theta} & R_{3_\theta} \end{bmatrix}.
\end{aligned}
$$

Multiplying both sides of this equality, by $\begin{bmatrix} I & 0 \\ -R_{3_\theta}^{-1} L_{2_\theta} & I \end{bmatrix}$ on the right and by the transpose on the left, and taking the (1,1) block yields

$$
\begin{aligned}
X_{\infty_\theta} &= \left(A_\theta - B_{2_\theta} R_{3_\theta}^{-1} L_{2_\theta}\right)' X_{\infty_{f(\theta)}}\left(A_\theta - B_{2_\theta} R_{3_\theta}^{-1} L_{2_\theta}\right) \\
&\quad - L_{\nabla_\theta}' \nabla_\theta^{-1} \nabla_\theta \nabla_\theta^{-1} L_{\nabla_\theta} + \left(C_\theta - D_{2_\theta} R_{3_\theta}^{-1} L_{2_\theta}\right)'\left(C_\theta - D_{2_\theta} R_{3_\theta}^{-1} L_{2_\theta}\right).
\end{aligned} \tag{84}
$$

Since $A_\theta - \bar{B}_\theta R_\theta^{-1} L_\theta = A_\theta - B_{2_\theta} R_{3_\theta}^{-1} L_{2_\theta} - \left(B_{1_\theta} - B_{2_\theta} R_{3_\theta}^{-1} L_{2_\theta}\right) \nabla_\theta^{-1} L_{\nabla_\theta}$ is assumed to be uniformly exponentially stable, we conclude that the triple

$$\left(\left(A_\theta - B_{2_\theta} R_{3_\theta}^{-1} L_{2_\theta}\right), \left(\nabla_\theta^{-1} L_{\nabla_\theta}\right), f\right)$$

is uniformly detectable. Since $\nabla_\theta \leq -\varrho I$, ∇_θ^{-1} is uniformly bounded. Since X_∞ is uniformly bounded, $\nabla_\theta^{-1} L_{\nabla_\theta}$ is uniformly bounded. Thus (84) is a Lyapunov equation, and Corollary 3.6 implies that

$$\xi(k+1) = \left(A_{\theta(k)} - B_{2_{\theta(k)}} R_{3_{\theta(k)}}^{-1} L_{2_{\theta(k)}}\right) \xi(k) \tag{85}$$

is a uniformly exponentially stable system. Therefore the control $u = u_\infty$ is uniformly exponentially stabilizing.

Since system (85) is uniformly exponentially stable, if $u = u_\infty$ and $w \in l_2$, then $x \in l_2$ and $\lim_{k\to\infty} x(k) = 0$. Thus, if $u = u_\infty$, then (28) implies that, for all N,

$$\begin{aligned} &\|z\|_{[0,N]}^2 - \gamma^2 \|w\|_{[0,N]}^2 + x'(N+1) X_{\infty_{f^{N+1}(\theta_o)}} x(N+1) \\ &\quad = x'(0) X_{\infty_{\theta_o}} x(0) + \sum_{k=0}^{N} \left(w(k) - w_\infty(k)\right)' \nabla_{f^k(\theta_o)} \left(w(k) - w_\infty(k)\right), \end{aligned} \tag{86}$$

where $w_\infty(k) := -\nabla_{f^k(\theta_o)}^{-1} L_{\nabla_{f^k(\theta_o)}} x(k)$. Since $x, w \in l_2$, it follows that $u, z \in l_2$. Furthermore, ∇ is bounded. Thus we can let $N \to \infty$ in (86), and for $x_o = 0$,

$$\|z\|_{l_2}^2 - \gamma \|w\|_{l_2}^2 = \sum_{k=0}^{\infty} \left(w(k) - w_\infty(k)\right)' \nabla_{f^k(\theta_o)} \left(w(k) - w_\infty(k)\right). \tag{87}$$

Since system (85) is stable and causal, the closed-loop system with $u = u_\infty$, viz.,

$$\begin{aligned} &\begin{bmatrix} x(k+1) \\ w(k) - w_\infty(k) \end{bmatrix} \\ &= \begin{bmatrix} \left(A_{\theta(k)} - B_{2_{\theta(k)}} R_{3_{\theta(k)}}^{-1} L_{2_{\theta(k)}}\right) & \left(B_{1_{\theta(k)}} - B_{2_{\theta(k)}} R_{3_{\theta(k)}}^{-1} R_{2_{\theta(k)}}\right) \\ \left(\nabla_{\theta(k)}^{-1} L_{\nabla_{\theta(k)}}\right) & I \end{bmatrix} \begin{bmatrix} x(k) \\ w(k) \end{bmatrix}, \end{aligned} \tag{88}$$

is l_2-stable and causal. The inverse of this system (see [36]) is

$$\begin{bmatrix} \xi(k+1) \\ w(k) \end{bmatrix} = \begin{bmatrix} \tilde{A}_{\theta(k)} & -\tilde{B}_{\theta(k)} \\ \tilde{C}_{\theta(k)} & \tilde{D}_{\theta(k)} \end{bmatrix} \begin{bmatrix} \xi(k) \\ w(k) - w_\infty(k) \end{bmatrix}$$

with

$$\begin{aligned} \tilde{A}_\theta &= A_\theta - B_{2_\theta} R_{3_\theta}^{-1} L_{2_\theta} - \left(B_{1_\theta} - B_{2_\theta} R_{3_\theta}^{-1} R_{2_\theta}\right) \nabla_\theta^{-1} L_{\nabla_\theta} \\ &= A_\theta - \bar{B}_\theta R_\theta^{-1} L_\theta, \\ \tilde{B}_\theta &= -\left(B_{1_\theta} - B_{2_\theta} R_{3_\theta}^{-1} R_{2_\theta}\right), \\ \tilde{C}_\theta &= -\nabla_\theta^{-1} L_{\nabla_\theta}, \\ \tilde{D}_\theta &= I. \end{aligned}$$

Since $\xi(k+1) = (A_{\theta(k)} - \bar{B}_{\theta(k)} R_{\theta(k)}^{-1} L_{\theta(k)})\xi(k)$ is uniformly exponentially stable, the inverse of system (88) is uniformly exponentially stable and hence l_2 stable. Thus there exists a $\delta > 0$ such that, for all $\theta_o \in \Theta$,

$$\|w\|_{l_2}^2 \leq \frac{1}{\delta} \|w - w_\infty\|_{l_2}^2 .$$

Since $\nabla \leq -\varrho I$, equation (87) implies that

$$\|z\|_{l_2}^2 - \gamma^2 \|w\|_{l_2}^2 \leq -\varrho \|w - w_\infty\|_{l_2}^2 \leq -\delta\varrho \|w\|_{l_2}^2 = -\varepsilon \|w\|_{l_2}^2 \ . \qquad \square$$

6. Controlling nonlinear systems with linear dynamically varying H^∞ controllers. In the preceding section, a technique for stabilizing an LDV system subject to an H^∞ disturbance rejection requirement was developed. Here, it will first be shown that the LDV controller for the *linearized* tracking error system can also be used to stabilize the *nonlinear* tracking error dynamics, in a scheme that works along every trajectory, provided that the initial tracking error is small enough (section 6.1). Next, some issues quite specific to the H^∞ implementation of the tracking scheme (to be published elsewhere) will be briefly surveyed.

6.1. Stability of a closed-loop nonlinear system. To make H^∞ design relevant to nonlinear tracking performance improvement, the guiding idea is to write the nonlinearity η in the tracking error dynamics (6) as a feedback from an output $z(k)$ to a disturbance $w(k)$. To this end, we introduce the factorization

(89)
$$\eta(x,u,\theta) = \begin{bmatrix} \eta_x(x,u,\theta) & \eta_u(x,u,\theta) \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} = B_{1_\theta} \tilde{\eta}(x,u,\theta) \begin{bmatrix} C_\theta & D_{2_\theta} \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix}.$$

Here we simply take

(90)
$$B_{1_\theta} := I_{n\times n}, \quad C_\theta := \begin{bmatrix} I_{n\times n} \\ 0_{m\times n} \end{bmatrix}, \quad D_{2_\theta} := \begin{bmatrix} 0_{n\times m} \\ I_{m\times m} \end{bmatrix},$$
$$\tilde{\eta}(x,u,\theta) := \begin{bmatrix} \eta_x(x,u,\theta) & \eta_u(x,u,\theta) \end{bmatrix},$$

where η_x and η_u are defined by (8) and (9). The error due to linearization can be modeled as a feedback from z to w:

$$\begin{aligned} \eta(x(k),u(k),\theta(k)) &= B_{1_{\theta(k)}} w(k), \\ w(k) &= \tilde{\eta}(x(k),u(k),\theta(k))\, z(k), \\ z(k) &= C_{\theta(k)} x(k) + D_{2_{\theta(k)}} u(k). \end{aligned}$$

Here $D_1 = 0$. Clearly $B_1 : \Theta \to \mathbb{R}^{n\times n}$, $C : \Theta \to \mathbb{R}^{(n+m)\times n}$, and $D_2 : \Theta \to \mathbb{R}^{(n+m)\times m}$ are continuous functions, and the triple (A, C, f) is detectable. Hence, assuming that (A, B_2, f) is stabilizable, Theorem 4.2 or Corollary 4.3 can be applied to generate a controller.

THEOREM 6.1. *Let (4) hold. Define A, B_1, B_2, C, D_1, and D_2 as above and assume that the triple (A, B_2, f) is stabilizable. Let F be the H^∞ controller such that $\sup_{w\in l_2} \|z\|_{l_2}/\|w\|_{l_2} < \gamma$ for some $\gamma < \infty$. Then there exists an $R_{Capture} > 0$ such that, if $u(k) = F_{\theta(k)}(\varphi(k) - \theta(k))$ and $|\varphi(0) - \theta(0)| < R_{Capture}$, then $|\varphi(k) - \theta(k)| \to 0$ as $k \to \infty$.*

Proof. Define $\bar{\eta}(\bar{x},\bar{u}) := \sup\{\|\tilde{\eta}(x,u,\theta)\| : |x| \le \bar{x}, |u| \le \bar{u}, \theta \in \Theta\}$. By (10) and (11), we obtain

$$\bar{\eta}(\bar{x},\bar{u}) \to 0 \quad \text{as} \quad \bar{x},\bar{u} \to 0. \tag{91}$$

It follows that there exist $x^*, u^* > 0$ such that $\bar{\eta}(x^*,u^*) \le \frac{1}{\gamma}$. Now define

$$h(x,u,\theta) := \begin{cases} \tilde{\eta}(x,u,\theta) & \text{for } |x| < x^* \text{ and } |u| < u^*, \\ 0_{n\times(n+m)} & \text{otherwise.} \end{cases} \tag{92}$$

Thus, for all x and u, $\sup_{\theta\in\Theta}\|h(x,u,\theta)\| \le \frac{1}{\gamma}$. Consider the following closed-loop LDV system:

$$\begin{aligned} \xi(k+1) &= A_{\theta(k)}\xi(k) + B_{1_{\theta(k)}}\omega(k) + B_{2_{\theta(k)}}\upsilon(k), \\ \zeta(k) &= C_{\theta(k)}\xi(k) + D_{2_{\theta(k)}}\upsilon(k), \\ \omega(k) &= h(\xi(k),\upsilon(k),\theta(k))\zeta(k), \\ \upsilon(k) &= F_{\theta(k)}\xi(k), \\ \theta(k+1) &= f(\theta(k)). \end{aligned} \tag{93}$$

Since $\sup_{w\in l_2}\|z\|_{l_2}/\|w\|_{l_2} < \gamma$, the small gain theorem of [32] implies that system (93) is externally l_2 stable. Since $A_\theta + B_{2_\theta}F_\theta$ is uniformly exponentially stable, system (93) is also internally l_2 stable. Therefore there exist a $G_x \ge 1$ and a $G_u > 0$ such that $\|\xi\|_{l_\infty} \le \|\xi\|_{l_2} < G_x|\xi(0)|$ and $\|\upsilon\|_{l_\infty} \le \|\upsilon\|_{l_2} < G_u|\xi(0)|$. Now set

$$R_{Capture} := \min\left(\frac{x^*}{G_x}, \frac{u^*}{G_u}\right).$$

If $|\xi(0)| \le R_{Capture}$, then

$$\|\xi\|_{l_\infty} < G_x|\xi(0)| \le x^*$$

and

$$\|\upsilon\|_{l_\infty} < G_u|\xi(0)| \le u^*.$$

By the above inequalities and (92), we conclude that, for all k, $h(\xi(k),\upsilon(k),\theta(k)) = \tilde{\eta}(\xi(k),\upsilon(k),\theta(k))$. Thus, if

$$|x(0)| < \min\left(\frac{x^*}{G_x}, \frac{u^*}{G_u}\right),$$

then, by the uniqueness of solutions to difference equations, the closed-loop LDV system

$$\begin{aligned} x(k+1) &= A_{\theta(k)}x(k) + B_{1_{\theta(k)}}w(k) + B_{2_{\theta(k)}}u(k), \\ z(k) &= C_{\theta(k)}x(k) + D_{2_{\theta(k)}}u(k), \\ w(k) &= \tilde{\eta}(x(k),u(k),\theta)z(k), \\ u(k) &= F_{\theta(k)}x(k), \\ \theta(k+1) &= f(\theta(k)) \end{aligned} \tag{94}$$

is l_2 stable, and furthermore, $\|x\|_{l_\infty} < x^*$ and $\|u\|_{l_\infty} < u^*$. Since (94) is the tracking error of the closed-loop nonlinear system, we conclude that $|x(k)| = |\varphi(k) - \theta(k)| \to 0$ as $k \to \infty$. ∎

6.2. Further considerations. Writing the nonlinearity $\eta(x, u, \theta)$ as a bounded feedback $\tilde{\eta}(x, u, \theta)$ from an output z to the input w (see (10), (11), (89)) yields an H^∞ design that attenuates the effect of the nonlinearity and hence amplifies the initial allowable tracking error. It is further possible to optimize this procedure by factoring $\tilde{\eta}$ in such a way that $\|\tilde{\eta}\| \leq 1$ (see [7] for details). The suboptimal H^∞ controller is continuous, and therefore an approximation of the LDV H^∞ controller can be constructed in the same way that an approximation of the LDV quadratic controller was constructed in [8]. The fact that the H^∞ controller is guaranteed to be continuous under the condition that it be suboptimal does not prove that the suboptimality condition for continuity is necessary. In fact, an example based on the Hénon map shows that the suboptimal controller becomes discontinuous as γ approaches γ_o, the optimal H^∞ tolerance.

7. Conclusion. Suboptimal H^∞ controllers for LDV systems have been developed. Like the LDV quadratic controllers, these H^∞ controllers are continuous functions. The H^∞ method has distinct advantages over the LQ method, in that H^∞ can be tuned to minimize the effect of linearization and it is possible to find a lower bound on the maximum allowable initial tracking error.

REFERENCES

[1] H. Abou-Kandil, G. Freiling, and G. Jank, *Solution and asymptotic behavior of coupled Riccati equations in jump linear systems*, IEEE Trans. Automat. Control, 39 (1994), pp. 1631–1636.

[2] H. Abou-Kandil, G. Freiling, and G. Jank, *On the solution of discrete-time Markovian jump linear quadratic control problems*, Automatica J. IFAC, 31 (1995), pp. 765–768.

[3] P. Apkarian, *Advanced gain-scheduling techniques for uncertain systems*, IEEE Trans. Control Syst. Technol., 6 (1998), pp. 21–32.

[4] P. Apkarian, P. Gahinet, and G. Becker *Self-scheduled H_∞ control of linear parameter-varying systems: Design example*, Automatica J. IFAC, 31 (1995), pp. 1251–1261.

[5] G. Becker and A. Packard, *Robust performance of linear parameterically varying systems using parametrically-dependent linear feedback*, Systems Control Lett., 23 (1995), pp. 205–215.

[6] G. Becker, A. Packard, and G. Balas, *Control of parametrically-dependent linear systems: A single quadratic Lyapunov approach*, in American Control Conference, 1993, pp. 2795–2799.

[7] S. Bohacek, *Controlling Systems with Complicated Asymptotic Dynamics with Linear Dynamically Varying Control*, Ph.D. thesis, University of Southern California, Los Angeles, CA, 1999.

[8] S. Bohacek and E. Jonckheere, *Linear dynamically varying LQ control of nonlinear systems over compact sets*, IEEE Trans. Automat. Control, 46 (2001), pp. 840–852.

[9] S. Bohacek and E. Jonckheere, *Linear dynamically varying H^∞ control chaos*, in *Nonlinear Control Systems Design Symposium (NOLCOS)*, Enschede, The Netherlands, IFAC, 1998, pp. 744–749.

[10] S. Bohacek and E. Jonckheere, *Linear dynamically varying (LDV) systems versus jump linear systems*, in Proceedings of the AACC American Control Conference, San Diego, CA, 1999, pp. 4024–4028.

[11] S. Bohacek and E. Jonckheere, *Analysis and synthesis of linear set valued dynamically varying (LSVDV) systems*, in Proceedings of the AACC American Control Conference, Chicago, IL, 2000, pp. 2770–2774.

[12] I. Borno and Z. Gajic, *Parallel algorithm for solving coupled algebraic Lyapunov equations of discrete-time jump linear systems*, Comput. Math. Appl., 30 (1995), pp. 1–4.

[13] W. Bown, *Mathematicians learn how to tame chaos*, New Scientist, 134 (1992), p. 16.

[14] S. Boyd, L. E. Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM, Philadelphia, PA, 1994.

[15] M. Fragoso, J. Ribeiro do Val, and D. Pinto, *Jump linear H_∞ control: The discrete-time case*, Control Theory Adv. Tech., 10 (1995), pp. 1459–1474.

[16] P. Gahinet, P. Apkarian, and M. Chilali, *Affine parameter-dependent Lyapunov function for real parameter uncertainty*, in Proceedings of the 33rd Conference on Decision and Control, Lake Buena Vista, FL, IEEE, 1994, pp. 2026–2031.
[17] M. Green and D. Limebeer, *Linear Robust Control*, Prentice–Hall, Englewood Cliffs, NJ, 1995.
[18] A. Halanay and V. Ionescu, *Time-Varying Discrete Linear Systems*, Birkhäuser-Verlag, Basel, Switzerland, 1994.
[19] Y. Ji and H. Chizeck, *Jump linear quadratic Gaussian control: Steady-state solution and testable conditions*, Control Theory Adv. Tech., 10 (1990), pp. 1459–1474.
[20] J. Burken,*Private communication*, NASA Dryden Flight Research Center, Edwards, CA, 1999.
[21] E. Jonckheere, P. Lohsoonthorn, and S. Bohacek, From Sioux-City to the X-33, in *Annu. Rev. Control*, 23 (1999), pp. 91–108.
[22] A. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Cambridge University Press, London, UK, 1995.
[23] I. E. Kose, F. Jabbari, and W. E. Schmitendorf, *A direct characterization of L_2-gain controllers for LPV systems*, in Proceedings of the 35th Conference on Decision and Control, Kobe, Japan, IEEE, 1996, pp. 3990–3995.
[24] J. R. Munkres, *Topology*, Prentice–Hall, Englewood Cliffs, NJ, 1975.
[25] A. Packard, *Gain scheduling via linear fractional transformations* Systems Control Lett., 22 (1994), pp. 79–92.
[26] M. A. Peters and P. A. Iglesias, *Minimum Entropy Control for Time-Varying Systems*. Birkhäuser, Boston, Cambridge, MA, 1997.
[27] S. Y. Pilyugin, *The Space of Dynamical Systems with the C^0-Topology*. Springer-Verlag, New York, 1991.
[28] J. S. Shamma and M. Athans, *Guaranteed properties of gain scheduled control for linear parameter-varying plants*, Automatica J. IFAC, 27 (1991), pp. 559–564.
[29] J. S. Shamma and D. Xiong, *Set-valued methods for linear parameter varying systems*, Automatica J. IFA, 35 (1999), pp. 1081–1089.
[30] A. Stoorvogel, *The H(infinity) Control Problem: A State Space Approach*, Prentice–Hall, Englewood Cliffs, NJ, 1992.
[31] A. A. Stoorvogel and A. J. T. Weeren, *The discrete-time Riccati equation related to the H_∞ control problem*, IEEE Trans. Automat. Control, 39 (1994), pp. 686–691.
[32] M. Vidyasagar, *Nonlinear Systems Analysis*, Prentice–Hall, Englewood Cliffs, NJ., 1993.
[33] F. Wu, X. H. Yang, A. Packard, and G. Becker, *Induced l-2 norm control for LPV systems with bounded parameter variation rates*, Internat. J. Robust Nonlinear Control, 6 (1996), pp. 983–998.
[34] J. Yu and A. Sideris, *H_∞ control with parameteric Lyapunov functions*, Systems Control Lett., 30 (1997), pp. 57–69.
[35] J. Yu and A. Sideris, *H_∞ control with parametric lyapunov functions*, in Proceedings of the 34th Conference on Decision and Control, New Orleans, LA, IEEE, 1995, pp. 2547–2552.
[36] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control*, Prentice–Hall, Upper Saddle River, NJ, 1996.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 4, pp. 1072–1086

A CONSTRUCTIVE SOLUTION TO INTERCONNECTION AND DECOMPOSITION PROBLEMS WITH MULTIDIMENSIONAL BEHAVIORS*

E. ZERZ† AND V. LOMADZE‡

Abstract. By a multidimensional behavior, we mean the solution space of a linear constant-coefficient system of partial difference or differential equations. Within the behavioral framework, a natural concept of interconnection has been introduced by J. C. Willems. The regular interconnection problem can be formulated as follows: Given a behavior (the plant), find—if possible—another behavior (a controller) such that their interconnection is not redundant and is equal to a certain desired behavior. We give a constructive solution to this problem, which is in terms of polynomial matrix equations that can be solved using the theory of Gröbner bases. We also study a dual problem concerning the existence of direct sum decompositions of multidimensional behaviors. Finally, we present a unified framework for studying both problems as special cases of constructing relative complements in the lattice of behaviors.

Key words. multidemensional systems, behavioral approach, regular interconnection, direct sum decomposition, multivariate polynomial modules, split exact sequences

AMS subject classifications. 93B25, 93C05, 93C20, 06C, 13C, 13D

PII. S0363012900374749

1. Introduction.

Feedback control is based on the interconnection of systems: Given a dynamical system (the plant), the goal of feedback control is to design another system (a controller) in such a way that the interconnection of the two systems has certain desired properties. As an example, consider a plant given in classical state space form

$$\dot{x} = Ax + Bu$$

and let the controller be specified by the feedback law $u = Fx + v$. Then the interconnection (the closed loop system) is

$$\dot{x} = (A + BF)x + Bv. \tag{1.1}$$

A typical aim of the controller design in this setting is spectral assignment; that is, a condition is given on the desired location of the eigenvalues of $A + BF$.

Note that interconnection means nothing more than combining the equations that determine plant and controller, respectively, and looking at their common solutions. For instance, combining the plant given by

$$\begin{bmatrix} \frac{d}{dt}I - A, & -B, & 0 \end{bmatrix} \begin{bmatrix} x \\ u \\ v \end{bmatrix} = 0$$

*Received by the editors July 3, 2000; accepted for publication (in revised form) May 2, 2001; published electronically November 28, 2001.
http://www.siam.org/journals/sicon/40-4/37474.html
†Department of Mathematics, University of Kaiserslautern, 67663 Kaiserslautern, Germany (zerz@mathematik.uni-kl.de).
‡Institute of Mathematics, M. Aleksidze str. 1, Tbilisi 380093, Georgia (loma@rmi.acnet.ge).

with the controller given by

$$\begin{bmatrix} F, & -I, & I \end{bmatrix} \begin{bmatrix} x \\ u \\ v \end{bmatrix} = 0$$

yields the interconnected system

$$\begin{bmatrix} \frac{d}{dt}I - A, & -B, & 0 \\ F, & -I, & I \end{bmatrix} \begin{bmatrix} x \\ u \\ v \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

from which we may eliminate u to get (1.1). It is worth noting that this interconnection is regular in a sense to be defined below.

The behavioral approach to systems theory developed by J. C. Willems—see [14] for a survey—provides an elegant framework for dealing with such interconnection problems as well as more general ones. A behavior is the set of signals w that obey a certain system law, say a system of linear differential equations, which may be written as

$$P_1(\tfrac{d}{dt})w = 0.$$

In the example above, $w = (x, u, v)^t$, where $(\cdot)^t$ denotes transposition, and

$$P_1\left(\tfrac{d}{dt}\right) = \begin{bmatrix} \frac{d}{dt}I - A, & -B, & 0 \end{bmatrix}.$$

Similarly, let the controller be determined by $P_2(\frac{d}{dt})w = 0$. Then the interconnection is determined by the system law

$$\begin{bmatrix} P_1 \\ P_2 \end{bmatrix} (\tfrac{d}{dt})\ w = 0.$$

In what follows, we will consider the multidimensional case, that is, systems of partial instead of ordinary differential equations.

A dual problem concerns direct sum decompositions of behaviors: Is it possible to write a behavior as the superposition of two smaller systems such that each trajectory of the overall behavior has a unique representation as the sum of two trajectories of the respective building blocks? A prominent example from classical one-dimensional systems theory is the so-called controllable-autonomous decomposition. Consider $\dot{x} = Ax + Bu$. In a suitably chosen basis of the state space, these equations take the form

$$\frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} A_1 & A_2 \\ 0 & A_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} B_1 \\ 0 \end{bmatrix} u,$$

where (A_1, B_1) is a reachable matrix pair. This is the well-known Kalman reachability decomposition. Now each trajectory $w = (x_1, x_2, u)^t$ can be uniquely written in the form $w = w_1 + w_2$, where w_1 belongs to the controllable part of the system (governed by $\dot{x}_1 = A_1x_1 + B_1u$, $x_2 = 0$) and w_2 belongs to a suitably defined autonomous part of the system.

For one-dimensional behaviors, i.e., systems governed by ordinary differential equations, the regular interconnection problem was formulated and solved by J. C. Willems [14, 15]. In the multidimensional situation, the problem was first tackled by

P. Rocha and J. Wood [10, 11], who provided a foundation for the control theory of multidimensional behaviors, based on prior work of U. Oberst [9]. Several authors have noted that the controllable-autonomous decomposition is, in general, no longer a direct sum decomposition in dimensions greater than one [6]. A more general treatment of decompositions of two-dimensional systems has been offered by M. E. Valcher and M. Bisiacco [2, 13]. Both problems can be treated within the unified framework of lattice theory; see [9, 12].

This paper is organized as follows. In section 2, we give some introductory material as well as the main mathematical tool for the subsequent discussion. We use this tool to solve the regular interconnection problem in section 3. Section 4 is devoted to the dual problem of direct sum decompositions. Finally, a unified approach to studying both problems is proposed in section 5. We conclude with three worked examples which are collected in section 6.

2. Preliminaries. Let K be a field, n a positive integer, and $\mathcal{D} = K[s_1, \ldots, s_n]$ the polynomial ring over K in n indeterminates $s_1, \ldots, s_n$.

By a linear n-dimensional system Σ with signal number q, we mean a submodule N of $\mathcal{D}^q$. Such a module admits a representation

$$N = \operatorname{im}(P^t) = P^t\mathcal{D}^g \quad \text{for some } P \in \mathcal{D}^{g\times q},$$

where g is a suitable integer and P^t is the transpose of P. The matrix P is called a representation of Σ.

The connection with behaviors, as described in the introduction, is as follows. Let $P \in \mathcal{D}^{g\times q}$ be given. Consider the solution space of the associated system of linear constant-coefficient partial differential equations, that is,

$$\mathcal{B} = \ker_{\mathcal{A}}(P) = \{w \in \mathcal{A}^q \mid P(\underline{\partial})w := P(\partial_1, \ldots, \partial_n)w = 0\}.$$

The signal space $\mathcal{A}$ is either the space of smooth functions or of distributions on $\mathbb{R}^n$. (The base field K is assumed to be either $\mathbb{R}$ or $\mathbb{C}$.) Oberst [9] showed that $\mathcal{B} \subseteq \mathcal{A}^q$ and the submodule of $\mathcal{D}^q$ generated by the columns of P^t are in fact equivalent data, and this motivates our definition of a system given above. Indeed,

$$\mathcal{B} = \ker_{\mathcal{A}}(P) \quad \rightarrow \quad M(\mathcal{B}) := \{p^t \in \mathcal{D}^q \mid p(\underline{\partial})w = 0 \ \forall w \in \mathcal{B}\}$$

and

$$\mathcal{B}(N) := \{w \in \mathcal{A}^q \mid p(\underline{\partial})w = 0 \ \forall p^t \in N\} \quad \leftarrow \quad N = \operatorname{im}(P^t) = P^t\mathcal{D}^g$$

are order-reversing bijections between behaviors in $\mathcal{A}^q$ and submodules of $\mathcal{D}^q$, and they are inverse to each other. We have the correspondences

$$\mathcal{B}_1 \cap \mathcal{B}_2 \quad \leftrightarrow \quad N_1 + N_2$$

(that is to say, $M(\mathcal{B}_1 \cap \mathcal{B}_2) = M(\mathcal{B}_1) + M(\mathcal{B}_2)$ and $\mathcal{B}(N_1 + N_2) = \mathcal{B}(N_1) \cap \mathcal{B}(N_2)$) and similarly

$$\mathcal{B}_1 + \mathcal{B}_2 \quad \leftrightarrow \quad N_1 \cap N_2.$$

A similar interpretation is valid for discrete systems, i.e., for behaviors given by difference instead of differential equations. For the sake of simplicity, we will restrict our analysis to the continuous case throughout this paper.

Given two behaviors $\mathcal{B}_1, \mathcal{B}_2 \subseteq \mathcal{A}^q$, their *interconnection* $\mathcal{B}$ is defined as [15]

$$\mathcal{B} = \mathcal{B}_1 \cap \mathcal{B}_2.$$

Thus, the interconnected behavior consists of the trajectories that obey the system laws of both $\mathcal{B}_1$ and $\mathcal{B}_2$; that is, if $\mathcal{B}_i = \ker_{\mathcal{A}}(P_i)$, then $\mathcal{B} = \ker_{\mathcal{A}}(P)$ with

$$P = \begin{bmatrix} P_1 \\ P_2 \end{bmatrix}. \tag{2.1}$$

The module-theoretic counterpart is the following: Given two systems Σ_1 and Σ_2 with the same signal number q, we define their interconnection $\Sigma := \Sigma_1 \wedge \Sigma_2$ by

$$N := N_1 + N_2.$$

A representation P of Σ is obtained from representations P_i of Σ_i by (2.1).

The interconnection is said to be *regular* if we have

$$N = N_1 \oplus N_2,$$

that is, if $N_1 \cap N_2 = \{0\}$. In terms of behaviors, this signifies

$$\mathcal{B}_1 + \mathcal{B}_2 = \mathcal{A}^q,$$

and in terms of representations, an interconnection (2.1) is regular if

$$\operatorname{rank}(P) = \operatorname{rank}(P_1) + \operatorname{rank}(P_2).$$

We say that a system Σ_1 is *less powerful* than Σ if $N_1 \subseteq N$. This implies the reverse relation for the corresponding behaviors: $\mathcal{B}_1 \supseteq \mathcal{B}$. In terms of representations, this means $P_1 = CP$ for some polynomial matrix C. Certainly, Σ_1 and Σ_2 themselves are always less powerful than $\Sigma_1 \wedge \Sigma_2$.

The basic mathematical tool for the following discussion is the notion of a *split exact sequence*. Let R be a commutative ring, and let L, L_1, L_2 be R-modules. Let $f : L_1 \to L$ and $g : L \to L_2$ be R-linear maps. The sequence

$$0 \to L_1 \xrightarrow{f} L \xrightarrow{g} L_2 \to 0 \tag{2.2}$$

is called *exact* if f is injective, g is surjective, and $\operatorname{im}(f) = \ker(g)$. An exact sequence (2.2) is said to be *split* if the following equivalent conditions are satisfied:

1. There exists a homomorphism $k : L \to L_1$ such that kf is the identity map on L_1, denoted by id_{L_1}.
2. There exists a homomorphism $h : L_2 \to L$ such that $gh = \mathrm{id}_{L_2}$.
3. $\operatorname{im}(f) = \ker(g)$ is a direct summand of L.

THEOREM 2.1. *Let R be a commutative ring. Let F_0, F_1 be finitely generated free R-modules, and let L, L_1, L_2 be arbitrary R-modules. Suppose that the following diagram is commutative with exact rows:*

$$\begin{array}{ccccccccc} & & F_1 & \xrightarrow{\alpha} & F_0 & \xrightarrow{\beta} & L_2 & \to & 0 \\ & & & & \pi \downarrow & & \| & & \\ 0 & \to & L_1 & \xrightarrow{f} & L & \xrightarrow{g} & L_2 & \to & 0. \end{array}$$

Then the following are equivalent:

1. *The lower sequence is split.*
2. *There exists a homomorphism* $\xi : F_0 \to L_1$ *such that*

$$f\xi\alpha = \pi\alpha. \tag{2.3}$$

Proof. $1 \Rightarrow 2$. Suppose there exists a homomorphism $h : L_2 \to L$ such that $gh = \mathrm{id}$. Then $g\pi = \beta = gh\beta$, and thus

$$\mathrm{im}(\pi - h\beta) \subseteq \ker(g) = \mathrm{im}(f).$$

Let B be a basis of F_0. Since f is injective, there exists, for any $b \in B$, a unique $\xi(b) \in L_1$ such that

$$f(\xi(b)) = (\pi - h\beta)(b).$$

This defines a homomorphism $\xi : F_0 \to L_1$ with

$$f\xi\alpha = (\pi - h\beta)\alpha = \pi\alpha.$$

$2 \Rightarrow 1$. Let $\xi : F_0 \to L_1$ be such that $(\pi - f\xi)\alpha = 0$, that is,

$$\ker(\beta) = \mathrm{im}(\alpha) \subseteq \ker(\pi - f\xi).$$

Then there is a well-defined homomorphism $h : L_2 \to L$ that assigns to each $l_2 = \beta(x)$, independently of the choice of such an $x \in F_0$,

$$h(l_2) := (\pi - f\xi)(x).$$

We have $gh(l_2) = g(\pi - f\xi)(x) = g\pi(x) = \beta(x) = l_2$ for all $l_2 \in L_2$. ∎

There is a close connection between the above theorem and the theory of extension modules. The reader is referred to [5] for a general introduction to this concept, and to [8] for its application to the regular interconnection problem.

3. The regular interconnection problem. We will consider the following control problem: Given Σ_1 and Σ, with Σ_1 less powerful than Σ, does there exist a system Σ_2 such that

$$\Sigma_1 \wedge \Sigma_2 = \Sigma$$

and the interconnection is regular? If yes, we say that Σ can be *achieved* from Σ_1 by regular interconnection. Moreover, in that case, we are looking for an explicit construction of such a Σ_2.

One should think of these systems as follows: Σ_1 is the given plant, Σ is a specified desired system, and Σ_2 is the controller to be constructed.

For this, let $P_1 \in \mathcal{D}^{g_1 \times q}$ and $P \in \mathcal{D}^{g \times q}$ be representations of Σ_1 and Σ, respectively, and define $N_1 := \mathrm{im}(P_1^t)$, $N := \mathrm{im}(P^t)$. By assumption, $N_1 \subseteq N$, that is, $P_1 = CP$ for some polynomial matrix C. Define $M := \mathrm{coker}(P^t) = \mathcal{D}^q/N$ and $M_1 := \mathrm{coker}(P_1^t)$.

The problem is to decide whether N_1 is a direct summand of N and, if so, to construct a complementary summand N_2. By the definition of a split exact sequence, the regular interconnection problem is solvable, i.e., N_1 is a direct summand of N if and only if the exact sequence

$$0 \to N_1 \hookrightarrow N \to N/N_1 \to 0$$

is split. In view of Theorem 2.1, consider the diagram

$$\begin{array}{ccccccccc} & & & & \mathcal{D}^k & \xrightarrow{A} & \mathcal{D}^g & \to & N/N_1 & \to & 0 \\ & & & & & & {\scriptstyle P^t}\downarrow & & \| & & \\ 0 & \to & N_1 & \hookrightarrow & N & \to & N/N_1 & \to & 0. & & \end{array}$$

The construction of A with the desired property, that is, exactness of the upper sequence, or equivalently

$$\mathcal{D}^g/\mathrm{im}(A) \cong N/N_1,$$

will be discussed in section 3.1. Note that the matrix representation of any homomorphism $\xi : \mathcal{D}^g \to N_1 = \mathrm{im}(P_1^t)$ has the form $P_1^t Y$, where $Y \in \mathcal{D}^{g_1 \times g}$. Thus we can write (2.3) in terms of the following polynomial matrix equation:

$$(P_1^t Y - P^t)A = 0. \tag{3.1}$$

The regular interconnection problem is solvable if and only if this linear matrix equation possesses a polynomial solution Y. In [8], this condition was derived using the concept of extension modules. Here, we have given an alternative proof without making use of that tool.

To check the solvability of (3.1), we rewrite it using the *Kronecker product* of two matrices; see, e.g., [3]. This is defined as follows: Let

$$A = \begin{bmatrix} a_{11} & \dots & a_{1l} \\ \vdots & & \vdots \\ a_{k1} & \dots & a_{kl} \end{bmatrix}$$

be a $k \times l$ matrix, and let B be an $m \times n$ matrix; then the Kronecker product $A \otimes B$ is the $km \times ln$ matrix defined by

$$A \otimes B = \begin{bmatrix} a_{11}B & \dots & a_{1l}B \\ \vdots & & \vdots \\ a_{k1}B & \dots & a_{kl}B \end{bmatrix}.$$

The vector $\mathrm{vec}(\cdot)$ is obtained from a matrix by stacking its columns into one long column vector, that is,

$$\mathrm{vec}(A) = [a_{11}, \dots, a_{k1}, \dots, a_{1l}, \dots, a_{kl}]^t.$$

For any three matrices A, B, Y of compatible size, we have

$$\mathrm{vec}(BYA) = (A^t \otimes B)\mathrm{vec}(Y).$$

Using this in (3.1), we get

$$(A^t \otimes P_1^t)\,\mathrm{vec}(Y) = \mathrm{vec}(P^t A). \tag{3.2}$$

This reduces the problem to checking whether $\mathrm{vec}(P^t A)$ is contained in the module spanned by the columns of $A^t \otimes P_1^t$. Also this question can be answered algorithmically, and the computations are based on Gröbner basis techniques, as discussed in the next section.

3.1. Computational issues. We need the following basic tools from the constructive theory of Gröbner bases, which was developed by Buchberger [4]; see also [1]. We list the algorithms briefly, and without detail.

Division with remainder. Given a polynomial matrix $P \in \mathcal{D}^{g\times q}$ and a vector $f \in \mathcal{D}^q$, this algorithm constructs $h \in \mathcal{D}^g$ and $r \in \mathcal{D}^q$ such that

$$f = P^t h + r \quad \text{and} \quad r \text{ is in normal form modulo } P. \tag{3.3}$$

Note that, in general, neither h nor the "remainder" r are uniquely determined.

Deciding module membership. We call P a Gröbner matrix if the columns of P^t are a Gröbner basis of the module $N = \mathrm{im}(P^t)$ that they generate. If P is a Gröbner matrix, then f admits a representation $f = P^t h$ (that is, $f \in N$) if and only if $r = 0$ in one, or equivalently, in all, representations (3.3). For any matrix P, we may compute a Gröbner matrix P_G such that $\mathrm{im}(P_G^t) = \mathrm{im}(P^t)$. Thus module membership is decidable for arbitrary polynomial matrices.

Computing minimal annihilators. Given a polynomial matrix $P \in \mathcal{D}^{g\times q}$, this algorithm yields a matrix Q such that $\ker(P^t) = \mathrm{im}(Q^t)$; that is,

1. $QP = 0$, and
2. any polynomial matrix Q_1 with $Q_1 P = 0$ has a representation $Q_1 = XQ$, where X is a polynomial matrix.

We call Q a minimal left annihilator of P.

Application to the regular interconnection problem. Given polynomial matrices $P \in \mathcal{D}^{g\times q}$ and $P_1 \in \mathcal{D}^{g_1\times q}$ with $\mathrm{im}(P_1^t) \subseteq \mathrm{im}(P^t)$, there exists a polynomial matrix $C \in \mathcal{D}^{g_1\times g}$ such that $P_1 = CP$. Such a C can be constructed by means of the division with remainder algorithm.

Next, we look for a matrix $A \in \mathcal{D}^{g\times k}$ (where k is suitably chosen) such that

$$\mathrm{coker}(A) = \mathcal{D}^g/\mathrm{im}(A) \cong \mathrm{im}(P^t)/\mathrm{im}(P_1^t) = P^t\mathcal{D}^g/P^tC^t\mathcal{D}^{g_1}.$$

We give the following easy lemma without proof.

LEMMA 3.1. *There is an isomorphism*

$$P^t\mathcal{D}^g/P^tC^t\mathcal{D}^{g_1} \cong \mathcal{D}^g/(\ker(P^t) + C^t\mathcal{D}^{g_1}).$$

Let Q be a minimal left annihilator of P, that is, $\ker(P^t) = \mathrm{im}(Q^t)$. Then the preceding lemma suggests the definition

$$A = \begin{bmatrix} Q^t, & C^t \end{bmatrix}, \tag{3.4}$$

and this will be used in Theorem 3.2 below.

3.2. Constructing controllers. In the case where the regular interconnection problem is solvable for given systems Σ and Σ_1, the following theorem yields a concrete formula for a controller Σ_2; that is, it allows us to construct a representation P_2 of Σ_2 directly from the data, i.e., from P and P_1. Note that the expression for A found in (3.4) is used to reformulate (3.1).

THEOREM 3.2. *Let P and P_1 be polynomial matrices with $\mathrm{im}(P_1^t) \subseteq \mathrm{im}(P^t)$. Let C be such that $P_1 = CP$, and let Q be a minimal left annihilator of P. The following are equivalent:*

1. *There exists a polynomial matrix P_2 such that*

$$\mathrm{im}(P^t) = \mathrm{im}(P_1^t) \oplus \mathrm{im}(P_2^t). \tag{3.5}$$

2. *The matrix equation*

$$(P_1^t Y - P^t) \begin{bmatrix} Q^t, & C^t \end{bmatrix} = 0 \tag{3.6}$$

possesses a polynomial solution Y.

Moreover, if such a Y *exists, then* $P_2^t := P_1^t Y - P^t$ *is a matrix satisfying* (3.5).

Proof. It suffices to show that $P_2^t = P_1^t Y - P^t$ satisfies (3.5). As $P_1^t = P^t C^t$, we have $P_2^t = P^t(C^t Y - I)$. It is easy to see that $\operatorname{im}(P^t) = \operatorname{im}(P_1^t) + \operatorname{im}(P_2^t)$ as

$$\begin{bmatrix} P_1^t, & P_2^t \end{bmatrix} = P^t \begin{bmatrix} C^t, & C^t Y - I \end{bmatrix}$$

and

$$P^t = \begin{bmatrix} P_1^t, & P_2^t \end{bmatrix} \begin{bmatrix} Y \\ -I \end{bmatrix}.$$

It remains to show that $\operatorname{im}(P_1^t) \cap \operatorname{im}(P_2^t) = \{0\}$. Suppose that

$$a = P_1^t x = P_2^t y = (P_1^t Y - P^t) y.$$

Define $z := Yy - x$; then

$$P^t C^t z = P_1^t z = P^t y.$$

Thus $y - C^t z \in \ker(P^t) = \operatorname{im}(Q^t)$, that is, $y \in \operatorname{im}[Q^t, C^t]$, and hence, due to (3.6), $a = (P_1^t Y - P^t) y = 0$. □

Note that in the situation of Theorem 3.2, (3.6) can be rewritten as

$$(P_1^t Y - P^t) Q^t = P_1^t Y Q^t = 0 \quad \text{and} \quad (P_1^t Y - P^t) C^t = P_1^t (Y C^t - I) = 0.$$

This has interesting consequences for the one-dimensional case ($n = 1$), as discussed in the following corollary. In particular, we recover a result of Willems [15]. Recall that a univariate polynomial matrix is said to be *left prime* if it possesses a polynomial right inverse.

COROLLARY 3.3. *Let* P, P_1, C, Q *be as described above. The regular interconnection problem is solvable if and only if there exists a polynomial matrix* Y *such that*

$$P_1^t Y Q^t = 0 \quad \textit{and} \quad P_1^t (Y C^t - I) = 0.$$

In the one-dimensional case ($n = 1$), *we may assume without loss of generality that* P *and* P_1 *have full row rank. Then* $Q = 0$ *and the equations simplify to*

$$Y C^t = I;$$

i.e., the regular interconnection problem is solvable if and only if C *is left prime. Certainly, left primeness of* P_1 *implies left primeness of* C. *Hence, the regular interconnection problem is always solvable provided that* P_1 *is left prime. In terms of behaviors, any subsystem of a controllable behavior* $\mathcal{B}_1$ *can be achieved from* $\mathcal{B}_1$ *by regular interconnection* [15, Theorem 6].

4. The direct sum decomposition problem. In the previous section, we have considered the direct sum decomposition

$$N_1 + N_2 = N \quad \text{and} \quad N_1 \cap N_2 = \{0\},$$

where N_1, N_2, N are submodules of $\mathcal{D}^q$. The dual problem, corresponding to a direct sum decomposition of the associated behaviors, is

$$N_1 \cap N_2 = N \quad \text{and} \quad N_1 + N_2 = \mathcal{D}^q.$$

THEOREM 4.1. *Let P_1 and P be polynomial matrices with*

$$N = \operatorname{im}(P^t) \subseteq N_1 = \operatorname{im}(P_1^t) \subseteq \mathcal{D}^q.$$

The following are equivalent:

1. *There exists a polynomial matrix P_2 such that*

$$\operatorname{im}(P_1^t) \cap \operatorname{im}(P_2^t) = \operatorname{im}(P^t) \quad \textit{and} \quad \operatorname{im}(P_1^t) + \operatorname{im}(P_2^t) = \mathcal{D}^q. \tag{4.1}$$

2. *There exists a polynomial matrix P_2 such that*

$$\operatorname{coker}(P^t) \cong \operatorname{coker}(P_1^t) \oplus \operatorname{coker}(P_2^t).$$

   In particular, $N_1/N \cong \operatorname{coker}(P_2^t)$ is a direct summand of $\mathcal{D}^q/N = \operatorname{coker}(P^t)$.

3. *There exist polynomial matrices X and Y such that*

$$P^t X + P_1^t Y P_1^t = P_1^t. \tag{4.2}$$

Proof. The implication $1 \Rightarrow 2$ is straightforward. For $2 \Rightarrow 3$, we again invoke Theorem 2.1. Consider the diagram

$$\begin{array}{ccccccccc} & & & & \mathcal{D}^{g_1} & \xrightarrow{P_1^t} & \mathcal{D}^q & \to & \operatorname{coker}(P_1^t) & \to & 0 \\ & & & & & & \| & & \| & & \\ 0 & \to & N_1 & \hookrightarrow & \mathcal{D}^q & \to & \mathcal{D}^q/N_1 & \to & 0. & & \end{array}$$

The splitting condition for the lower sequence reads

$$(P_1^t Y - I)P_1^t = 0.$$

However, we are actually interested in the splitting of

$$0 \to N_1/N \hookrightarrow \mathcal{D}^q/N \to \mathcal{D}^q/N_1 \to 0;$$

that is, the condition becomes

$$(P_1^t Y - I)P_1^t \equiv 0 \quad \text{modulo} \quad N = \operatorname{im}(P^t).$$

That is to say,

$$(P_1^t Y - I)P_1^t + P^t X = 0$$

for some polynomial matrix X.

Finally, for showing $3 \Rightarrow 1$, suppose that (4.2) is satisfied. We show that

$$P_2^t = \begin{bmatrix} P_1^t Y - I, & P^t \end{bmatrix}$$

satisfies (4.1). By construction, $\mathrm{im}(P^t) \subseteq \mathrm{im}(P_1^t) \cap \mathrm{im}(P_2^t)$. For the converse direction, let $a = P_1^t x = P_2^t y = (P_1^t Y - I)y_1 + P^t y_2$. This implies that $y_1 \in \mathrm{im}(P_1^t)$; say, $y_1 = P_1^t z$. Then

$$a = P_1^t x = (P_1^t Y P_1^t - P_1^t)z + P^t y_2 = -P^t X z + P^t y_2,$$

showing that $a \in \mathrm{im}(P^t)$ as desired.

To see that $\mathcal{D}^q = \mathrm{im}(P_1^t) + \mathrm{im}(P_2^t)$, it suffices to note that any $x \in \mathcal{D}^q$ can be written as $x = P_1^t Y x + (I - P_1^t Y)x$. The first summand is contained in $\mathrm{im}(P_1^t)$, and the second in $\mathrm{im}(P_2^t)$. □

Again, condition (4.2) can be tested by means of the Kronecker product. Rewrite (4.2) as

$$(I \otimes P^t)\mathrm{vec}(X) + (P_1 \otimes P_1^t)\mathrm{vec}(Y) = \mathrm{vec}(P_1^t).$$

The test amounts to checking whether $\mathrm{vec}(P_1^t)$ is in the module spanned by the columns of

$$\left[\ I \otimes P^t, \quad P_1 \otimes P_1^t\ \right].$$

4.1. Constructing complementary behaviors. It is worth noting that the proof of Theorem 4.1 contains an actual construction procedure for the desired complementary cokernel module (if it exists). The following corollary summarizes the resulting decomposition result for behaviors

$$\mathcal{B} = \ker_{\mathcal{A}}(P) = \{w \in \mathcal{A}^q \mid P(\partial_1, \ldots, \partial_n)w = 0\},$$

where $P \in \mathcal{D}^{g \times q}$. The signal space $\mathcal{A}$ is supposed to satisfy Oberst's duality [9]. This is true, e.g., for the spaces of smooth functions or of distributions on $\mathbb{R}^n$.

COROLLARY 4.2. *Let $\mathcal{B}_1 \subseteq \mathcal{B} \subseteq \mathcal{A}^q$ be behaviors with kernel representations P_1 and P, respectively. The following are equivalent:*

1. *There exists a behavior $\mathcal{B}_2$ such that $\mathcal{B}_1 \oplus \mathcal{B}_2 = \mathcal{B}$.*
2. *There exist polynomial matrices X and Y such that*

$$P^t X + P_1^t Y P_1^t = P_1^t.$$

Moreover, if such matrices X, Y exist, a complementary behavior may be constructed as follows:

$$\mathcal{B}_2 = \ker_{\mathcal{A}} \begin{bmatrix} Y^t P_1 - I \\ P \end{bmatrix} = \ker_{\mathcal{A}}(Y^t P_1 - I) \cap \mathcal{B}.$$

For the two-dimensional case ($n = 2$), an equivalent condition has been derived by Bisiacco and Valcher [2]; see also Valcher's earlier work [13]. The following corollary concerns the one-dimensional case. We recover a well-known decomposition result on controllable subsystems being direct summands.

COROLLARY 4.3. *Let $n = 1$. Then we may assume that P_1 has full row rank. Let $P = EP_1$; then* (4.2) *reduces to*

$$E^t X + Y P_1^t = I.$$

In particular, when P_1 is left prime, this equation is always solvable (take $X = 0$ and let Y be a polynomial left inverse of P_1^t). Thus any controllable subbehavior of $\mathcal{B}$ is a direct summand of $\mathcal{B}$.

5. A unified view of interconnection and decomposition. The two problems discussed in this paper are special cases of finding *relative complements* in the lattice of behaviors [12] in $\mathcal{A}^q$ or in the lattice of submodules of $\mathcal{D}^q$, respectively. The set of submodules of $\mathcal{D}^q$ is partially ordered by inclusion, and it becomes a lattice via

$$\inf(N_1, N_2) = N_1 \cap N_2 \quad \text{and} \quad \sup(N_1, N_2) = N_1 + N_2.$$

Given three modules $N_l \subseteq N_1 \subseteq N_u \subseteq \mathcal{D}^q$, the problem is to decide whether there exists a relative complement N_2 of N_1, that is, a submodule of $\mathcal{D}^q$ with

$$\inf(N_1, N_2) = N_l \quad \text{and} \quad \sup(N_1, N_2) = N_u.$$

In terms of matrix representations, this problem has a solution if and only if there exist polynomial matrices X and Y such that

$$P_l^t X + (P_1^t Y - P_u^t) A = 0,$$

where $A = [Q^t, C^t]$ with $P_1 = CP_u$ and Q being a minimal left annihilator of P_u. A solution is then given by

$$P_2^t = \begin{bmatrix} P_1^t Y - P_u^t, & P_l^t \end{bmatrix}.$$

The problem reduces to the regular interconnection problem when $N_l = \{0\}$ (i.e., $P_l = 0$) and to the direct sum decomposition problem if $N_u = \mathcal{D}^q$ (i.e., without loss of generality, $P_u = I$ and $A = P_1^t$). Specializing even more, we may decide whether a given module $N_1 = \operatorname{im}(P_1^t)$ possesses a complement, that is, a module N_2 with $N_1 \oplus N_2 = \mathcal{D}^q$. This is true if and only if the equation

$$P_1^t Y P_1^t = P_1^t \tag{5.1}$$

has a polynomial solution Y. This is equivalent to the splitting of

$$0 \to N_1 \hookrightarrow \mathcal{D}^q \to \mathcal{D}^q / N_1 \to 0$$

or

$$\operatorname{im}(P_1^t) \oplus \operatorname{coker}(P_1^t) \cong \mathcal{D}^q.$$

In particular, $\operatorname{im}(P^t)$ is projective and hence free, due to the Quillen–Suslin theorem (see, e.g., [16]). Then we may assume without loss of generality that P_1 has full row rank, and thus (5.1) simplifies to

$$Y P_1^t = I.$$

These considerations are summarized in the following corollary.

COROLLARY 5.1. *Let N_1 be a submodule of $\mathcal{D}^q$. The following are equivalent:*

1. *N_1 has a complement; that is, there exists a submodule N_2 of $\mathcal{D}^q$ such that $N_1 \oplus N_2 = \mathcal{D}^q$.*
2. *For any polynomial matrix P_1 with $N_1 = \operatorname{im}(P_1^t)$, there exists a polynomial matrix Y such that*

$$P_1^t Y P_1^t = P_1^t.$$

3. *There exist polynomial matrices* P_1, Y *such that* $N_1 = \mathrm{im}(P_1^t)$ *and*

$$YP_1^t = I.$$

Moreover, in that case, $N_2 = \mathrm{im}(P_2^t)$ *with* $P_2^t = I - P_1^t Y$ *is a complement of* N_1.

In terms of behaviors, we say that $\mathcal{B}_1$ has a complement if there exists $\mathcal{B}_2$ such that $\mathcal{B}_1 \oplus \mathcal{B}_2 = \mathcal{A}^q$. In other words, the question is whether the zero behavior can be achieved from a given behavior $\mathcal{B}_1$ by regular interconnection. Let us call this property *controllability to zero.* Parts of the following result can also be found in [11].

COROLLARY 5.2. *Let* $\mathcal{B}_1$ *be a behavior in* $\mathcal{A}^q$. *The following are equivalent:*

1. $\mathcal{B}_1$ *is controllable to zero.*
2. $\mathcal{B}_1$ *has a complement, i.e., there exists a behavior* $\mathcal{B}_2$ *with* $\mathcal{B}_1 \oplus \mathcal{B}_2 = \mathcal{A}^q$.
3. $\mathcal{B}_1$ *is strongly controllable, i.e., it possesses a zero left prime representation matrix; that is,* $\mathcal{B}_1 = \ker_{\mathcal{A}}(P_1)$ *with* $YP_1^t = I$ *for some polynomial matrix* Y.
4. *For any kernel representation* P_1 *of* $\mathcal{B}_1$,

$$P_1^t Y P_1^t = P_1^t$$

has a polynomial solution Y.

Moreover, if such a Y *exists, then* $\mathcal{B}_2 = \ker_{\mathcal{A}}(P_2)$ *with* $P_2 = I - Y^t P_1$ *is a complement of* $\mathcal{B}_1$.

6. Examples. The first two examples are taken from [11], and the third one is taken from [2].

Example 1. Let

$$P_1 = \begin{bmatrix} s_1^2 - 1, & s_1 - s_2 \end{bmatrix}$$

and

$$P = \begin{bmatrix} s_1 + 1 & s_2 \\ 0 & -s_1 \end{bmatrix}.$$

Certainly, P_1 is less powerful than P, as

$$P_1 = \begin{bmatrix} s_1 - 1, & s_2 - 1 \end{bmatrix} P.$$

As P has full row rank, we may take $Q = 0$ and $A = C^t$; thus

$$A = \begin{bmatrix} s_1 - 1 \\ s_2 - 1 \end{bmatrix}.$$

We compute

$$A^t \otimes P_1^t = \begin{bmatrix} (s_1 - 1)(s_1^2 - 1) & (s_2 - 1)(s_1^2 - 1) \\ (s_1 - 1)(s_1 - s_2) & (s_2 - 1)(s_1 - s_2) \end{bmatrix}$$

and

$$\mathrm{vec}(P^t A) = \begin{bmatrix} s_1^2 - 1 \\ s_1 - s_2 \end{bmatrix}.$$

Equation (3.2) possesses the solution

$$\begin{bmatrix} \frac{1}{s_1 - 1} - \frac{t(s_2 - 1)}{s_1 - 1} \\ t \end{bmatrix}$$

over the field of rational functions (t is a rational parameter). It is impossible to choose t such that the solution becomes polynomial. Thus (3.2) has no polynomial solution, and we conclude that the regular interconnection problem is not solvable in this example.

Example 2. Consider

$$P_1 = \begin{bmatrix} s_1 s_2, & s_1 + 1, & s_2 \end{bmatrix}$$

and

$$P = \begin{bmatrix} 0 & s_1 + s_2 + 1 & s_2 \\ s_1 s_2 & -s_1^2 - s_1 s_2 + 1 & -s_1 s_2 + s_2 \\ -s_1^2 - s_1 s_2 & s_1 + s_2 & 0 \end{bmatrix}.$$

Clearly, P_1 is less powerful than P, as $P_1 = CP$ with

$$C = \begin{bmatrix} s_1, & 1, & 0 \end{bmatrix}.$$

Using the computer algebra system SINGULAR [7], we obtain the following reduced expression for A:

$$A = \begin{bmatrix} -s_2 & s_1 \\ 1 & 1 \\ s_2 & 0 \end{bmatrix}.$$

We compute

$$A^t \otimes P_1^t = \begin{bmatrix} -s_1 s_2^2 & s_1 s_2 & s_1 s_2^2 \\ -s_1 s_2 - s_2 & s_1 + 1 & s_1 s_2 + s_2 \\ -s_2^2 & s_2 & s_2^2 \\ s_1^2 s_2 & s_1 s_2 & 0 \\ s_1^2 + s_1 & s_1 + 1 & 0 \\ s_1 s_2 & s_2 & 0 \end{bmatrix}$$

and

$$\operatorname{vec}(P^t A) = \begin{bmatrix} s_1 s_2 - s_1 s_2^2 - s_1^2 s_2 \\ -s_1 s_2 - s_2 - s_1^2 + 1 \\ -s_2^2 - s_1 s_2 + s_2 \\ s_1 s_2 \\ s_1 + 1 \\ s_2 \end{bmatrix}.$$

It is easily checked that (3.2) is solvable; in fact, a solution is given by

$$Y = \begin{bmatrix} 1, & 1 - s_1, & 0 \end{bmatrix}.$$

As a controller, we get

$$P_2 = \begin{bmatrix} s_1 s_2 & -s_2 & 0 \\ -s_1^2 s_2 & s_1 s_2 & 0 \\ s_1^2 + s_1 s_2 & -s_1 - s_2 & 0 \end{bmatrix}.$$

Noting that the second row is a multiple of the first one, we may replace this solution by

$$P_2 = \begin{bmatrix} s_1 s_2 & -s_2 & 0 \\ s_1^2 & -s_1 & 0 \end{bmatrix}.$$

Using different methods, Rocha and Wood [11] derive another controller for the present example.

Example 3. Let

$$P_1 = [s_1 + s_2 + 1] \quad \text{and} \quad P = \begin{bmatrix} s_1 + 1 \\ s_2 + 1 \end{bmatrix} P_1.$$

In order to test whether $\ker_{\mathcal{A}}(P_1)$ is a direct summand of $\ker_{\mathcal{A}}(P)$, we have to check whether there exist polynomials x_1, x_2, y such that (compare (4.2))

$$(s_1 + s_2 + 1) \begin{bmatrix} s_1 + 1, & s_2 + 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + y(s_1 + s_2 + 1)^2 = s_1 + s_2 + 1.$$

This is obviously solvable; take

$$x_1 = 1, \quad x_2 = 1, \quad y = -1.$$

To find a complementary behavior $\ker_{\mathcal{A}}(P_2)$ as indicated in the proof of Theorem 4.1, we compute

$$P_2 = \begin{bmatrix} yP_1 - 1 \\ P \end{bmatrix} = \begin{bmatrix} -(s_1 + s_2 + 2) \\ (s_1 + 1)(s_1 + s_2 + 1) \\ (s_2 + 1)(s_1 + s_2 + 1) \end{bmatrix}.$$

This representation may be simplified by computing a Gröbner matrix P_{2G} of P_2. We obtain the equivalent solution

$$P_{2G} = \begin{bmatrix} s_1 + 1 \\ s_2 + 1 \end{bmatrix}.$$

7. Conclusion. The behavioral approach to systems theory provides a neat setting for modeling the interconnection of systems. We have considered the following basic linear control problem: Given a system Σ_1 (the plant), does there exist a system Σ_2 (a controller) such that the regular interconnection $\Sigma_1 \wedge \Sigma_2$ equals a certain prescribed system Σ (a desired controlled system)? Based on a homological approach, we have given an algebraic criterion for the solvability of this problem. This criterion boils down to a certain linear matrix equation when dealing with concrete representations of the systems Σ and Σ_1.

Using different methods, Rocha and Wood [11] give another solution of the regular interconnection problem. Their approach is based on a characterization of behaviors with direct sum decompositions, which is in terms of zero skew-coprimeness of polynomial matrices. For bivariate matrices, an analogous characterization comes from Bisiacco and Valcher [2]; see also [13]. The condition used there is actually equivalent to the one given in Corollary 4.2 of the present paper. Indeed, the regular interconnection problem and the direct sum decomposition problems are dual. The present paper provides a unified framework for treating both in terms of relative complements in the lattice of behaviors.

Moreover, our solution to the regular interconnection problem features an appealing simplicity. In particular, the polynomial matrix equation (3.1) allows us to decide whether the problem is solvable, by means of basic tools from the theory of Gröbner bases: computing minimal annihilators, division with remainder, and testing module membership. Moreover, the solution is constructive, as it provides a concrete formula for a controller (if one exists).

REFERENCES

[1] T. Becker and V. Weispfenning, *Gröbner Bases, A Computational Approach to Commutative Algebra*, Springer-Verlag, New York, 1993.

[2] M. Bisiacco and M. E. Valcher, *Direct sum decompositions of two-dimensional behaviors*, in Proceedings of the 14th International Symposium of Mathematical Theory of Networks and Systems, Perpignan, France, 2000.

[3] J. W. Brewer, *Kronecker products and matrix calculus in system theory*, IEEE Trans. Circuits Systems, 25 (1978), pp. 772–781.

[4] B. Buchberger, *Gröbner bases: An algorithmic method in polynomial ideal theory*, in Multidimensional Systems Theory, N. K. Bose, ed., D. Reidel, Boston, Dordrecht, 1985, pp. 184–232.

[5] D. Eisenbud, *Commutative Algebra with a View Toward Algebraic Geometry*, Springer-Verlag, New York, 1995.

[6] E. Fornasini, P. Rocha, and S. Zampieri, *State space realization of* 2*-D finite-dimensional behaviours*, SIAM J. Control Optim., 31 (1993), pp. 1502–1517.

[7] G. M. Greuel, G. Pfister, and H. Schönemann, *Singular version* 1.2 *user manual*, in Reports on Computer Algebra 21, Centre for Computer Algebra, University of Kaiserslautern, 1998; also available online from http://www.mathematik.uni-kl.de/∼zca/Singular.

[8] V. Lomadze and E. Zerz, *Control and interconnection revisited: The linear multidimensional case*, in Proceedings of the Second International Workshop on Multidimensional (ND) Systems, Czocha Castle, Poland, Technical University Press, Zielona Gora, Poland, 2000.

[9] U. Oberst, *Multidimensional constant linear systems*, Acta Appl. Math., 20 (1990), pp. 1–175.

[10] P. Rocha and J. Wood, *A foundation for the control theory of nD behaviours*, in Proceedings of the 13th International Symposium of Mathematical Theory of Networks and Systems, Padova, Italy, 1998.

[11] P. Rocha and J. Wood, *Trajectory control and interconnection of* 1*D and nD systems*, SIAM J. Control Optim., 40 (2001), pp. 107–134.

[12] S. Shankar, *The lattice structure of behaviors*, SIAM J. Control Optim., 39 (2001), pp. 1817–1832.

[13] M. E. Valcher, *On the decomposition of two-dimensional behaviors*, Multidimens. Systems Signal Process., 11 (2000), pp. 49–65.

[14] J. C. Willems, *Paradigms and puzzles in the theory of dynamical systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 259–294.

[15] J. C. Willems, *On interconnections, control, and feedback*, IEEE Trans. Automat. Control, 42 (1997), pp. 326–339.

[16] D. C. Youla and P. F. Pickel, *The Quillen–Suslin theorem and the structure of n-dimensional elementary polynomial matrices*, IEEE Trans. Circuits Systems, 31 (1984), pp. 513–518.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 4, pp. 1087–1106

ON OPEN- AND CLOSED-LOOP BANG-BANG CONTROL IN NONZERO-SUM DIFFERENTIAL GAMES*

GEERT JAN OLSDER†

Abstract. The Nash equilibria of two two-person nonzero-sum differential games with hard constraints on the controls are studied. For both games the open-loop as well as the closed-loop solutions, and their relationships, are discussed. As is well-known for "smooth" nonzero-sum games, these solutions are generally different. Because of the constraints, the optimal controls are of the bang-bang type, and the solutions of the two problems under consideration are nonsmooth. One deals with non-Lipschitzian differential equations (considering the problem as an optimal control problem for one player while the bang-bang feedback control of the other player is assumed to be fixed), and the corresponding value functions possess singular surfaces. General conditions for the existence and uniqueness of the feedback solutions in this framework are not yet known. It is shown that in the two examples the open-loop and closed-loop solutions differ. As a by-product, the paper aims at a modest exploration of singular surfaces in nonzero-sum games.

Key words. bang-bang control, nonzero-sum game, differential game, feedback, singular surface, switching surface, value function, open-loop control, Nash equilibrium

AMS subject classifications. 90D25,90D10,49K30,49N35

PII. S0363012900373252

1. Introduction.

This paper deals with open-loop as well as closed-loop (equivalently, feedback) bang-bang control in nonzero-sum games. The author knows of only one other paper in this direction, viz., [14]. The goal originally set was to extend the definition of viscosity solutions (see [1],[2]) to nonzero-sum differential games in which discontinuities in the solutions appear. For that purpose some simple nonzero-sum differential games with bang-bang solutions have been formulated in the hope that open-loop as well as feedback solutions can be obtained that can be intuitively understood. With such solutions in mind one might then get a feeling for the "suitability" of possible definitions of viscosity solutions in this context. With respect to the equilibrium concept, only Nash solutions are considered.

As it soon turned out, this goal was set too high. What remains is the presentation of two examples for which the optimal open-loop as well as the optimal feedback solutions are given. The latter solution is surrounded by some question marks, however, for both examples. The optimal open-loop solutions by themselves have some interesting singular curves and other features. These examples may have an interest by themselves and can function as benchmark problems for further studies. What has been shown is that the synthesis of the optimal open-loop solutions does not lead to the optimal feedback solutions. Thus the optimal feedback solutions, generally hard to obtain and with no results on uniqueness, are fundamentally different from the optimal open-loop ones.

The models that we will consider can formally be described by

$$\dot{x} = f(x, u_1, u_2)$$

*Received by the editors June 2, 2000; accepted for publication (in revised form) June 6, 2001; published electronically November 28, 2001.
http://www.siam.org/journals/sicon/40-4/37325.html
†Faculty of Information Technology and Systems, Delft University of Technology, P.O. Box 5031, 2600 GA Delft, the Netherlands (g.j.olsder@its.tudelft.nl).

with a given initial condition, and where the state $x \in \mathrm{R}^n$ and the controls $u_i \in \mathrm{R}^{m_i}$. The dot on x refers to the time derivative; $\dot{x} = \frac{\mathrm{d}x}{\mathrm{d}t}$, where time itself is (as usual) indicated by the variable t. Decision maker i chooses $u_i(\cdot)$, $i = 1, 2$. Sometimes one has restrictions on the controls of the form $u_i \in U_i \subset \mathrm{R}^{m_i}$; U_i is called the admissible region for u_i. This model will be considered on an interval $0 \leq t \leq T$, where T is defined as

$$T = \inf\{t | l(t, x(t)) = 0\}$$

for a given scalar function l. Thus the final time T is not necessarily fixed, but can depend on the controls chosen. The decision makers have different cost functions, which they try to minimize: $\min_{u_i} J_i(u_1, u_2)$, $i = 1, 2$, where

$$J_i(u_1, u_2) = \left(\int_0^T g_i(x, u_1, u_2)\mathrm{d}t + q_i(x(T)) \right).$$

(If one would add a minus sign to the cost functions, the minimization becomes a maximization; therefore there is no essential difference in whether the players are maximizing or minimizing.) The problem just formulated belongs to the realm of game theory. To emphasize the time aspect in these problems one also talks about the theory of differential, or dynamic, games. If it happens that $g_1 = -g_2$ and $q_1 = -q_2$, then one talks about a zero-sum game. The two problems to be discussed in this paper are both *nonzero-sum*. In the literature (as well as in this paper) the notation varies somewhat. Instead of the controls u_1 and u_2 one also sees the notation u and v; this is the case in problem statement 2 of this paper, for instance.

The problem statement is not yet complete. One needs an equilibrium concept. The one adopted here is the Nash equilibrium (u_1^*, u_2^*), which satisfies, by definition [3],

(1.1) $J_1(u_1^*, u_2^*) \leq J_1(u_1, u_2^*)$, $\quad J_2(u_1^*, u_2^*) \leq J_2(u_1^*, u_2)$ $\quad \forall$ admissible u_1 and u_2.

For an additional remark on "admissible controls," see Remark 1.1 below.

The last item needed to make the problem statement complete is the information on which the players base their decisions. Both players know the model, the initial condition, their own and their opponent's cost functions, and the equilibrium concept according to which the game will be played. If the control functions are furthermore based only on time, written as $u_i(t)$, then we talk about open-loop solutions. If the control functions can (also) depend on the current state x (i.e., at time t), then we write $u_i(t, x)$ and such controls are called feedback or closed-loop controls, or simply strategies. In the case of closed-loop control, at time t the player has the state $x(t)$ at his disposition; in the case of open-loop control, the player knows only the time t upon which he must base his decision. Apart from robustness considerations, it is also necessary to make the distinction between open-loop and closed-loop controls, since the corresponding optimal solutions will generally be different; this was first observed in [15], [16] (see further [3]).

Please note the following. The open-loop solution depends on the initial condition, although this is not explicitly shown in the conventional notation $u_i(t)$. (We could have written $u_i(t, x(0))$.) To be precise, and strictly speaking in contradiction with the foregoing, the feedback solution $u_i(t, x)$ will, by definition, *not* depend on the initial state. If we allow such a dependence, which could be written as $u_i(t, x, x(0))$, a plethora of informationally nonunique Nash equilibria arises; see [3], section 6.3.2.

If the functions f, l, and g_i are independent of t, then the feedback controls (as well as the value functions to be introduced later in this section) will also be independent of t, i.e., we can write $u_i(x)$.

The usual necessary and/or sufficient conditions for the optimal controls to satisfy are either given in terms of the so-called maximum principle of Pontryagin or in terms of the Hamilton–Jacobi–Bellman (HJB) theory, the latter of which essentially is a mathematical consequence of the principle of dynamic programming. Well-known references are [4] and [5]. We will give these conditions in some detail for optimal control problems (i.e., problems with only one player). The reason why it is useful to consider optimal control theory is that in a game with the Nash equilibrium concept, each player tries to solve an optimal control problem, keeping the control or strategy of the other player fixed. For the application of the maximum principle, one defines (for a single player, with obvious notation) the Hamiltonian

$$H = \lambda(t) f(x,u) + g(x,u),$$

where $\lambda(t) \in \mathrm{R}^n$, and where λf is the innerproduct of the vectors λ and f. The costate function $\lambda(t)$ satisfies $\dot{\lambda} = -\frac{\partial H}{\partial x}$ with boundary condition $\lambda(T) = \frac{\mathrm{d}q(x(T))}{\mathrm{d}x}$ along $l(t,x) = 0$. The optimal u satisfies $u^* = \arg\max_u H$. With this approach one obtains u^* as a function of t, i.e., as an open-loop control. In contrast, application of the HJB theory yields the optimal u as a function of the current time and the current state: $u^*(t,x)$. It is obtained by introduction of the so-called value function V, which is, by definition,

$$V(t,x) = \min_u \left(\int_t^T g(x,u)\mathrm{d}t + q(x(T)) \right),$$

with the assumption that at time t the state is x (these t and x are the arguments of V). It satisfies the HJB PDE

$$-\frac{\partial V}{\partial t} = \min_u \left(\frac{\partial V}{\partial x} f + g \right),$$

with boundary condition $V(T,x) = q(x(T))$ along $l(t,x) = 0$.

REMARK 1.1. *The conditions on f, l, g, and q under which the maximum principle and/or the HJB theory are applicable have not been formulated explicitly; the reader is referred to the literature. In addition, one must also carefully define the class of admissible u_1 and u_2 in* (1.1). *The formalization is involved, and for the so-called regularity conditions, especially for feedback controls, the reader is referred to* [10].

REMARK 1.2. *The difference in these two approaches (optimal open-loop solutions with conditions based on the Pontryagin maximum principle and optimal closed-loop solutions with conditions based on the HJB equation) also shows up if one tries to apply Pontryagin when all players would play closed-loop strategies. For the games to be discussed in the paper, the Hamiltonian and the costate vector corresponding to the first player are indicated by H_1, respectively λ. For the second player these quantities are, respectively, H_2 and μ. Expressed in these quantities, the difference in the conditions given for the closed-loop case is that the differential equations for the costate vectors must be adapted to $\dot{\lambda} = -\frac{\partial H_1}{\partial x} - \frac{\partial H_1}{\partial u_2}\frac{\partial u_2}{\partial x}$ and $\dot{\mu} = -\frac{\partial H_2}{\partial x} - \frac{\partial H_2}{\partial u_1}\frac{\partial u_1}{\partial x}$.*

For the so-called linear quadratic games (see [3]*) with no restrictions on the controls, more explicit results exist in terms of Riccati differential equations; the form of these equations depends on whether open-loop or closed-loop solutions are considered.*

As formulated above, the HJB equation is a necessary condition for the optimal control and value function to satisfy. A sufficiency condition in terms of the HJB equation is known under the name of verification theorem. For that purpose an essential condition is that $V(t,x)$ must be twice differentiable in x. Solutions of many optimal control problems do not have such smooth value functions; in particular, when bang-bang controls are involved, this is usually not the case. One speaks of bang-bang control if the optimal control jumps from one boundary point of the admissible region to another. In such cases, V is often only piecewise continuous or piecewise continuously differentiable. The boundaries between areas in the (t,x) space where V is smooth are called singular surfaces or singular lines. Especially in zero-sum differential game theory such singular surfaces are a well-studied subject [7],[11]. See also [12] for some in-depth discussions on singular surfaces. In nonzero-sum differential games such singular surfaces are largely unexplored. As a by-product, this paper yields one of the first (see also [14]), though modest, explorations in this direction.

It is worth mentioning that other definitions of closed-loop equilibrium solutions exist. One such equilibrium is based on the threat or punishment principle that one often encounters in the theory of repeated games. Briefly, if one player decides to deviate from his optimal strategy within the context of this equilibrium, then the other player immediately plays worst-case against this deviating player. In other words, the threat is that the other player will try to maximize the cost function of the deviating player (who himself wants to minimize this function), thereby totally disregarding his own cost function. More information on this equilibrium concept, which does not satisfy the dynamic programming principle, can be found in [8], [9].

2. Problem statement 1. The problem studied is a nonzero-sum differential game with two players. The model is

$$\dot{x} = (1-x)u_1 - xu_2, \tag{2.1}$$

and the criteria are

$$\max_{u_1} \int_0^T (c_1 x - u_1)\mathrm{d}t, \quad \max_{u_2} \int_0^T (c_2(1-x) - u_2)\mathrm{d}t. \tag{2.2}$$

The state x as well as the controls u_i are one-dimensional quantities. The choice of the controls is subject to

$$0 \le u_i(t) \le 1, \qquad i = 1,2. \tag{2.3}$$

The quantities c_i are positive constants. The final time T is supposed to be fixed. This model and both criteria are known to both players (they have full information). We are interested in the Nash equilibrium solutions of this game. In the next section such open-loop solutions will be studied, and in section 4 we get to the heart of this paper, the study of feedback solutions.

This model was mentioned in [13]. A possible interpretation is as follows. Players P1 and P2 are firms on the same market and produce the same product. The number of customers allotted to P1 at time t is $x(t)$; the number of customers allotted to P2 at time t are those remaining, i.e., $1-x(t)$. The total number of customers is assumed to be constant and has been normalized to 1. The controls u_i refer to advertising intensities (amount of money spent on advertising per unit of time). If P1 advertises, then his number of customers will increase proportionally to the advertising rate and the number of customers not already allotted to him (the term $(1-x)u_1$ in the

model). On the other hand, advertising by P2 will decrease the number of customers of P1, also proportionally to the advertising rate and the number of customers not already allotted to P2 (the term $-xu_2$ in the model). The criteria represent the profits made by each player. This profit, per unit of time, is proportional to the number of customers allotted to the player concerned (with a proportionality factor c_i) minus the amount of money spent on advertising.

3. Optimal trajectories—The open-loop case. Consider the two Hamiltonians

$$H_1 = \lambda_1(u_1(1-x) - u_2x) + c_1x - u_1, \tag{3.1}$$

$$H_2 = \mu_1(u_1(1-x) - u_2x) + c_2(1-x) - u_2. \tag{3.2}$$

The costate variables λ_i satisfy

$$\dot{\lambda}_1 = -\frac{\partial H_1}{\partial x} = -(-\lambda_1u_1 - \lambda_1u_2) - c_1, \qquad \lambda_1(T) = 0, \tag{3.3}$$

$$\dot{\mu}_1 = -\frac{\partial H_2}{\partial x} = -(-\mu_1u_1 - \mu_1u_2) + c_2, \qquad \mu_1(T) = 0. \tag{3.4}$$

Notice that

$$\mu_1 \equiv -\frac{c_2\lambda_1}{c_1}. \tag{3.5}$$

The optimal controls must satisfy

$$u_1 = \text{Heav}(\lambda_1(1-x) - 1), \tag{3.6}$$

$$u_2 = \text{Heav}(-\mu_1x - 1), \tag{3.7}$$

where Heav stands for the Heaviside function, i.e.,

$$\text{Heav}(x) = \begin{cases} 0 \text{ if } x < 0, \\ \text{undetermined if } x = 0, \\ 1 \text{ if } x > 0. \end{cases}$$

The switching function for u_1 is

$$\lambda_1(t)(1 - x(t)) - 1.$$

Close to T we have $u_i = 0,\ i = 1, 2$, and the switching curve, to be denoted by l_1, in the (t, x)-plane is

$$\lambda_1(t)(1 - x(t)) - 1 = -c_1(t - T)(1 - x) - 1 = 0.$$

We have to check whether this switching curve is a real one (i.e., instead of a real switch one might get shattering or a singular arc). It easily follows that the switching function indeed will change sign (hence shattering is not possible, provided u_2 remains zero). Moreover, it follows from some elementary though tedious calculations not shown here that neither control will have another switch (i.e., in retrograde time u_1 switches from 0 to 1 and u_2 does not switch at all), provided that $c_1 < 4$.

A similar reasoning holds with respect to the switching curve (l_2) for u_2, which satisfies

$$-\mu_1(t)x(t) - 1 = -c_2(t - T)x - 1 = 0.$$

These two curves intersect at

$$x_{\rm s} = \frac{c_1}{c_1+c_2}, \quad t_{\rm s} = T - \frac{c_1+c_2}{c_1 c_2}.$$

Suppose that this point lies in the area of interest, i.e., $0 < x_{\rm s} < 1$ and $0 < t_s < T$. A singular arc ($l_{\rm s}$) exists to left of the point $t_{\rm s}, x_{\rm s}$, along which

$$u_1 = \frac{c_1^2 c_2}{(c_1+c_2)^2}, \quad u_2 = \frac{c_1 c_2^2}{(c_1+c_2)^2}, \tag{3.8}$$

provided that these values lie in the interval $[0,1]$, which is satisfied for $0 \le c_i \le 4$. This singular arc has been obtained by studying

$$\lambda_1(1-x) - 1 \equiv 0, \quad -\mu_1 x - 1 \equiv 0,$$

and calculating their time derivatives:

$$\lambda_1 u_2 - c_1(1-x) \equiv 0, \quad c_2 x + \mu_1 u_1 \equiv 0.$$

Together with (3.5) it is then easily shown that the singular arc given above is the only one possible. Along the same lines it can be shown that a singular arc with respect to only one control (its switching function being zero and the switching function of the other control being nonzero) does not exist.

4. Optimal trajectories—The feedback case.

4.1. By means of HJB.

4.1.1. Synthesis approach. The meaning of "synthesis" is the following. Since the whole (t,x) space (at least the relevant part) is covered with open-loop trajectories (see Figure 4.1) to each point (t,x) there corresponds a unique trajectory (in forward time direction) with corresponding local u_i values. If these values are written as $u_i(t,x)$, we speak of synthesis. The central question in this subsection will be whether these feedback functions constitute the *optimal* feedback strategies. The answer will turn out to be no. The space R^2, with one axis being the t-axis and the other one the x-axis, will be denoted by Ω.

The two HJB equations are

$$\begin{aligned} -\frac{\partial V_1}{\partial t} &= \max_{u_1}\left(\frac{\partial V_1}{\partial x}((1-x)u_1 - x u_2) + c_1 x - u_1\right) \\ &= \left[\frac{\partial V_1}{\partial x}(1-x) - 1\right]^+ - x\frac{\partial V_1}{\partial x}\mathrm{Heav}\left(-\frac{\partial V_2}{\partial x}x - 1\right) + c_1 x, \end{aligned} \tag{4.1}$$

$$\begin{aligned} -\frac{\partial V_2}{\partial t} &= \max_{u_2}\left(\frac{\partial V_2}{\partial x}((1-x)u_1 - x u_2) + c_2(1-x) - u_2\right) \\ &= \left[-x\frac{\partial V_2}{\partial x} - 1\right]^+ + \frac{\partial V_2}{\partial x}(1-x)\mathrm{Heav}\left(\frac{\partial V_1}{\partial x}(1-x) - 1\right) \\ &\quad + c_2(1-x), \end{aligned} \tag{4.2}$$

where $[a]^+ = 0$ if $a \le 0$, and $[a]^+ = a$ if $a > 0$. Formally $\frac{\partial V_i}{\partial x}$, $i = 1,2$, satisfy

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial V_1}{\partial x}\right) = \frac{\partial V_1}{\partial x}(u_1+u_2) - c_1 + \frac{\partial V_1}{\partial x}x\frac{\partial u_2}{\partial x}, \quad \frac{\partial V_1(T,x)}{\partial x} = 0, \tag{4.3}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial V_2}{\partial x}\right) = \frac{\partial V_2}{\partial x}(u_1+u_2) + c_2 - \frac{\partial V_2}{\partial x}(1-x)\frac{\partial u_1}{\partial x}, \quad \frac{\partial V_2(T,x)}{\partial x} = 0, \tag{4.4}$$

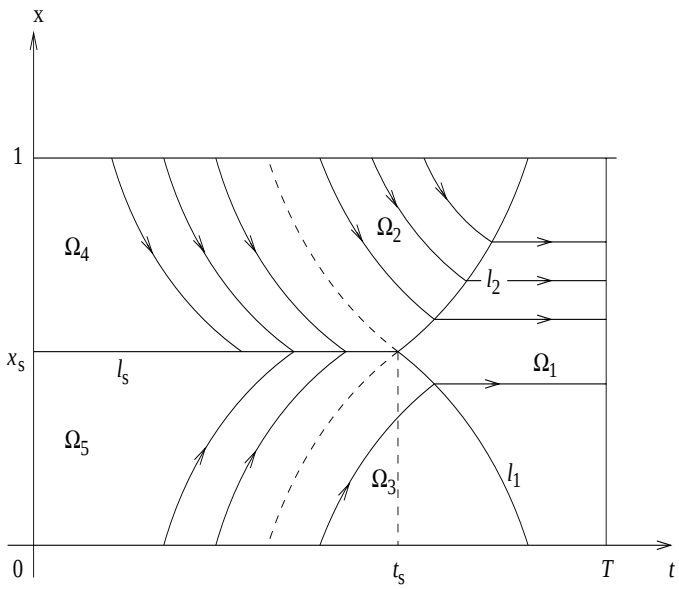


FIG. 4.1. *The open-loop trajectories in the (state, time) plane.*

where

$$u_2 = \text{Heav}\left(-\frac{\partial V_2}{\partial x}x - 1\right), \tag{4.5}$$

$$u_1 = \text{Heav}\left(\frac{\partial V_1}{\partial x}(1-x) - 1\right). \tag{4.6}$$

Above we wrote "formally," since the partial derivatives $\frac{\partial u_i}{\partial x}$ do not necessarily exist.

Since the open-loop solutions cover the whole Ω space, we can calculate the cost to go from each point in Ω. If we obtain identities by substituting these costs into (4.1) and (4.2), then the conclusion would be that the optimal open-loop solutions also form a set of optimal feedback solutions (tacitly assuming that an extension of the verification theorem, without the differentiability assumptions, will hold for this situation). For this purpose, Ω will be split up into five subregions as indicated in Figure 4.1; Ω_4, for instance, consists of those starting points for which first $u_1 = 0$, $u_2 = 1$, then a singular part and, during the last part, both controls are zero. For a point $(\tilde{t}, \tilde{x}) \in \Omega_4$ it is a straightforward calculation to show that

(4.7)
$$V_1(\tilde{t}, \tilde{x}) = c_1\left(\tilde{x} - \frac{c_1}{c_1 + c_2}\right) + \frac{c_1^3}{(c_1 + c_2)^2}(T - t^*) + \frac{c_1}{c_1 + c_2},$$

(4.8)
$$V_2(\tilde{t}, \tilde{x}) = (c_2 - 1)\ln\left(\tilde{x}\frac{c_1 + c_2}{c_1}\right) + c_2\left(-\tilde{x} + \frac{c_1}{c_1 + c_2}\right) + \frac{c_2^3}{(c_1 + c_2)^2}(T - t^*) + \frac{c_2}{c_1 + c_2}.$$

where $t^* = \tilde{t} + \ln(\tilde{x}\frac{c_1+c_2}{c_1})$. To simplify the calculations somewhat, suppose $c_1 = c_2 = 2$. Then

$$V_1(\tilde{t},\tilde{x}) = 2\tilde{x} - \frac{1}{2}\ln(2\tilde{x}) + \frac{1}{2}(T-\tilde{t}) - \frac{1}{2}, \tag{4.9}$$

$$V_2(\tilde{t},\tilde{x}) = -2\tilde{x} + \frac{1}{2}\ln(2\tilde{x}) + \frac{1}{2}(T-\tilde{t}) + \frac{3}{2}. \tag{4.10}$$

These equations are now substituted into (4.1), together with $u_2 = 1$ (which is the open-loop u_2-solution in Ω_4), and into (4.2), together with $u_1 = 0$ (which is the open-loop u_1-solution in Ω_4), resulting in

$$\begin{aligned} \frac{1}{2} &= \max_{u_1}\left(\left(2 - \frac{1}{2\tilde{x}}\right)((1-\tilde{x})u_1 - \tilde{x}) + 2\tilde{x} - u_1\right) \\ &= \max_{u_1}\left(u_1\left(\frac{3}{2} - 2\tilde{x} - \frac{1}{2\tilde{x}}\right)\right) + \frac{1}{2}, \end{aligned} \tag{4.11}$$

$$\begin{aligned} \frac{1}{2} &= \max_{u_2}\left(\left(-2 + \frac{1}{2\tilde{x}}\right)(-\tilde{x}u_2) + 2(1-\tilde{x}) - u_2\right) \\ &= \max_{u_2}\left(u_2\left(2\tilde{x} - \frac{3}{2}\right)\right) + 2(1-\tilde{x}). \end{aligned} \tag{4.12}$$

Equation (4.11) is an identity; (4.12), however, is not. Equation (4.12) is an identity only if $\tilde{x} > \frac{3}{4}$ (with $u_2 = 1$). For $\frac{1}{2} < \tilde{x} < \frac{3}{4}$ it yields $u_2 = 0$, however, which does not lead to an identity and which does not coincide with the open-loop solution. Thus we have shown that the optimal open-loop solution does not constitute an optimal feedback solution.

4.1.2. Optimal feedback solution. If one would change the control values along the singular arc from the values given in (3.8) to $u_1 = u_2 = 0$ (educated guess), then the optimal trajectories obtained do not change in the (x,t)-space. Calculation of the value functions with this change leads to (taking again $c_1 = c_2 = 2$ and a point $(\tilde{x},\tilde{t}) \in \Omega_4$)

$$V_1(\tilde{t},\tilde{x}) = T - 1 - \tilde{t} + 2\tilde{x} - \ln(2\tilde{x}), \tag{4.13}$$

$$V_2(\tilde{t},\tilde{x}) = T + 1 - \tilde{t} - 2\tilde{x}. \tag{4.14}$$

If one does the same exercise again as in the previous subsection, i.e., substitution of these value functions into (4.1) and (4.2), then it turns out that two identities result. Hence one may conclude that a set of optimal Nash feedback strategies has been found. Nothing is known about its uniqueness.

5. Other approaches. In this section three other, as yet not very fruitful, approaches are briefly indicated.

5.1. Brute force numerical solutions. Numerical integration of (4.1) and (4.2) by means of simple integration schemes did not seem to lead to any kind of convergence of the numerical outcome if the stepsizes were made smaller and smaller.

5.2. By means of the maximum principle. In this subsection a method is given for integrating (4.3) and (4.4) backward in time. The functions u_1 and u_2 are determined by the right-hand sides of (4.1) and (4.2). The crucial step is that δ-functions (the derivatives of u_i with respect to x) appear and that they will be

replaced by rectangles of width ϵ and height $\frac{1}{\epsilon}$. After the calculations have been made including this ϵ, we take $\lim_{\epsilon \downarrow 0}$.

Consider (4.4), with $u_1 = u_2 = 0$, assuming that we integrate backwards starting from $t = T$. Once we hit l_1 at $t = t_{1s}$, the right-hand side contains $\frac{\partial u_1}{\partial x}$, which will be interpreted as a negative δ-function (along a line $t =$ constant, u_1 first equals 1 and subsequently 0 for increasing x). Thus we obtain

$$\frac{\partial V_2}{\partial x}(t_{1s}-) = e^{-(1-x(t_{1s}))}\frac{\partial V_2}{\partial x}(t_{1s}+).$$

Thus $\frac{\partial V_2}{\partial x}$ has a jump at $t = t_{1s}$. Technically we can handle this jump, but is it not known what its real meaning for the solution of the problem might be. Imagining a horizontal singular arc in the feedback case, just as l_{s}, then integrating along this l_{s}, one is faced with an ongoing δ-function. This does not make sense at all. It is very possible that other kinds of singular surfaces will play a role. Such surfaces are well-known in zero-sum pursuit evasion games; see [7] and [11].

5.3. The unpaved road of viscosity solutions. The standard definitions of viscosity solutions do not apply here due to the discontinuous right-hand sides of (4.1) and (4.2). In order to smooth the discontinuities, one can add noise to the system equations (and then let the noise intensity go to zero). This approach does not seem attractive from an analytic point of view.

6. Problem statement 2. Consider the model

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u + \frac{1}{2}\begin{pmatrix} 0 \\ 1 \end{pmatrix} v. \tag{6.1}$$

The constraints on the scalar controls are $|u| \leq 1$ and $|v| \leq 1$. The final time T is defined as the first instant at which $x_1 = x_2$. Starting from an arbitrary initial condition, the u-player wants to minimize T and the v-player wants to maximize $x_1(T)$. Thus a nonzero-sum differential game has been formulated. This game is time-invariant in the sense that calendar time does not enter into the problem statement. (This in contrast to the previous problem, where the calendar time is present in terms of the fixed final time T.) Hence the value functions will be functions of the state only (and not also of time, as was the case in the previous problem). No physical or economical meaning is envisaged here. It is a generalization of the optimal control problem in Example 5.2 of [3]. We will concentrate on the solution in the south-east of the state space, defined by $x_1 > x_2$. The solution in the north-west is a point-mirrored (through the origin) copy of the one to be obtained in the south-east.

7. Analysis. If the costate vectors are given by $\lambda = (\lambda_1, \lambda_2)$ and $\mu = (\mu_1, \mu_2)$ for the u-, respectively v-, player, the Hamiltonians are

$$H_1 = 1 + \lambda_1 x_2 + \lambda_2\left(u + \frac{1}{2}v\right), \tag{7.1}$$

$$H_2 = \mu_1 x_2 + \mu_2\left(u + \frac{1}{2}v\right). \tag{7.2}$$

The optimal controls satisfy

$$u^* = -\mathrm{sgn}(\lambda_2), \quad v^* = \mathrm{sgn}(\mu_2).$$

In the open-loop case, the differential equations for the costate variables are

$$\dot{\lambda}_1 = 0, \quad \dot{\lambda}_2 = -\lambda_1, \quad \dot{\mu}_1 = 0, \quad \dot{\mu}_2 = -\mu_1.$$

In the feedback case, the differential equations for the costate variables are (provided they make sense)

$$\dot{\lambda}_1 = -\frac{1}{2}\lambda_2 \frac{\partial v}{\partial x_1}, \quad \dot{\lambda}_2 = -\lambda_1 - \frac{1}{2}\lambda_2 \frac{\partial v}{\partial x_2}, \tag{7.3}$$

$$\dot{\mu}_1 = -\mu_2 \frac{\partial u}{\partial x_1}, \quad \dot{\mu}_2 = -\mu_1 - \mu_2 \frac{\partial u}{\partial x_2}. \tag{7.4}$$

Let the final state $x(T)$ be parametrized by $x_1 = x_2 = a$. Termination can be enforced by the u-player as long as $a \leq \frac{1}{2}$. With a "cooperative" v-player, termination can occur for all $a \leq \frac{3}{2}$. Later on we will see that the v-player will cooperate in this sense of terminating the game—it is in his own interest, and thus it is not pure cooperation.

Suppose $a < 0$. Then close to the end we will have $u^* = v^* = 1$ (obvious, but which also can be seen as the outcome of a trivial static nonzero-sum game, assuming that u and v are constant during the last part of the game), which leads to

$$\lambda_1(T) = \frac{\partial V_1}{\partial x_1} = \frac{1}{\frac{3}{2} - a}, \quad \lambda_2(T) = \frac{\partial V_1}{\partial x_2} = \frac{-1}{\frac{3}{2} - a},$$

$$\mu_1(T) = \frac{\partial V_2}{\partial x_1} = \frac{1}{1 - \frac{2}{3}a}, \quad \mu_2(T) = \frac{\partial V_2}{\partial x_2} = \frac{-\frac{2}{3}a}{1 - \frac{2}{3}a}.$$

As an example of how these terminal conditions have been calculated, see Figure 7.1. Suppose that T_1 is the terminal point and that a perturbation Δx_1 occurs such that the state is temporarily in point B. From there onward, the game will end in the new endpoint T. In this state space figure, the velocity vector $\dot{x}_1 = x_2$, $\dot{x}_2 = \frac{3}{2}$ has been superimposed (the origin of this vector being situated at point B). Now we get

$$\tan\beta = \frac{-x_2}{\frac{3}{2}} = \frac{\overline{\mathrm{TC}}}{\overline{\mathrm{TA}}} = \frac{\Delta x_1 - \overline{\mathrm{T_1A}}}{\overline{\mathrm{T_1A}}},$$

which leads to

$$\overline{\mathrm{T_1A}} = \overline{\mathrm{TA}} = \frac{\frac{3}{2}\Delta x_1}{-x_2 + \frac{3}{2}}.$$

Substitution of these expressions into

$$\frac{\Delta V_1}{\Delta x_1} = \frac{\text{time needed to go from B to T}}{\Delta x_1} = \frac{\frac{2}{3}\overline{\mathrm{TA}}}{\Delta x_1}, \quad \frac{\Delta V_2}{\Delta x_1} = \frac{\overline{\mathrm{T_1A}}}{\Delta x_1},$$

leads immediately to the expressions for $\lambda_1(T)$ and $\mu_1(T)$ given above. A bar over a line segment denotes its length in these formulas. One can use the same figure for deriving the expressions for $\lambda_2(T)$ and $\mu_2(T)$ if one starts with T_2 as the original unperturbed termination point.

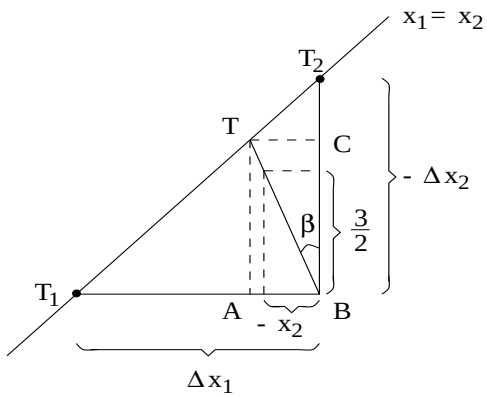


FIG. 7.1. *Perturbations at termination point; all quantities related to length are positively valued.*

Now solving for the costate variables, the switching functions become

$$\lambda_2(t) = \frac{-1}{\frac{3}{2} - a}(t - T + 1), \quad \mu_2(t) = \frac{-1}{1 - \frac{2}{3}a}\left(t - T + \frac{2}{3}a\right).$$

The control v will not have a switch. The control u will have a switch for $t = T - 1$.

Suppose now that $0 < a < \frac{1}{2}$. Then close to the end we have $u^* = 1,\ v^* = -1$ and, performing an analysis similar to that given above, we obtain

$$\lambda_1(T) = \frac{2}{1 - 2a}, \quad \lambda_2(T) = \frac{-2}{1 - 2a},$$
$$\mu_1(T) = \frac{1}{1 - 2a}, \quad \mu_2(T) = \frac{-2a}{1 - 2a}.$$

The switching functions now become

$$\lambda_2(t) = -\frac{2}{1 - 2a}(t - T + 1), \quad \mu_2(t) = -\frac{t - T + 2a}{1 - 2a}.$$

The control u will have a switch for $t = T - 1$, and the control v will have a switch for $t = T - 2a$.

8. Optimal trajectories—The open-loop case. Three cases will be distinguished: $a < 0$, $0 < a < \frac{1}{2}$, and $\frac{1}{2} < a < \frac{3}{2}$.

8.1. The case $a < 0$. The optimal trajectories during $T - 1 \le t \le T$ are, with $u^* = v^* = 1$,

$$x_1(t) = \frac{3}{4}(t - T)^2 + a(t - T) + a, \tag{8.1}$$

$$x_2(t) = \frac{3}{2}(t - T) + a. \tag{8.2}$$

At $t = T - 1$, a switch (for u) occurs. The switching line is given by $x_1 = \frac{3}{4},\ x_2 = a - \frac{3}{2}$. In retrograde time the trajectories continue with $u^* = -1,\ v^* = 1$, leading to

$$x_1(t) = -\frac{1}{4}(t - T + 1)^2 + \left(a - \frac{3}{2}\right)(t - T + 1) + \frac{3}{4}, \tag{8.3}$$

$$x_2(t) = -\frac{1}{2}(t - T + 1) + a - \frac{3}{2}. \tag{8.4}$$

8.2. The case $0 < a < \frac{1}{2}$. Close to the line of termination we have $u^* = 1$, $v^* = -1$, and the corresponding trajectories are

$$x_1(t) = \frac{1}{4}(t-T)^2 + a(t-T) + a, \tag{8.5}$$

$$x_2(t) = \frac{1}{2}(t-T) + a. \tag{8.6}$$

A switch (for v) occurs at $t = T-2a$; the switching line is given by $x_1 = -a^2+a$, $x_2 = 0$. In retrograde time the trajectories continue with $u^* = v^* = 1$, and another switch (now for u) occurs for $t = T-1$. At $t = T-1$ we are at

$$x_1(T-1) = \frac{3}{4}(-1+2a)^2 - a^2 + a, \tag{8.7}$$

$$x_2(T-1) = \frac{3}{2}(-1+2a). \tag{8.8}$$

Elimination of a leads to the switching curve $x_1 = \frac{2}{9}x_2^2 + \frac{1}{4}$, with $-\frac{3}{2} \le x_2 \le 0$. In further retrogade time the trajectories are determined by $u = -1$, $v = 1$. The trajectories of this case and the previous one fill the whole state space (i.e., the south-eastern part). This is shown in Figure 8.1. Whether these trajectories are really the optimal ones will be discussed in the next subsection. In Figure 8.1, the curve GA is a barrier for the u-player, but not for the v-player.

8.3. The case $\frac{1}{2} < a < \frac{3}{2}$. The situation is somewhat subtle here. Of the interval $\frac{1}{2} < a \le \frac{3}{2}$, only the point $a = \frac{3}{2}$ will indicate a termination point. On this interval the u-player by himself cannot force a termination; he needs the help of player v. It is in the interest of the latter to terminate at the largest possible x_1 (remember the cost function). The optimal control v^* is not unique here. This nonuniqueness will not influence the outcome of the v-player; it will, however, influence the outcome of the u-player. In order to continue with a unique v^*, assume that the v-player goes as slowly as possible to the point $a = \frac{3}{2}$. This is the worst case from the u-player point of view. Thus $u^* = 1$, $v^* = -1$ until the trajectory hits the parabola $x_1 = \frac{1}{3}x_2^2 + \frac{3}{4}$. Once this parabola has been hit, the trajectory follows this parabola ($u^* = v^* = 1$) until the termination point has been reached. The u-player may prefer terminating the game at the point characterized by $a = \frac{3}{2}$ and hence playing $u = 1$, rather than playing $u = -1$, which will lead to a terminal point characterized by $a \le \frac{1}{2}$, which may lead to a higher T. Hence a dispersal line (the direction in which to go is to be chosen by the u-player) can be expected to exist.

In Figure 8.2 all results obtained so far have been put together; the trajectories terminating on $a = x_1 = x_2$, $\frac{1}{2} < a \le \frac{3}{2}$, form a kind of patchwork in the previous Figure 8.1. The optimal trajectories are provided with arrows (indicating the direction in forward time). Point A corresponds to $a = \frac{1}{2}$, point B to $a = \frac{3}{2}$. The curve GKCH is the switching curve for u; the line segment OG is the switching curve for v. The curve MN is the dispersal line; its location has not been calculated explicitly. On this line, the u-player determines whether to go north or south. (Remember that the u-player minimizes T.) In the cone NABDM, the v-control is nonunique; the v-player can always make sure that B will be the terminal point. The trajectories drawn refer to the slowest possible such trajectories, i.e., first $u = 1$ and $v = -1$, until the barrier

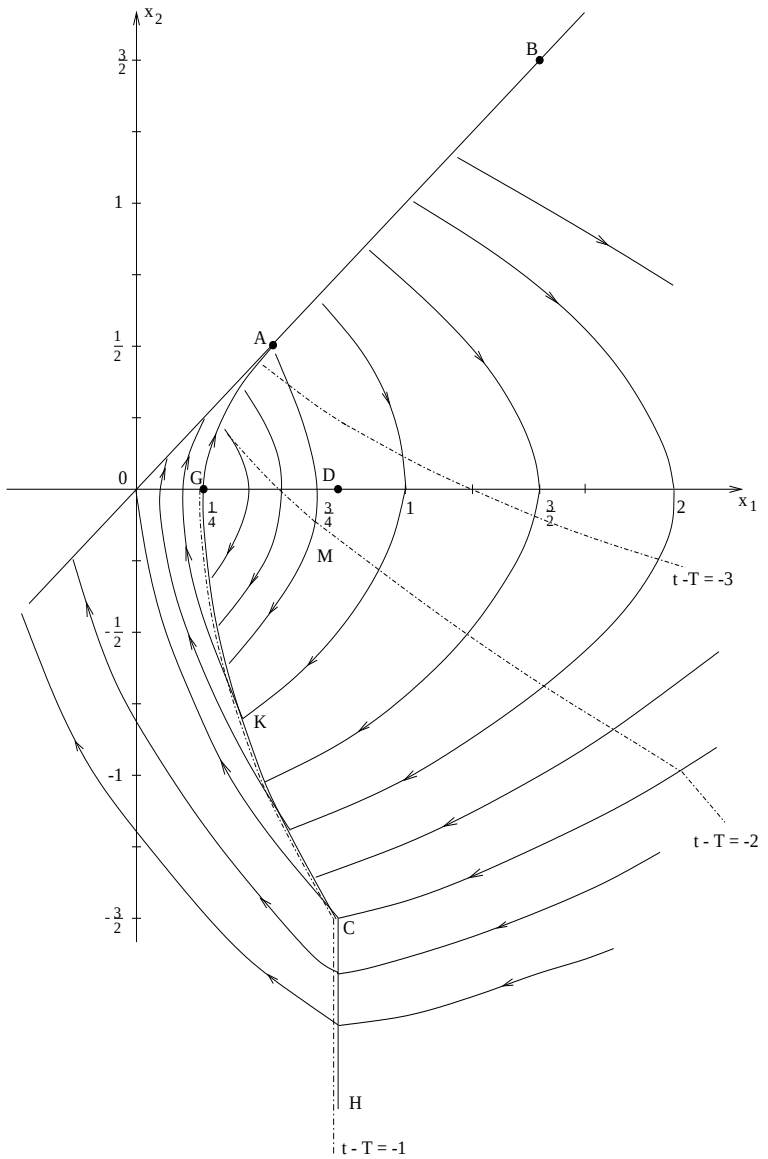


FIG. 8.1. *The open-loop trajectories in the state space, terminating with* $a < \frac{1}{2}$.

BD is hit. The curve DM is part of the parabola $x_1 = x_2^2 + \frac{3}{4}$ (with u=1, $v = -1$). Along the barrier BD, $u = 1$ and $v = 1$. One can calculate the coordinates of the point M by realizing that the times needed to go in either direction towards the line $x_1 = x_2$ must be equal. This leads to $x_2(M) = -\sqrt{9/52}$ and $x_1(M) = x_2^2(M) + \frac{3}{4}$. Both GA and DB are barriers for both players. There is a third barrier, for the v-player only; it is part of the parabola above and through point M, with formula $x_1 = -x_2^2 + c$ (corresponding to $u = -1$, $v = 1$), with c chosen such that point M is indeed part of this parabola. Some level curves of constant time to go (for the u-player) are indicated by means of dotted curves.

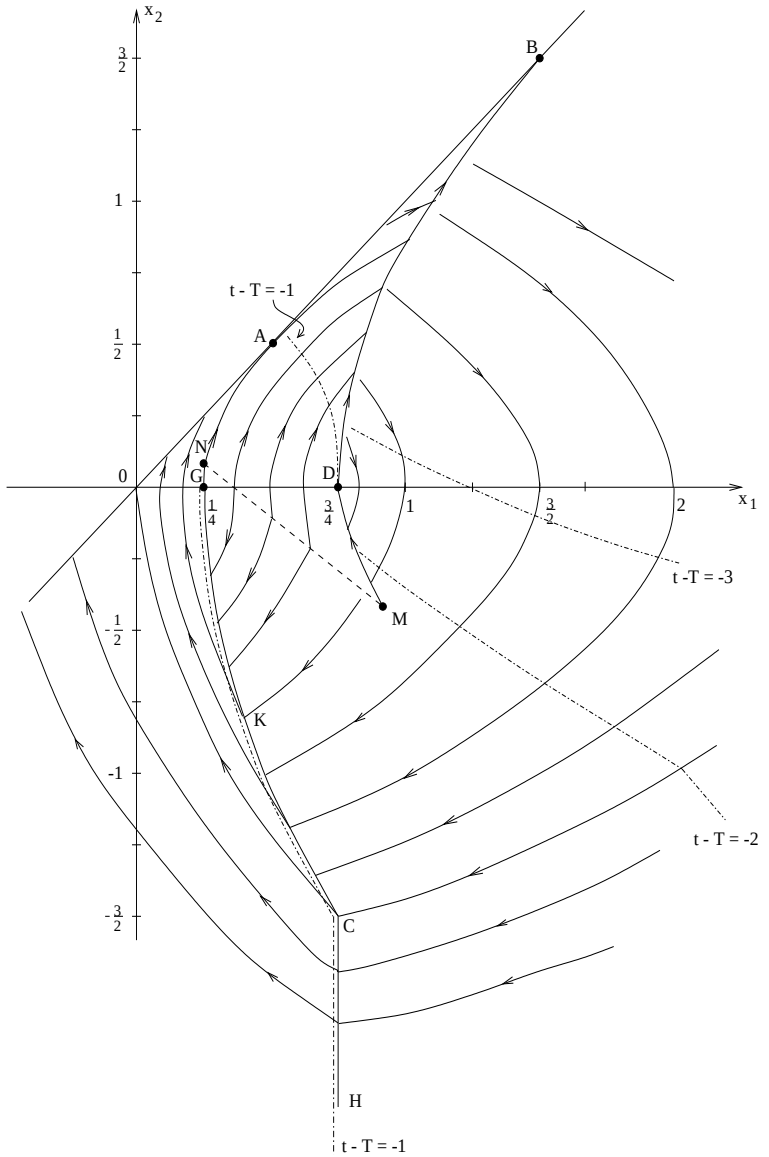


FIG. 8.2. *The optimal open-loop trajectories in the state space.*

9. Optimal trajectories—The feedback case.

9.1. By means of the maximum principle.

9.1.1. $a < 0$. At $t = T - 1$ the control u switches from $+1$ to -1 in retrograde time. Will this have an influence on μ_2, and hence on v? Towards that end consider

$$\dot{\mu}_1 = -\mu_2 \frac{\partial u}{\partial x_1}, \quad \dot{\mu}_2 = -\mu_1 - \mu_2 \frac{\partial u}{\partial x_2}.$$

At the switch we make the plausible, though mathematically not well-funded, assumption that

$$\frac{\partial u}{\partial x_1} = -2\delta\left(x_1 - \frac{3}{4}\right), \quad \frac{\partial u}{\partial x_2} = 0,$$

where δ denotes the Dirac-function. This leads to

$$\mu_2((T-1)^-) = \mu_2((T-1)^+) = 1,$$

$$\mu_1((T-1)^-) = \mu_1((T-1)^+) + 2 = \frac{1}{1-\frac{2}{3}a} + 2 = \frac{9-4a}{3-2a} > 0,$$

$$\mu_2(t) = -\frac{9-4a}{3-2a}(t-T+1) + 1 \quad \text{for } t - T < -1.$$

The notation $\mu(t^+)$ stands for $\lim_{s \downarrow t} \mu(s)$, and similarly $\mu(t^-)$ stands for $\lim_{s \uparrow t} \mu(s)$. Because $\mu_2(t) > 0$ for $t < T-1$, v will have no switch (from $+1$ to -1) for $t < T-1$.

9.1.2. $0 < a < \frac{1}{2}$. The control v switches (in retrograde time from -1 to $+1$) at $t = T - 2a$, which gives rise to the switching line OG and to a δ-function in (7.3). Again, we make a plausible, though not well-funded, assumption:

$$\frac{\partial v}{\partial x_1} = 0, \quad \frac{\partial v}{\partial x_2} = -2\delta(x_2).$$

This leads to

$$\lambda_1(t) = \frac{2}{1-2a} \quad \forall t \le T,$$

$$\lambda_2(t) = -\frac{2}{1-2a}(t-T+1), \qquad T - 2a < t \le T,$$

$$\lambda_2((T-2a)^-) = \frac{\lambda_2((T-2a)^+)}{e} = -\frac{2}{e},$$

$$\lambda_2(t) = \frac{-2}{1-2a}(t-T+2a) - \frac{2}{e} \quad \forall t < T - 2a.$$

The control u will face a switch when $\lambda_2 = 0$, i.e., for $t = T - 2a - \frac{1-2a}{e}$. It is easily shown that the corresponding curve has the equation

$$x_1 = \left(\frac{1}{3} - \frac{e^2}{9}\right)x_2^2 + \frac{1}{4}, \qquad -\frac{3}{2e} \le x_2 \le 0, \tag{9.1}$$

which is different from the curve GKC in Figure 8.2.

REMARK 9.1. *In the latter case, some analysis, not shown here, has been performed for the construction of the further backward trajectories starting from the last switching curve* $x_1 = (\frac{1}{3} - \frac{e^2}{9})x_2^2 + \frac{1}{4}$. *A new switching curve, now again for* v, *seems to result (depending on* a; *for a certain range of* a *no more switches arise). However, the trajectories do not pass this new switching curve, but are reflected. This phenomenon has not yet been investigated and/or clarified. Also a "void" (see* [7]) *seems to show up here.*

In subsection 9.3 *we will see that this approach with* δ*-functions does not lead to the correct answers, at least not when these* δ*-functions are considered as limits of rectangles with arbitrarily small basis and with area* 1.

REMARK 9.2. *It follows directly from* (7.3) *that* u *cannot be singular on a time interval with positive length. The same remark holds for* v.

9.2. By means of HJB–The synthesis approach. The value functions $V_i(x_1, x_2)$, $i = 1, 2$, provided they exist, satisfy the coupled set of HJB equations

$$1 + \frac{\partial V_1}{\partial x_1} x_2 - \left| \frac{\partial V_1}{\partial x_2} \right| + \frac{1}{2} \frac{\partial V_1}{\partial x_2} \operatorname{sgn} \frac{\partial V_2}{\partial x_2} = 0, \tag{9.2}$$

$$\frac{\partial V_2}{\partial x_1} x_2 + \frac{1}{2} \left| \frac{\partial V_2}{\partial x_2} \right| - \frac{\partial V_2}{\partial x_2} \operatorname{sgn} \frac{\partial V_1}{\partial x_2} = 0. \tag{9.3}$$

The boundary conditions are

$$V_1(x_1 = x_2) = 0, \quad V_2(x_1 = x_2) = x_1. \tag{9.4}$$

What we want to know is whether the synthesized open-loop solutions also form feedback solutions. For this example the synthesized controls will have the format $u(x_1, x_2)$ and $v(x_1, x_2)$. The central question of this section is whether these feedback functions will constitute the feedback optimal strategies.

In principle, we can calculate such synthesized V_i functions by tracing the open-loop trajectories and calculate the time till termination (for the u-player) and the termination point (for the v-player). Take, for instance, a point $(x_1 = p,\ x_2 = q)$ in the neighborhood of the point $(x_1 = 1,\ x_2 = -1)$. In this area $\frac{\partial V_1}{\partial x_2} > 0$ and $\frac{\partial V_2}{\partial x_2} > 0$, and hence the set of two HJB equations becomes two uncoupled ones:

$$1 + \frac{\partial V_1}{\partial x_1} x_2 - \frac{1}{2} \frac{\partial V_1}{\partial x_2} = 0, \tag{9.5}$$

$$\frac{\partial V_2}{\partial x_1} x_2 - \frac{1}{2} \frac{\partial V_2}{\partial x_2} = 0. \tag{9.6}$$

The (open-loop) optimal trajectory through the point (p, q) intersects the switching curve CKG at a point $(x_{1\mathrm{s}},\ x_{2\mathrm{s}})$, say. Then

$$x_{2\mathrm{s}} = -\sqrt{\frac{9}{11}\left(p + q^2 - \frac{1}{4}\right)}.$$

The time needed to go from $(x_1 = p,\ x_2 = q)$ to $(x_{1\mathrm{s}},\ x_{2\mathrm{s}})$ equals $2(q - x_{2\mathrm{s}})$, and hence

$$V_1(p, q) = 1 + 2\left(q + \sqrt{\frac{9}{11}\left(p + q^2 - \frac{1}{4}\right)}\right).$$

Now it easily follows that this V_1 is indeed a solution of (9.5). Along the same lines,

$$V_2(p, q) = \frac{1 - \sqrt{\frac{4}{11}(p + q^2) - \frac{1}{11}}}{2},$$

and this function turns out to be a solution of (9.6).

Right away we cannot conclude that the V's found are the correct value functions, since the verification theorem, a sufficiency result, does not hold. (This verification theorem (see, e.g., [7]) would apply if the value functions found would be twice differentiable, which is not true in our situation.) Despite the fact that V_1 satisfies (9.5) and V_2 (9.6), the following argument will give rise to some doubts as to whether these

value functions are the ones we are looking for. For that purpose consider initial points in the first quadrant, to the right of curve BD, e.g., point ($x_1 = \frac{3}{2},\ x_2 = \frac{1}{2}$). It pays here for the v-player to play $v = -1$ rather than $v = +1$ according to the open-loop solution. With $v = -1$ the trajectory moves in direction south-south-east, which leads to a better result for the v-player than the indicated open-loop solution (with $v = +1$), which moves in the direction south-east. In this part of the state space it therefore looks plausible that the optimal feedback strategies are $u = -1,\ v = -1$, which have a smooth continuation with the trajectories in the fourth quadrant as indicated in Figure 8.2.

Some further ad hoc calculations seem to point to the following feedback solution. Starting from Figure 8.2, the switching curve GKC is replaced by the parabola $x_1 = -\frac{2}{3}x_2^2 + \frac{1}{4}$ for $-\frac{1}{2} \le x_2 \le 0$ (see subsection 9.3 for the derivation) and the parabola $x_1 = \frac{1}{3}x_2^2$ for $-\frac{3}{2} \le x_2 \le -\frac{1}{2}$. The patchwork surrounded by the letters NABDG remains roughly as it is (though the position of the curve NM will be slightly higher). In the north-east of the state space, the optimal strategies are $u = -1$ and $v = -1$ as made plausible in the previous paragraph.

9.3. By means of additional common sense. The real optimal feedback solution as we think of it is given in Figure 9.1. Based on the remarks made earlier, $v = -1$ for all points x for which $x_2 > 0$, and $v = +1$ for all points x for which $x_2 < 0$. The switching curve GE for u, with formula $x_1 = -\frac{2}{3}x_2^2 + \frac{1}{4}$, has been obtained as follows. It is assumed that the trajectory starting from x_1, x_2 with $\frac{1}{4} \le x_1 \le \frac{1}{3}$ and $x_2 = 0$ will do this with $u = -1$ and $v = +1$. At a certain time, indicated by the point $x_2 = \gamma$, the optimal u control will switch to $+1$, and then the trajectory will continue with $u = v = +1$ until it reaches $x_2 = 0$, where v will now switch to the value -1, and finally the trajectory ends on $a = x_1 = x_2,\ 0 \le a \le \frac{1}{2}$. Minimizing the total time needed with respect to γ leads to the curve GE. Note that the equation for this curve is different from (9.1). This shows that one cannot formally proceed with δ-functions as we did previously; that may lead to wrong answers.

What remains to be shown is that the trajectories starting at x with $\frac{1}{3} < x_1 < 3$ and $x_2 = 0$ will continue with $u = -1$ and $v = +1$ until they reach the parabola EFC with equation $x_1 = \frac{1}{3}x_2^2$. This parabola is a semiuniversal surface as introduced in [7, pp. 196, 197]. Other options for optimal trajectories are indicated in Figure 9.2. The curved cone enclosed by FE with equation $x_1 = -x_2^2 + \frac{1}{3}$ (i.e., the optimal trajectory upstream point E in Figure 9.1) and by EC with equation $x_1 = \frac{1}{3}x_2^2$ can be considered as a void. Point E in Figure 9.2 is the same as in Figure 9.1. Suppose the point J is an arbitrary initial point within this void. Three possible candidate optimal trajectories are given in the figure which all end up in point E. Which one is the optimal one? It turns out to be the trajectory from J to P (with $u = -1,\ v = +1$) and then to E (with $u = v = +1$). This follows if one applies the approach with Green's theorem as given in [6, pp. 120, 121].

Finally, the patchwork consisting of the optimal trajectories that terminate at $a = \frac{3}{2}$ has to be hung in the already constructed field of solutions. Note that the dispersal line M'N' has a slightly different position when compared to the dispersal line MN of Figure 8.2. It is easy to show that the x_2 coordinate of point M' equals $\frac{2}{3}(1 - \sqrt{2})$.

Now we would like to show that (9.2) and (9.3) hold for these claimed optimal feedback trajectories. This is not straightforward, however. Take, for instance, again a point (p, q) in the neighborhood of $(x_1 = 1,\ x_2 = -1)$. In this area $V_2 \equiv 0$ and subsequently $\frac{\partial V_2}{\partial x_2} = 0$, and (9.2) is not well-defined. Both V_1, which equals

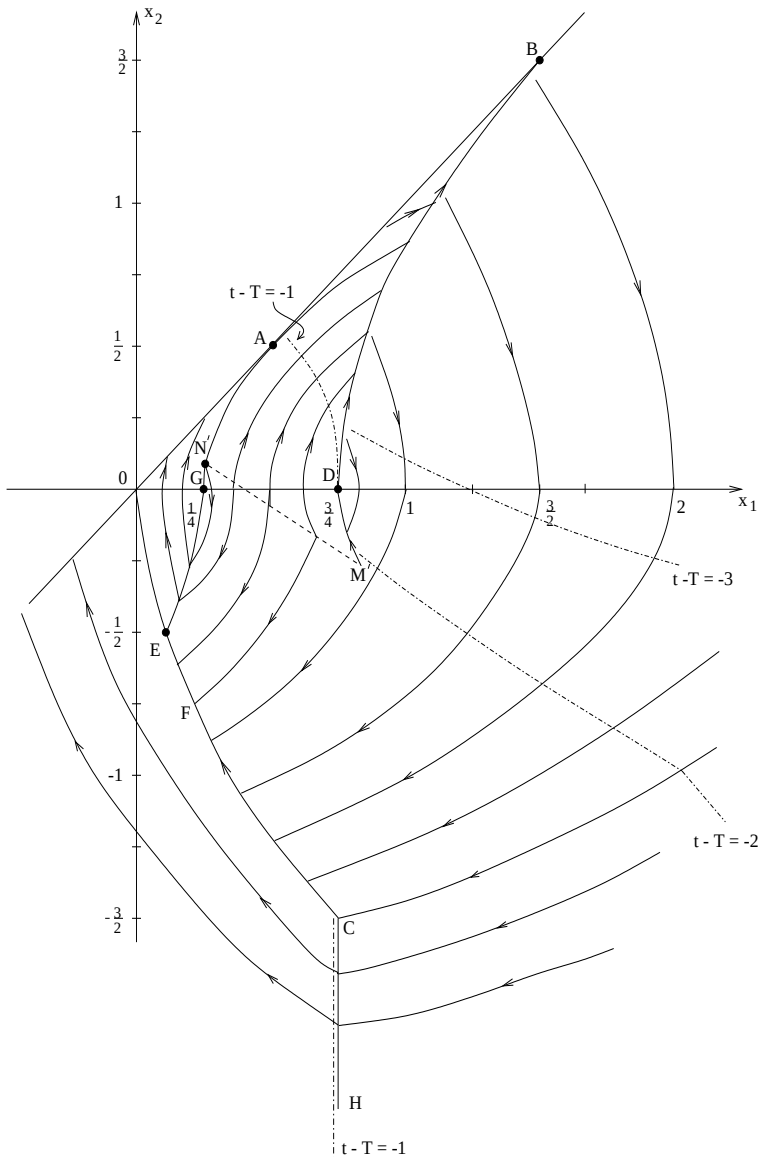


FIG. 9.1. *The optimal feedback solution.*

$2q + \frac{4}{3}\sqrt{3(p+q^2)}$, and $V_2 \equiv 0$ do satisfy (9.5) (respectively, (9.6)), however. Hence if we choose $\text{sgn}(\frac{\partial V_2}{\partial x_2}) = \text{sgn}(0) = 1$, then everything seems to fit. Suppose next that we take a point (p, q) in the neighborhood of $(x_1 = \frac{3}{2},\ x_2 = 1)$. For this area, $V_1 = \frac{2}{3} + \frac{4}{3}\sqrt{3p+q^2}$, $V_2 \equiv 0$, and hence again $\frac{\partial V_2}{\partial x_2} = 0$. We try $\text{sgn}(\frac{\partial V_2}{\partial x_2}) = \alpha$, for some suitable α with $-1 \leq \alpha \leq 1$ to be determined. A little analysis then shows that for $\alpha = -1$ equation (9.2) becomes an identity.

A conclusion of all this analysis is that the set of equations (9.2) and (9.3), together with the boundary conditions (9.4), allows at least two different solution sets (V_1, V_2), one sketched in Figure 8.2 and the other in Figure 9.1. An important question to be answered is whether a definition of viscosity solutions can be found such that a unique

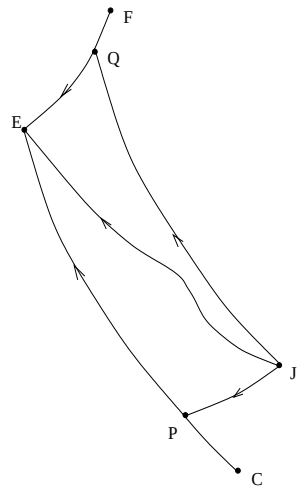


FIG. 9.2. *Comparing candidate optimal trajectories by Green's function method.*

solution of the PDEs and the boundary conditions results (and appropriate for the optimal feedback solutions). See [1], [2] for viscosity solution concepts for problems not as wild (i.e., discontinuous) as those considered in the current paper. Another reference in which differential equations are studied with non-Lipschitzian right-hand sides is [10]. (If we keep the feedback bang-bang control of one player fixed and view the remaining problem as an optimal control problem for the other player, then the model is such a non-Lipschitzian differential equation.)

10. Conclusions. Two simple, at least in their formulation, nonzero-sum dynamic games with bang-bang control have been discussed. Both have a unique optimal open-loop solution. For the first one, we found an optimal feedback solution which is almost identical to the synthesized open-loop solution; the difference is in the control values along a singular arc. For the second problem, two candidates for the optimal feedback solutions have been found. Neither jumping nor corner conditions for the costate equations along (or across) singular arcs nor appropriate extensions of the viscosity solution concept have been attempted to answer questions related to uniqueness of the optimality of the feedback solutions constructed. The solutions to these examples may help in formulating such concepts and with related analysis in the future.

It so happened that for both examples one could not uniquely solve the controls $u_i,\ i = 1, 2$, from $\dot{x} = f(x, u_1, u_2)$ and thus express them in $\dot{x}$ and x. In other words, given a trajectory in the (t, x) space (or in the x space for time-independent problems), there may be many controls which realize this given trajectory. Whether this feature leads to specific difficulties and/or phenomena is not known.

REFERENCES

[1] M. BARDI, M.G. CRANDALL, L.C. EVANS, AND H.M. SONER, *Viscosity Solutions and Applications*, Springer-Verlag, New York, 1995.

[2] M. BARDI AND I. CAPUZZO-DOLCETTA, *Optimal Control and Viscosity Solutions of Hamilton-Jacob-Bellman Equations*, Birkhäuser Boston, Cambridge, MA, 1997.

[3] T. BAŞAR AND G.J. OLSDER, *Dynamic Noncooperative Game Theory*, 2nd ed., SIAM, Philadelphia, 1998.

[4] A.E. Bryson and Y.-C. Ho, *Applied Optimal Control*, Ginn, Boston, 1969.
[5] W.H. Fleming and R.W. Rishel, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.
[6] H. Hermes and J.P. LaSalle, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.
[7] R. Isaacs, *Differential Games*, Wiley, New York, 1965.
[8] A.F. Kleimenov, *Nonantagonist Positional Differential Games*, Nauka, Ekaterinburg, Russia, 1993.
[9] A.F. Kleimenov, *Solutions in a non-antagonistic positional differential game*, J. Appl. Math. Mech., 61 (1997), pp. 717–723.
[10] N.N. Krasovskii and A.I. Subbotin, *Game-Theoretical Control Problems*, Springer-Verlag, New York, Berlin, 1988.
[11] J. Lewin, *Differential Games: Theory and Methods for Solving Game Problems with Singular Surfaces*, Springer-Verlag, New York, 1994.
[12] A.A. Melikyan, *Generalized Characteristics of First Order PDE's: Applications in Optimal Control and Differential Games*, Birkhäuser Boston, Cambridge, MA, 1998.
[13] G.J. Olsder, *Some thoughts about simple advertising models as differential games and the structure of coalitions*, in Directions in Large Scale Systems, Y.-C Ho and S.K. Mitter, eds., Plenum Press, New York, 1976, pp. 187–206.
[14] I.G. Sarma and U.R. Prasad, *Switching surfaces in N-person differential games*, J. Optim. Theory Appl., 10 (1972), pp. 160–177.
[15] A.W. Starr and Y.-C. Ho, *Nonzero-sum differential games*, J. Optim. Theory Appl., 3 (1969), pp. 184–206.
[16] A.W. Starr and Y.-C. Ho, *Further properties of nonzero-sum differential games*, J. Optim. Theory Appl., 3 (1969), pp. 207–219.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 4, pp. 1107–1133

ANALYSIS AND OPTIMIZATION OF NONSMOOTH ARCHES*

A. IGNAT†, J. SPREKELS‡, AND D. TIBA‡§

Abstract. It is our aim to present a new treatment for some classical models of arches and for their optimization. In particular, our approach allows us to study nonsmooth arches, while the standard assumptions from the literature require $W^{3,\infty}$-regularity for the parametric representation. Moreover, by a duality-type argument, the deformation of the arches may be explicitly expressed by integral formulas.

As examples for the shape optimization problems under study, we mention the design of the middle curve of a clamped arch such that, under a prescribed load, the obtained deflection satisfies certain desired properties. In all cases, no smoothness is required for the design parameters.

Key words. Lipschitz arches, flexural arches, shape optimization

AMS subject classifications. 49Q10, 35J35, 34A55

PII. S0363012900374427

1. Introduction. If $\varphi\colon [0,1] \to \mathbb{R}^2$ is a smooth clamped arch and $c\colon [0,1] \to \mathbb{R}$ denotes its curvature, then the classical Kirchhoff–Love model (with normalized mechanical constants) is given by

$$\begin{aligned} &\int_0^1 \left[\frac{1}{\varepsilon}(v_1' - c\,v_2)(u_1' - c\,u_2) + (v_2' + c\,v_1)'(u_2' + c\,u_1)'\right] ds \\ &= \int_0^1 (f_1\,u_1 + f_2\,u_2)\,ds \qquad \forall\, u_1 \in H_0^1(0,1)\,, \quad \forall\, u_2 \in H_0^2(0,1)\,. \end{aligned} \tag{1.1}$$

Here, $\sqrt{\varepsilon}$ is the constant thickness of the arch; $v_1 \in H_0^1(0,1)$, $v_2 \in H_0^2(0,1)$ are the tangential, respectively the normal, components of the deformation in the local coordinate system associated with the arch; and $[f_1, f_2]$ is a similar representation of the forces, including the internal and external loading of the arch, which are assumed to act in the same plane.

A thorough presentation via Dirichlet's principle and Korn's inequality of the existence and the uniqueness of the solution for (1.1) may be found in Ciarlet [11, p. 432]. In Chenais and Paumier [8] the "locking" problem, in connection with the numerical approximation of (1.1) and of shells, is discussed: If the discretization parameter is of the same order as ε, then the obtained numerical approximation may be meaningless, and special finite element schemes are necessary in order to solve (1.1).

In section 1, we introduce a new variational formulation for (1.1), based on optimal control theory, which is valid also for Lipschitz (or, by reparametrization—see Remark 2.8—absolutely continuous) mappings φ. Using duality-type arguments, we derive explicit integration rules for (1.1). If φ is smooth, we show that our solution satisfies

*Received by the editors June 19, 2000; accepted for publication (in revised form) May 3, 2001; published electronically November 28, 2001.

http://www.siam.org/journals/sicon/40-4/37442.html

†Faculty of Computer Science, University "Al. I. Cuza," str. Berthelot 16, RO–6600 Iaşi, Romania (ancai@thor.infoiasi.ro).

‡Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstrasse 39, D–10117 Berlin, Germany (sprekels@wias-berlin.de, tiba@wias-berlin.de).

§Permanent address: Institute of Mathematics, Romanian Academy of Sciences, P.O. Box 1–764, RO–70700 Bucharest, Romania (dtiba@imar.ro).

(1.1). In the general case, if φ is approximated by a sequence of smooth functions φ_δ with $\delta \to 0$ (obtained by a regularization via Friedrichs mollifiers), the approximation remains valid for the corresponding solutions as well.

This shows that our variational formulation is a natural extension of (1.1) to the case of nonsmooth arches. It also provides, by its explicit character, a complete solution of the above-mentioned "locking" problem in dimension one. We also study the behavior for $\varepsilon \to 0$ and obtain, under the weak optimal control formulation of (1.1), the analogue of flexural models in the sense of Ciarlet [12]. Some of the results of this section were announced without proofs in Sprekels and Tiba [22]. Our arguments use neither the Dirichlet principle nor the Korn inequality. Moreover, although the arch may have an infinity of corners, we do not impose transmission conditions as were used by Geymonat and Sanchez-Palencia [14]—they are implicitly contained in our approach. Models for shells and rods, under low geometrical regularity conditions, are also discussed in Blouza and Le Dret [5] and Chapelle [7].

In order to make the basic ideas more transparent, we now present a very simple example of how our variational approach based on optimal control theory works. To this end, let us consider the fourth order boundary value problem:

$$y'''' = f \quad \text{in }]0,1[\,, \tag{1.2}$$

$$y(0) = y(1) = 0\,, \tag{1.3}$$

$$y''(0) = y''(1) = 0\,, \tag{1.4}$$

with $f \in L^2(0,1)$.

The usual variational approach to (1.2)–(1.4) is given by the minimization of the energy,

$$\operatorname*{Min}_{y \in H^2(0,1) \cap H_0^1(0,1)} \left\{ \frac{1}{2} \int_0^1 (y'')^2 \, ds - \int_0^1 f\, y \, ds \right\}. \tag{1.5}$$

We rewrite (1.5) as an unconstrained optimal control problem, namely,

$$\operatorname{Min} \left\{ \frac{1}{2} \int_0^1 \underline{z}^2 \, ds - \int_0^1 f\, y \, ds \right\}, \tag{1.6}$$

$$y'' = \underline{z} \quad \text{in }]0,1[\,, \tag{1.7}$$

$$y(0) = y(1) = 0\,. \tag{1.8}$$

Relations (1.6)–(1.8) define a standard control problem with the newly introduced unknown $\underline{z} \in L^2(0,1)$ playing the role of the control. By coercivity and strict convexity, the existence of a unique optimal pair $[y^*, \underline{z}^*] \in [H^2(0,1) \cap H_0^1(0,1)] \times L^2(0,1)$ follows immediately. The first order optimality conditions for (1.6)–(1.8) are expressed by (1.7), (1.8) (where $y = y^*$, $\underline{z} = \underline{z}^*$), the adjoint system

$$p'' = f \quad \text{in }]0,1[\,, \tag{1.9}$$

$$p(0) = p(1) = 0\,, \tag{1.10}$$

and the Pontryagin maximum principle

$$\underline{z}^* = p \quad \text{in } [0,1]\,. \tag{1.11}$$

This can easily be inferred from the Euler equation associated with (1.6). Eliminating p from (1.7)–(1.11), we obtain the usual decomposition of (1.2)–(1.4) as a system of two second order differential equations. That is, (1.2)–(1.4), or (1.5), or (1.6)–(1.8), or (1.7)–(1.11) all constitute equivalent formulations of the same problem. Notice as well that (1.9)–(1.11) yield the regularity $\underline{z}^* \in H^2(0,1) \cap H_0^1(0,1)$, although the control space is only $L^2(0,1)$. (Such regularity properties are specific for unconstrained control problems.) This shows that the solution to (1.6)–(1.8) induces a strong solution for (1.2) with maximal regularity corresponding to $f \in L^2(0,1)$.

Next, we further modify (1.6) by integrating twice by parts in the second integral. If we denote by $g \in H^2(0,1) \cap H_0^1(0,1)$ the unique solution to $g'' = f, g(0) = g(1) = 0$, then the cost functional (1.6) can be rewritten as

$$\frac{1}{2}\int_0^1 \underline{z}^2\,ds - \int_0^1 f\,y\,ds = \frac{1}{2}\int_0^1 \underline{z}^2\,ds - \int_0^1 g\,\underline{z}\,ds = \frac{1}{2}\int_0^1 (\underline{z}-g)^2\,ds - \frac{1}{2}\int_0^1 g^2\,ds.$$

We redenote $\underline{z} - g$ again by $\underline{z}$. Then the control problem (1.6)–(1.8) becomes

$$\text{Min}\left\{\frac{1}{2}\int_0^1 \underline{z}^2\,ds\right\}, \tag{1.6}'$$

$$y'' = \underline{z} + g \quad \text{in }]0,1[\,, \tag{1.7}'$$

$$y(0) = y(1) = 0\,. \tag{1.8}'$$

Variants of such transformations and other mathematical modifications will be applied in section 2 to (1.1). In this respect, the mapping g (or g_1, g_2 defined in (2.2), (2.3)) will be used to write the explicit integration rules.

Moreover, in view of how the control problem (1.6)–(1.8) was introduced by starting from the quadratic functional (1.5), in our approach to problem (1.1) the following differential control system will play a key role:

$$v_1' - c\,v_2 = z_1 \quad \text{in }]0,1[\,, \tag{1.12}$$

$$v_2' + c\,v_1 = z_2 \quad \text{in }]0,1[\,. \tag{1.13}$$

The corresponding correct boundary conditions, the control mappings z_1, z_2, and the adequate notion of a weak solution to (1.12), (1.13) will be defined in detail in section 2.

In section 3, we use the optimal control formulation from the previous section in its equivalent form obtained by a variant of Pontryagin's maximum principle. For given $[f_1, f_2]$, we study the shape optimization problem of finding φ in a closed bounded subset of the space of Lipschitz arches, such that the obtained deflection $[v_1, v_2]$ has certain desired properties.

It should be noted that in this setting the considered optimization problem appears as a nonconvex control-into-coefficients problem. We prove the existence of the minimizer and derive the first order optimality conditions by computing the directional derivative of the cost. Similar problems were studied by Rousselet, Piekarski, and Myslinski [17], Chenais and Rousselet [9], and Chenais, Rousselet, and Benedict [10] under differentiability assumptions.

The last section collects numerical experiments related to arches and to their optimization. For simple input functions, the deformations can be computed by MAPLE. In the optimization case, local gradient methods are combined with some

global search, due to the nonconvexity of the problem. We have succeeded in finding, in some examples, global minimum points which have been theoretically justified a posteriori.

Finally, we point out that the core of our methods is a variety of special decompositions of (1.1) obtained via the first order optimality conditions for appropriately defined control problems. In this respect, the present work continues the investigations of Sprekels and Tiba from [18, 19, 20, 21, 22]. In particular, similar results may be obtained in the case of clamped plate models involving a discontinuous thickness (see Sprekels and Tiba [23]). The main tools that we use here are control theory and duality.

2. The control approach. Let $\theta(t)$ denote the angle between the tangent vector to the arch (given by φ') and the horizontal axis. If φ is smooth, then $\theta' = c$ (see [11, p. 432]). If $\varphi \in (W^{1,\infty}(0,1))^2$, then $\theta \in L^\infty(0,1)$, and this is the assumption we impose in what follows. Note that in this case the variational formulation (1.1) is not meaningful. However, it is still possible to define mild solutions for the system (1.12), (1.13) by the variation of constants formula (see (2.4)).

To this end, we introduce the fundamental matrix W (see Pontryagin [16, p. 110]) of the homogeneous linear ODE system $v_1' = c\,v_2$, $v_2' = -c\,v_1$:

$$W(t) = \begin{pmatrix} \cos\theta(t) & \sin\theta(t) \\ -\sin\theta(t) & \cos\theta(t) \end{pmatrix}, \tag{2.1}$$

which is meaningful for $\theta \in L^\infty(0,1)$. The affine part of the control system (compare with (1.7)′) is here given by the functions l, h, g_1, g_2 that are constructed from $f_1, f_2 \in L^2(0,1)$ as follows:

$$g_1 = \varepsilon\, l, \qquad -g_2'' = h, \qquad g_2(0) = g_2(1) = 0, \tag{2.2}$$

$$\begin{bmatrix} l \\ h \end{bmatrix}(t) = -\int_0^t W(t)\, W^{-1}(s) \begin{bmatrix} f_1(s) \\ f_2(s) \end{bmatrix} ds. \tag{2.3}$$

While (2.3) is similar to the definition of g from (1.7)′ and uses the state operator (2.4), relation (2.2) takes into account the ε and the supplementary derivative from the second integrand in (1.1). We then define the control system corresponding to (1.12), (1.13) and with z_1, z_2 replaced by $u + g_1, z + g_2$, respectively:

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}(t) := \int_0^t W(t)\, W^{-1}(s) \begin{bmatrix} u + g_1 \\ z + g_2 \end{bmatrix}(s)\, ds. \tag{2.4}$$

Since (2.4) takes into account just the initial conditions appearing in (1.1), the unconstrained problem (1.6)′–(1.8)′ is replaced by the constrained optimal control problem

$$\mathrm{Min}\left\{ \frac{1}{2\varepsilon} \int_0^1 u^2\, ds + \frac{1}{2} \int_0^1 (z')^2\, ds \right\}, \tag{P_ε}$$

subject to $u \in L^2(0,1)$, $z \in H_0^1(0,1)$, such that the mappings $[v_1, v_2] \in (L^\infty(0,1))^2$, given by (2.4), satisfy $v_1(1) = v_2(1) = 0$ in the sense that

$$\int_0^1 W^{-1}(s) \begin{bmatrix} u(s) + g_1(s) \\ z(s) + g_2(s) \end{bmatrix} ds = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{2.5}$$

Clearly, $u = -g_1$, $z = -g_2$ give an admissible control pair for (P_ε). From the coercivity and strict convexity of the cost, there follows the existence of a unique minimizer $[u_\varepsilon, z_\varepsilon] \in L^2(0,1) \times H_0^1(0,1)$.

Denote by $S \subset L^2(0,1) \times H_0^1(0,1)$ the closed subspace of admissible variations for (P_ε). Then, $[\mu, \xi] \in S$ if and only if

$$\int_0^1 W^{-1}(s) \begin{bmatrix} \mu(s) \\ \xi(s) \end{bmatrix} ds = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{2.6}$$

The Euler equation associated with $[u_\varepsilon, z_\varepsilon]$ is

$$\frac{1}{\varepsilon} \int_0^1 u_\varepsilon \, \mu \, ds + \int_0^1 z'_\varepsilon \, \xi' \, ds = 0 \quad \forall \, [\mu, \xi] \in S. \tag{2.7}$$

In particular, (2.7) says that $[u_\varepsilon, z_\varepsilon] \in S_\varepsilon^\perp$, where $S_\varepsilon^\perp$ denotes the orthogonal subspace of $S \subset L^2(0,1) \times H_0^1(0,1)$ with respect to the modified scalar product defined by the left-hand side of (2.7).

Remark 2.1. If $\theta \in W^{1,1}(0,1)$, then $c \in L^1(0,1)$, and relation (2.4) can be written in differential form as

$$v_1' - c\, v_2 = u + g_1 \qquad \text{a.e. in } (0,1), \tag{2.8}$$

$$v_2' + c\, v_1 = z + g_2 \qquad \text{a.e. in } (0,1). \tag{2.9}$$

Relation (2.4) gives the "mild" solution of (2.8), (2.9) with null initial conditions in the sense of semigroup theory; see Bénilan [4], Barbu [3]. If (2.8), (2.9) give the state equations of the control problem (P_ε), then (2.5) is a state constraint. It is expressed directly in the form of a control constraint, since the system (2.8), (2.9) is integrated by (2.4), and $W(t)$ is a nonsingular matrix.

We denote by $[v_1^\varepsilon, v_2^\varepsilon] \in (L^\infty(0,1))^2$ the optimal state of (P_ε), obtained from $[u_\varepsilon, z_\varepsilon]$ via (2.4).

THEOREM 2.2. *If $\varphi \in (W^{3,\infty}(0,1))^2$, then $[v_1^\varepsilon, v_2^\varepsilon]$ is the solution to* (1.1).

Proof. Under this regularity assumption, (2.4) can be written in the form (2.8), (2.9).

For any $u_1 \in H_0^1(0,1)$, $u_2 \in H_0^2(0,1)$, we introduce

$$\tilde{\mu} = u_1' - c\, u_2 \in L^2(0,1), \tag{2.10}$$

$$\tilde{\xi} = u_2' + c\, u_1 \in H_0^1(0,1), \tag{2.11}$$

and we have, consequently,

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}(t) = \int_0^t W(t)\, W^{-1}(s) \begin{bmatrix} \tilde{\mu} \\ \tilde{\xi} \end{bmatrix}(s)\, ds. \tag{2.12}$$

Since u_1, u_2 vanish at both ends of $[0,1]$, it follows from (2.12) and (2.6) that $[\tilde{\mu}, \tilde{\xi}] \in S$. Hence they may be used in (2.7). Taking into account relations (2.8), (2.9) satisfied by $v_1^\varepsilon, v_2^\varepsilon$, as well as (2.10), (2.11), and (2.2), we obtain that

$$\begin{aligned} 0 &= \frac{1}{\varepsilon} \int_0^1 ((v_1^\varepsilon)' - c\, v_2^\varepsilon - g_1)(u_1' - c u_2) ds + \int_0^1 ((v_2^\varepsilon)' + c\, v_1^\varepsilon - g_2)'\, (u_2' + c\, u_1)'\, ds \\ &= \frac{1}{\varepsilon} \int_0^1 \big((v_1^\varepsilon)' - c\, v_2^\varepsilon\big)(u_1' - c\, u_2)\, ds + \int_0^1 ((v_2^\varepsilon)' + c\, v_1^\varepsilon)'(u_2' + c\, u_1)'\, ds \\ &\quad - \int_0^1 l(u_1' - c\, u_2)\, ds - \int_0^1 h(u_2' + c\, u_1)\, ds. \end{aligned}$$

By the regularity assumption, (2.3) can be rewritten in the differential form (2.8), (2.9), and we can infer that

$$\int_0^1 l(u_1' - c\,u_2)\,ds + \int_0^1 h(u_2' - c\,u_1)\,ds$$
$$= -\int_0^1 u_1(l' - c\,h)\,ds - \int_0^1 u_2(h' + c\,l)\,ds = \int_0^1 (f_1\,u_1 + f_2\,u_2)\,ds.$$

The last two relations give (1.1) and the proof is finished. □

Remark 2.3. The approach via problem (P_ε) is constructive and does not use either Dirichlet's principle or Korn's inequality. As the formulation of (P_ε) is valid for $\theta \in L^\infty(0,1)$, this method may give solutions even in nonsmooth situations when Korn's inequality is not valid. For such cases, we refer to Geymonat and Gilardi [13].

In the general case, the following extension of Theorem 2.2 holds true.

THEOREM 2.4. *If* $\varphi \in (W^{1,\infty}(0,1))^2$, *then we have for any* $[\mu,\xi] \in S$

$$\frac{1}{\varepsilon}\int_0^1 (u_\varepsilon + g_1)\mu\,ds + \int_0^1 (z_\varepsilon + g_2)'\,\xi'\,ds = \int_0^1 (f_1\,u_1 + f_2\,u_2)\,ds\,, \tag{2.13}$$

with $u_1, u_2 \in L^\infty(0,1)$ *given by*

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}(s) = -\int_s^1 W(s)\,W^{-1}(t)\begin{bmatrix} \mu(t) \\ \xi(t) \end{bmatrix} dt \qquad \text{for almost every } s \in (0,1). \tag{2.14}$$

Proof. Since $[u_\varepsilon, z_\varepsilon] \in S_\varepsilon^\perp$ (see (2.7)), we obtain that

$$\frac{1}{\varepsilon}\int_0^1 (u_\varepsilon + g_1)\mu\,ds + \int_0^1 (z_\varepsilon + g_1)'\xi'\,ds$$
$$= \frac{1}{\varepsilon}\int_0^1 g_1\,\mu\,ds - \int_0^1 \xi\,g_2''\,ds = \int_0^1 [\mu,\xi]\begin{bmatrix} l \\ h \end{bmatrix} dt$$
$$= -\int_0^1 [\mu,\xi](t)\int_0^t W(t)\,W^{-1}(s)\begin{bmatrix} f_1(s) \\ f_2(s) \end{bmatrix} ds\,dt$$
$$= -\int_0^1\int_0^t \big[f_1(s), f_2(s)\big]W(s)\,W^{-1}(t)\begin{bmatrix} \mu(t) \\ \xi(t) \end{bmatrix} ds\,dt,$$

due to the orthogonality of the matrix $W(t)$ and to (2.2), (2.3). Fubini's theorem and (2.14) imply the result. □

Remark 2.5. It is possible to prove Theorem 2.2 via Theorem 2.4. These results show that the problem (P_ε) provides a notion of weak solution for the arch problem which is a natural extension of the classical one. This will be further justified below in Theorem 3.2 and Remark 3.6, via an approximation argument.

We now introduce the mappings $w_1, w_2 \in H^2(0,1) \cap H_0^1(0,1)$ given by

$$w_1''(s) = \sin\theta(s) \qquad \text{a.e. in } (0,1), \tag{2.15}$$

$$w_2''(s) = -\cos\theta(s) \qquad \text{a.e. in } (0,1). \tag{2.16}$$

Taking into account (2.1), relation (2.6) can be rewritten as

$$\int_0^1 \Big[\mu(s)\,\cos\theta(s) - \xi(s)\,\sin\theta(s)\Big]\,ds = 0,$$

$$\int_0^1 \Big[\mu(s)\,\sin\theta(s)\,-\,\xi(s)\,\cos\theta(s)\Big]\,ds\,=\,0.$$

Replacing the coefficients of $\xi(s)$ according to (2.15), (2.16) and integrating once by parts, relation (2.6) may be put into the equivalent form

$$\frac{1}{\varepsilon}\int_0^1 \varepsilon\,\cos\theta(s)\,\mu(s)\,ds\,+\,\int_0^1 w_1'(s)\,\xi'(s)\,ds\,=\,0, \tag{2.17}$$

$$\frac{1}{\varepsilon}\int_0^1 \varepsilon\,\sin\theta(s)\,\mu(s)\,ds\,+\,\int_0^1 w_2'(s)\,\xi'(s)\,ds\,=\,0. \tag{2.18}$$

From the definition of S using the modified scalar product from (2.7) it follows that the (linearly independent) vectors $[\varepsilon\cos\theta(\cdot), w_1(\cdot)]$ and $[\varepsilon\sin\theta(\cdot), w_2(\cdot)]$ provide a basis of the two-dimensional space $S_\varepsilon^\perp$.

In addition, from relations (2.5) and (2.6) we can infer that $[u_\varepsilon+g_1\,,\,z_\varepsilon+g_2]\in S$. Consequently, relation (2.7) gives us that

$$[u_\varepsilon, z_\varepsilon]\,=\,-\,\mathrm{proj}_{S_\varepsilon^\perp}[g_1, g_2], \tag{2.19}$$

where the projection is computed in the norm generated by the modified scalar product from (2.7).

Then, (2.17)–(2.19) yield that

$$[u_\varepsilon, z_\varepsilon]\,=\,\lambda_1^\varepsilon\,[\varepsilon\,\cos\theta\,,\,w_1]+\lambda_2^\varepsilon\,[\varepsilon\,\sin\theta\,,\,w_2] \tag{2.19$'$}$$

for some $\lambda_1^\varepsilon, \lambda_2^\varepsilon\in\mathbb{R}$. By virtue of the definition of the projection operator, and owing to (2.19), (2.19)$'$, we see that $(\lambda_1^\varepsilon,\,\lambda_2^\varepsilon)$ is the unique minimizer of the unconstrained optimization problem

$$\begin{aligned}(\mathrm{D}_\varepsilon)\qquad \min_{\lambda_1,\,\lambda_2\,\in\,\mathbb{R}}\ \Bigg\{&\frac{1}{2\varepsilon}\int_0^1\Big(\lambda_1\,\varepsilon\,\cos\theta(s)\,+\,\lambda_2\,\varepsilon\,\sin\theta(s)\,+\,\varepsilon\,l(s)\Big)^2\,ds\\ &+\,\frac{1}{2}\int_0^1\Big[(\lambda_1\,w_1\,+\,\lambda_2\,w_2\,+\,g_2)'\Big]^2\,ds\Bigg\}.\end{aligned}$$

Problem (D_ε) can be solved explicitly by imposing that the derivatives of the quadratic form with respect to λ_1, λ_2, are zero at the optimum point. This gives a linear algebraic system with a strictly positive determinant (by the Cauchy–Schwarz inequality and the structure of the basis of $S_\varepsilon^\perp$). We indicate the system for subsequent use:

$$\begin{aligned}&\varepsilon\,\lambda_1\int_0^1\cos^2\,\theta(s)\,ds\,+\,\lambda_1\,\big|w_1\big|^2_{H_0^1(0,1)}\,+\,\varepsilon\,\lambda_2\int_0^1\cos\theta(s)\,\sin\theta(s)\,ds\\ &\quad+\,\lambda_2\int_0^1 w_1'(s)\,w_2'(s)\,ds\,+\,\varepsilon\int_0^1 l(s)\,\cos\theta(s)\,ds\,+\,\int_0^1 g_2'(s)\,w_1'(s)\,ds\,=\,0,\\ &\varepsilon\,\lambda_1\int_0^1\cos\theta(s)\,\sin\theta(s)\,ds\,+\,\lambda_1\int_0^1 w_1'(s)\,w_2'(s)\,ds\,+\,\varepsilon\,\lambda_2\int_0^1\sin^2\theta(s)\,ds\\ &\quad+\,\lambda_2\big|w_2\big|^2_{H_0^1(0,1)}\,+\,\varepsilon\int_0^1 l(s)\sin\theta(s)\,ds\,+\,\int_0^1 g_2'(s)\,w_2'(s)\,ds\,=\,0.\end{aligned} \tag{2.20}$$

We have proved the following result.

THEOREM 2.6. *The solution of* (1.1) *(or of* (P_ε)*, if* $\theta \in L^\infty(0,1)$*) is given by* (2.19)′ *and* (2.4)*, with* $(\lambda_1^\varepsilon, \lambda_2^\varepsilon)$ *being the unique solution of* (D_ε)*, and with* w_1, w_2*,* g_1, g_2 *defined by* (2.2), (2.3), (2.15), (2.16).

Remark 2.7. In optimization theory, (D_ε) is the dual problem of (P_ε). Its complete solution is possible since the constraints from (P_ε) are affine and finite-dimensional. In simple examples of mappings θ, f_1, f_2, explicit formulas can be derived for the deformation $[v_1, v_2]$. In the general situation, numerical approximation is needed just to evaluate the occurring integrals. See section 4 for examples. In particular, Theorem 2.6 provides a complete solution of the "locking" problem discussed by Chenais and Paumier [8], in dimension one.

Remark 2.8. We also notice that, if $\tilde{\varphi}\colon [a,b] \to \mathbb{R}^2$ is an absolutely continuous Jordan arc of length one such that $\tilde{\varphi}' \neq 0$ a.e. in (a,b), then, by the usual reparametrization via the arc length function $s\colon [a,b] \to [0,1]$, $s(0) = 0$, $s'(\cdot) = \left|\tilde{\varphi}'(\cdot)\right|_{\mathbb{R}^2}$, we get that $\varphi(t) = \tilde{\varphi}(s^{-1}(t))$ satisfies $\left|\varphi'(t)\right|_{\mathbb{R}^2} = 1$ for almost every $t \in (0,1)$, i.e., it is Lipschitzian, and our results still apply.

Remark 2.9. If $\theta \in L^\infty(0,1)$, then $v_1^\varepsilon, v_2^\varepsilon$ as defined by Theorem 2.6 (see (2.4)) belong to $L^\infty(0,1)$. However, their global Cartesian representation is

$$W(t)^{-1} \begin{bmatrix} v_1^\varepsilon \\ v_2^\varepsilon \end{bmatrix}(t)$$

and belongs to $(W^{1,2}(0,1))^2$. This means that the lack of smoothness is due to the local coordinates (θ is defined a.e. and may have jumps) and that the constructed deformation is continuous.

The next result gives a characterization of the solution of the problem (P_ε) (or, equivalently, of the problem (D_ε)) as a system of first order differential equations which will be used frequently in what follows. Implicitly, it provides a nonstandard decomposition of (1.1) in the case of nonsmooth coefficients. Basically, this is given by the first order necessary conditions for (P_ε), but the form is different from the classical Pontryagin principle.

THEOREM 2.10. *The optimality system for the problem* (P_ε) *is given by*

$$\text{(2.21)}\quad \begin{bmatrix} v_1^\varepsilon \\ v_2^\varepsilon \end{bmatrix}(t) = \int_0^t W(t)\, W^{-1}(s) \begin{bmatrix} u_\varepsilon(s) + g_1(s) \\ z_\varepsilon(s) + g_2(s) \end{bmatrix} ds \quad \textit{for almost every } t \in (0,1)\,,$$

$$\text{(2.22)}\quad \int_0^1 W^{-1}(s) \begin{bmatrix} u_\varepsilon(s) + g_1(s) \\ z_\varepsilon(s) + g_2(s) \end{bmatrix} ds = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

$$\text{(2.23)}\quad \begin{bmatrix} p_\varepsilon \\ q_\varepsilon \end{bmatrix}(t) = W(t) \begin{bmatrix} \lambda_1^\varepsilon \\ \lambda_2^\varepsilon \end{bmatrix} \quad \textit{for almost every } t \in (0,1)\,,$$

$$\text{(2.24)}\quad u_\varepsilon = \varepsilon\, p_\varepsilon \qquad \textit{a.e. in } (0,1)\,,$$

$$\text{(2.25)}\quad z_\varepsilon'' = -\, q_\varepsilon \quad \textit{a.e. in } (0,1)\,, \quad z_\varepsilon(0) = z_\varepsilon(1) = 0.$$

Under smoothness hypotheses, it is possible to write (2.21)–(2.25) in differential form. (Compare with (3.1)–(3.5) and (3.7)–(3.16) in the next section.)

Proof. Assume first that u_ε, z_ε satisfy (2.21)–(2.25) with some $\lambda_1^\varepsilon, \lambda_2^\varepsilon \in \mathbb{R}$, $p_\varepsilon, q_\varepsilon, v_1^\varepsilon, v_2^\varepsilon \in L^\infty(0,1)$. Then clearly, $[u_\varepsilon + g_1, z_\varepsilon + g_2] \in S$, i.e., $[u_\varepsilon\,, z_\varepsilon]$ is admissible for (P_ε). Using (2.23)–(2.25), the definition of S, and the orthogonality of

$W(t)$, we find that for any $[\mu, \xi] \in S$

$$\frac{1}{\varepsilon}\int_0^1 u_\varepsilon\,\mu\,ds \;+\; \int_0^1 z_\varepsilon'\,\xi'\,ds = \int_0^1 p_\varepsilon\,\mu\,ds + \int_0^1 q_\varepsilon\,\xi\,ds = \int_0^1 [\mu,\xi]W(s)\begin{bmatrix}\lambda_1^\varepsilon\\ \lambda_2^\varepsilon\end{bmatrix} ds$$

$$= [\lambda_1^\varepsilon, \lambda_2^\varepsilon]\int_0^1 W(s)^{-1}\begin{bmatrix}\mu\\ \xi\end{bmatrix}(s)\,ds = 0.$$

Consequently, $[u_\varepsilon\,,\, z_\varepsilon] \in S_\varepsilon^\perp$. Together with the admissibility of $[u_\varepsilon, z_\varepsilon]$, noticed above, this immediately gives that $[u_\varepsilon\,,\, z_\varepsilon]$ is the unique minimizer of (P_ε).

Conversely, we remark that (2.23)–(2.25) give a complete description of the two-dimensional space $S_\varepsilon^\perp$ when $\lambda_1, \lambda_2 \in \mathbb{R}$ are arbitrary. By (2.6), we know that the optimal control $[u_\varepsilon\,,\, z_\varepsilon]$ belongs to $S_\varepsilon^\perp$. Hence, there are $\lambda_1^\varepsilon, \lambda_2^\varepsilon \in \mathbb{R}$ such that $[u_\varepsilon\,,\, z_\varepsilon]$ can be represented via (2.23)–(2.25). (This is, in fact, the same representation as in (2.19)′.) Moreover, $[u_\varepsilon\,,\, z_\varepsilon]$ also satisfy (2.21), (2.22) by their admissibility for (P_ε). This ends the proof. □

As a first application of Theorem 2.10, we study the behavior for $\varepsilon \to 0$ of the problem (P_ε). Since arches are special cases of cylindrical shells, after passing to the limit a "flexural" model will be obtained (Ciarlet [12]). The treatment that we indicate below is valid under the weak regularity condition $\theta \in L^\infty(0,1)$. We shall also assume that θ is nonconstant in $[0,1]$, i.e., the arch is not a bar. For constant θ the results also remain valid, but some adaptation of the argument is necessary, since the dimension of $S_\varepsilon^\perp$ reduces to one in this case, if $\varepsilon = 0$.

THEOREM 2.11. *As $\varepsilon \searrow 0$, the mappings $v_1^\varepsilon, v_2^\varepsilon, p_\varepsilon, q_\varepsilon$ are bounded in $L^\infty(0,1)$, $\lambda_1^\varepsilon\,,\, \lambda_2^\varepsilon$ are bounded in $\mathbb{R}$, z_ε is bounded in $H^2(0,1)$, and u_ε strongly converges to 0 in $L^\infty(0,1)$. If we denote without ε their weak or weak* limits (on a subsequence) in the corresponding spaces, then these satisfy the conditions*

$$\begin{bmatrix}v_1\\ v_2\end{bmatrix}(t) = \int_0^t W(t)\,W^{-1}(s)\begin{bmatrix}0\\ z(s)+g_2(s)\end{bmatrix} ds\,,$$

$$\int_0^1 W^{-1}(s)\begin{bmatrix}0\\ z(s)+g_2(s)\end{bmatrix} ds = \begin{bmatrix}0\\ 0\end{bmatrix},$$

$$\begin{bmatrix}p\\ q\end{bmatrix}(t) = W(t)\begin{bmatrix}\lambda_1\\ \lambda_2\end{bmatrix},$$

$$z'' = -q\,, \qquad z(0) = z(1) = 0\,.$$

Proof. The explicit calculus indicated in Theorem 2.6 and (2.20) shows directly that $\lambda_1^\varepsilon, \lambda_2^\varepsilon$ are bounded in $\mathbb{R}$ for $\varepsilon \to 0$. For instance, the determinant of the system is

$$\left[\varepsilon\int_0^1 \cos^2\theta(s)\,ds + \Big|w_1\Big|^2_{H_0^1(0,1)}\right]\cdot\left[\varepsilon\int_0^1 \sin^2\theta(s)\,ds + \Big|w_2\Big|^2_{H_0^1(0,1)}\right]$$

$$-\left[\varepsilon\int_0^1 \cos\theta(s)\cdot\sin\theta(s)\,ds + \int_0^1 w_1'(s)w_2'(s)\,ds\right]^2 \underset{\varepsilon\to 0}{\longrightarrow} \Big|w_1\Big|^2_{H_0^1}\cdot\Big|w_2\Big|^2_{H_0^1}$$

$$-\Big\langle w_1, w_2\Big\rangle^2_{H_0^1} > 0.$$

Here the assumption that θ is nonconstant is necessary, since for $\varepsilon = 0$ and θ constant the vectors used in (2.19)′ become proportional. (In this case only one parameter λ is necessary and a simpler argument works.)

Thus, by (2.23), p_ε and q_ε are bounded in $L^\infty(0,1)$. Relation (2.24) gives $u_\varepsilon \to 0$ strongly in $L^\infty(0,1)$, and (2.25) shows that z_ε is bounded in $H^2(0,1)$, for instance. By (2.21), we see that $v_1^\varepsilon, v_2^\varepsilon$ are bounded in $L^\infty(0,1)$ as well. Definition (2.2) gives that g_1 depends on ε (and has the strong limit 0 in $L^\infty(0,1)$), while g_2 is independent of ε.

Finally, we can pass to the limit in (2.21)–(2.25) on a subsequence, and we obtain the desired conclusion. □

Remark 2.12. The system obtained by Theorem 2.11 characterizes, in the sense of Theorem 2.10, the following constrained optimal control problem:

$$\mathrm{Min}\left\{\frac{1}{2}|z|^2_{H^1_0(0,1)}\right\},$$

subject to $z \in H^1_0(0,1)$, such that the mappings

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}(t) \,=\, \int_0^t W(t)\, W^{-1}(s) \begin{bmatrix} 0 \\ z(s) + g_2(s) \end{bmatrix} ds$$

satisfy $v_1(1) = v_2(1) = 0$ in the sense that

$$\int_0^1 W^{-1}(s) \begin{bmatrix} 0 \\ z(s) + g_2(s) \end{bmatrix} ds \,=\, \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

The structure of this problem is very similar to (P_ε), and the proof follows closely that of Theorem 2.10, by considering the subspace $Z \subset H^1_0(0,1)$, defined by

$$\int_0^1 W^{-1}(s) \begin{bmatrix} 0 \\ \xi(s) \end{bmatrix} ds \,=\, \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

and its orthogonal subspace $Z^\perp$. If $\theta \in L^\infty(0,1)$ is not constant, $Z^\perp$ has dimension two, and we can argue as above.

Remark 2.13. If $\theta \in W^{2,\infty}(0,1)$, then one can show, as in Theorem 2.2, that v_1, v_2 defined in Remark 2.12 satisfy the "flexural" arch model:

$$\int_0^1 (v_2' \,+\, c v_1)'\,(u_2' \,+\, c\, u_1)'\, ds \,=\, \int_0^1 (f_1\, u_1 \,+\, f_2\, u_2)\, ds$$

$$\forall\ (u_1, u_2) \in V_F \,=\, \Big\{(u_1, u_2) \in H^1_0(0,1) \,\times\, H^2_0(0,1)\,;\, u_1' \,-\, c\, u_2 \,=\, 0\Big\},$$

$$(v_1, v_2) \in V_F.$$

Such asymptotic properties have been discussed in detail by Ciarlet [12] for the case of shells. Theorem 2.11 shows that they remain valid for nonsmooth arches and under our variational formulation via optimal control theory.

3. Optimization of nonsmooth arches. One advantage of the method presented in the previous section is that in the study of related optimization problems nonsmooth arches may be taken into consideration. Let $\mathcal{K} \subset L^\infty(0,1)$ be a closed subset. We shall study the model problem

$$(\mathrm{Q}) \qquad \mathop{\mathrm{Min}}_{\theta\,\in\,\mathcal{K}} \left\{\frac{1}{2}\int_0^1 v_2^2\, ds\right\},$$

subject to

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}(t) = \int_0^t W_\theta(t)\, W_\theta^{-1}(s) \begin{bmatrix} u(s) + g_1(s) \\ z(s) + g_2(s) \end{bmatrix} ds \quad \text{for almost every } t \in (0,1), \tag{3.1}$$

$$\int_0^1 W_\theta^{-1}(s) \begin{bmatrix} u(s) + g_1(s) \\ z(s) + g_2(s) \end{bmatrix} ds = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tag{3.2}$$

$$\begin{bmatrix} p \\ q \end{bmatrix}(t) = W_\theta(t) \begin{bmatrix} \lambda_1^\varepsilon \\ \lambda_2^\varepsilon \end{bmatrix} \quad \text{for almost every } t \in (0,1), \tag{3.3}$$

$$u = \varepsilon\, p \quad \text{a.e. in } (0,1), \tag{3.4}$$

$$z'' = -\,q \quad \text{a.e. in } (0,1), \quad z(0) = z(1) = 0. \tag{3.5}$$

The matrix W_θ is given by (2.1), and the new notation just puts into evidence the dependence on the arch (characterized by θ). The state system (3.1)–(3.5) is exactly the decomposition of the Kirchhoff–Love model provided by Theorem 2.10. It should be noted that all the quantities appearing in it (including the data g_1, g_2 defined by (2.2), (2.3)) depend on θ. This is due to W_θ and to the fact that $[f_1, f_2]$ (the load) depends on θ by the local choice of the coordinates system. In what follows, we shall write $v_1(\theta), v_2(\theta), \lambda_1(\theta), \lambda_2(\theta)$, etc. ($\varepsilon$ is fixed now).

Remark 3.1. The shape optimization problem (Q) is a nonconvex control-into-coefficients problem. In the given subset $\mathcal{K}$, the arch that minimizes the normal deflection (in the L^2-norm) is sought. This is a natural safety requirement. Various other cost functionals may be studied as well.

THEOREM 3.2. *If $\theta_n \to \theta$ in $L^\infty(0,1)$ and $f_i(\theta_n) \to f_i(\theta)$ in $L^1(0,1), i = 1,2$, then $W_{\theta_n} \to W_\theta$ in $(L^\infty(0,1))^4$, $\lambda_1(\theta_n) \to \lambda_1(\theta)$, $\lambda_2(\theta_n) \to \lambda_2(\theta)$, $g_1(\theta_n) \to g_1(\theta)$, $h(\theta_n) \to h(\theta)$, and $l(\theta_n) \to l(\theta)$ in $L^\infty(0,1), g_2(\theta_n) \to g_2(\theta)$ in $W^{2,\infty}(0,1)$, $p(\theta_n) \to p(\theta)$, $u(\theta_n) \to u(\theta)$, and $q(\theta_n) \to q(\theta)$ in $L^\infty(0,1)$, $z(\theta_n) \to z(\theta)$ in $W^{2,\infty}(0,1)$, and $v_1(\theta_n) \to v_1(\theta)$, $v_2(\theta_n) \to v_2(\theta)$ in $L^\infty(0,1)$. If $\theta_n \to \theta$ in $C[0,1]$, then the above convergences are also valid in $C[0,1]$ and $C^2[0,1]$, respectively.*

Proof. If $\theta_n \to \theta$ in $L^\infty(0,1)$, then $\cos\theta_n \to \cos\theta$ and $\sin\theta_n \to \sin\theta$ in $L^\infty(0,1)$. Consequently, $W_{\theta_n} \to W_\theta$, $W_{\theta_n}^{-1} \to W_\theta^{-1}$, strongly in $(L^\infty(0,1))^4$. Moreover, (2.15), (2.16) show that $w_1(\theta_n) \to w_1(\theta)$ and $w_2(\theta_n) \to w_2(\theta)$ in $W^{2,\infty}(0,1)$. If $\Delta(\theta_n)$ is the determinant associated with the system (2.20) (written for θ_n), a direct calculus gives that $\Delta(\theta_n) \to \Delta(\theta)$.

From the relation (2.3) we infer that for almost every $t \in (0,1)$

$$\begin{aligned}
&\left| \begin{bmatrix} l(\theta_n) \\ h(\theta_n) \end{bmatrix}(t) - \begin{bmatrix} l(\theta) \\ h(\theta) \end{bmatrix}(t) \right|_{\mathbb{R}^2} \\
&\le \left| W_{\theta_n} - W_\theta \right|_{(L^\infty(0,1))^4} \left| W_{\theta_n}^{-1} \right|_{(L^\infty(0,1))^4} \left| \begin{bmatrix} f_1(\theta_n) \\ f_2(\theta_n) \end{bmatrix} \right|_{(L^1(0,1))^2} \\
&\quad + \left| W_\theta \right|_{(L^\infty(0,1))^4} \left| W_{\theta_n}^{-1} - W_\theta^{-1} \right|_{(L^\infty(0,1))^4} \left| \begin{bmatrix} f_1(\theta_n) \\ f_2(\theta_n) \end{bmatrix} \right|_{(L^1(0,1))^2} \\
&\quad + \left| W_\theta \right|_{(L^\infty(0,1))^4} \left| W_\theta^{-1} \right|_{(L^\infty(0,1))^4} \left| \begin{bmatrix} f_1(\theta_n) \\ f_2(\theta_n) \end{bmatrix} - \begin{bmatrix} f_1(\theta) \\ f_2(\theta) \end{bmatrix} \right|_{(L^1(0,1))^2}.
\end{aligned} \tag{3.6}$$

It follows that $l(\theta_n) \to l(\theta)$, $h(\theta_n) \to h(\theta)$, strongly in $L^\infty(0,1)$. By (2.2), the same is valid for $g_1(\theta_n) \to g_1(\theta)$, while $g_2(\theta_n) \to g_2(\theta)$ strongly in $W^{2,\infty}(0,1)$. Then one obtains $\lambda_1(\theta_n) \to \lambda_1(\theta)$ and $\lambda_2(\theta_n) \to \lambda_2(\theta)$ from (2.20).

Equations (3.3)–(3.5) give the assertion for $p(\theta_n)$, $q(\theta_n)$, $u(\theta_n)$, $z(\theta_n)$. The argument for the convergence $v_1(\theta_n) \to v_1(\theta)$, $v_2(\theta_n) \to v_2(\theta)$, strongly in $L^\infty(0,1)$, is similar to that in the inequality (3.6). If $\theta_n \to \theta$ in $C[0,1]$, the proof follows the same lines, with minor modifications. □

COROLLARY 3.3. *The shape optimization problem* (Q) *has at least one solution if* $\mathcal{K}$ *is compact in* $L^\infty(0,1)$.

Proof. This is a direct consequence of Theorem 3.2, observing that it is possible to pass to the limit in (3.2) and in the cost functional, if $\theta_n \to \theta$ strongly in $L^\infty(0,1)$. □

Remark 3.4. In addition to Remark 2.9, we notice that the convergence of the global Cartesian representation of the displacement

$$W_{\theta_n}^{-1}(t) \begin{bmatrix} v_1(\theta_n) \\ v_2(\theta_n) \end{bmatrix} (t)$$

is valid in $(W^{1,\infty}(0,1))^2$. Here, we also use the fact that by (3.4) the solution $[u,z]$ of the problem (P_ε) belongs to $(L^\infty(0,1))^2$.

Remark 3.5. If the curvature c corresponding to the arches associated with $\theta \in \mathcal{K}$ is bounded in some $L^r(0,1)$, $r > 1$, then $\mathcal{K}$ is compact in $C[0,1]$. This shows that the compactness assumption from Theorem 3.2 and Corollary 3.3 is very weak in comparison with those used in the existing literature.

Remark 3.6. For any $\theta \in L^\infty(0,1)$, we may define a smooth sequence θ_n converging to θ in $L^r(0,1)\,\forall\, r \geq 1$ by a regularization process with a Friedrichs mollifier. Then, keeping $[f_1, f_2] \in (L^2(0,1))^2$ fixed, it is possible to modify (3.6) and the other arguments in the proof of Theorem 3.2 to show that for the corresponding solutions we have $v_n^1 \to v^1$, $v_n^2 \to v^2$ in $L^r(0,1)\,\forall\, r \geq 1$. If θ is continuous, the obtained convergences are uniform. We also note that the global Cartesian representation

$$W_{\theta_n}^{-1}(t) \begin{bmatrix} v_n^1 \\ v_n^2 \end{bmatrix} (t)$$

is convergent in $(W^{1,r}(0,1))^2\,\forall\, r \geq 1$. Since for θ_n the corresponding solution of (P_ε) then coincides with the solution of (1.1) (by Theorem 2.2), we see that for any $\theta \in L^\infty(0,1)$ the optimal state of (P_ε) can be approximated by usual solutions of (1.1).

The remainder of this section is devoted to the sensitivity analysis of the Kirchhoff–Love model. We proceed in two steps. First, we assume that $c \in L^1(0,1)$ and that, consequently, $\theta \in W^{1,1}(0,1)$, and we compute the gradient of the cost in this case. Then, we use an approximation argument to reduce the general case $\theta \in L^\infty(0,1)$ to the previous one.

Under the assumption $c \in L^1(0,1)$ and recalling definition (2.1) of W_θ as a fundamental matrix, the state system (3.1)–(3.5) for problem (Q) can be written in differential form:

$$v_1' - c\, v_2 = u + g_1\,, \tag{3.7}$$

$$v_2' + c\, v_1 = z + g_2\,, \tag{3.8}$$

$$v_1(0) = v_2(0) = 0\,, \tag{3.9}$$

$$v_1(1) = v_2(1) = 0\,, \tag{3.10}$$

$$p' - c\,q = 0\,, \tag{3.11}$$

$$q' + c\,p = 0\,, \tag{3.12}$$

$$p(0) = \lambda_1 \cos\theta(0) + \lambda_2 \sin\theta(0)\,, \quad q(0) = -\lambda_1 \sin\theta(0) + \lambda_2 \cos\theta(0)\,, \tag{3.13}$$

$$u = \varepsilon\, p\,, \tag{3.14}$$

$$z'' = -q\,, \tag{3.15}$$

$$z(0) = z(1) = 0\,. \tag{3.16}$$

We shall denote by $v_1(c), v_2(c), \ldots$ the dependence of the solution of (3.7)–(3.16) on $c \in L^1(0,1)$, which is now considered instead of the related dependence on θ. We study its Gâteaux differentiability, and we take variations of the form $c + \delta\, d$ with $d \in L^1(0,1), \delta \in \mathbb{R}$ "small."

The definitions of g_1, g_2, given in (2.2) and (2.3), can also be rewritten in differential form:

$$g_1 = \varepsilon\, l\,, \tag{3.17}$$

$$g_2'' = -h\,, \tag{3.18}$$

$$g_2(0) = g_2(1) = 0\,, \tag{3.19}$$

$$l' - c\,h = -f_1\,, \tag{3.20}$$

$$h' + c\,l = -f_2\,, \tag{3.21}$$

$$l(0) = h(0) = 0\,. \tag{3.22}$$

We have, by (3.20), (3.21),

$$\frac{l(c+\delta\, d)' - l(c)'}{\delta} - (c+\delta\, d)\frac{h(c+\delta\, d) - h(c)}{\delta} = d\,h(c) - \frac{f_1(c+\delta\, d) - f_1(c)}{\delta}\,, \tag{3.23}$$

$$\frac{l(c+\delta\, d)' - h(c)'}{\delta} + (c+\delta\, d)\frac{l(c+\delta\, d) - l(c)}{\delta} = -d\,l(c) - \frac{f_2(c+\delta\, d) - f_2(c)}{\delta}\,. \tag{3.24}$$

We interpret $f_1, f_2 : L^1(0,1) \to L^1(0,1)$ as nonlinear operators, and we assume that they are Gâteaux differentiable. Multiplying (3.23), (3.24) by $[\frac{l(c+\delta\, d)-l(c)}{\delta}, \frac{h(c+\delta\, d)-h(c)}{\delta}]$ and integrating over $[0,t]$, we find that

$$\begin{aligned} \frac{1}{2}\left\| \begin{bmatrix} \dfrac{l(c+\delta\, d) - l(c)}{\delta} \\ \dfrac{h(c+\delta\, d) - h(c)}{\delta} \end{bmatrix}(t) \right\|^2_{\mathbb{R}^2} \\ \leq \int_0^t \left\langle \begin{matrix} d\,h(c) - \dfrac{f_1(c+\delta\, d) - f_1(c)}{\delta} \\ -d\,l(c) - \dfrac{f_2(c+\delta\, d) - f_2(c)}{\delta} \end{matrix} ; \begin{matrix} \dfrac{l(c+\delta\, d) - l(c)}{\delta} \\ \dfrac{h(c+\delta\, d) - h(c)}{\delta} \end{matrix} \right\rangle_{\mathbb{R}^2} ds \end{aligned} \tag{3.25}$$

with obvious notations for the norm and the scalar product in $\mathbb{R}^2$.

The Brezis [6] variant of Gronwall's lemma and (3.25) imply that $\{\frac{l(c+\delta\, d)-l(c)}{\delta}\}$, $\{\frac{h(c+\delta\, d)-h(c)}{\delta}\}$ are bounded in $L^\infty(0,1)$ for $\delta \to 0$. From (3.23), (3.24), we see that the boundedness is even valid in $W^{1,1}(0,1)$, and we also have equi-uniform continuity due to the equi-absolute integrability of $\{\frac{f_i(c+\delta\, d)-f_i(c)}{\delta}\}$, $i = 1,2$. Consequently, by taking a subsequence, we get convergence and the Gâteaux differentiability of $l(c), h(c)$ in $L^2(0,1)$, for instance. Relations (3.17)–(3.19) then show that $g_1(\cdot) : L^1(0,1) \to L^2(0,1), g_2(\cdot) : L^1(0,1) \to W^{2,2}(0,1)$ are also Gâteaux differentiable.

The auxiliary mappings w_1, w_2 defined in (2.15), (2.16) are clearly Gâteaux differentiable. Recalling that $\theta' = c$ and assuming that the perturbation $\tilde{\theta'}_\delta = c + \delta d$ satisfies $\tilde{\theta}_\delta(0) = \theta(0) + \delta\, \eta(0)$, if $\bar{w}_1, \bar{w}_2$ denote the directional derivatives at c in the direction d, we see that

$$\bar{w}_1'' = \left(\eta(0) + \int_0^t d(s)\, ds\right) \cos\left(\theta(0) + \int_0^t c(s)\, ds\right),\ \bar{w}_1(0) = \bar{w}_1(1) = 0\,, \tag{3.26}$$

$$\bar{w}_2'' = \left(\eta(0) + \int_0^t d(s)\, ds\right) \sin\left(\theta(0) + \int_0^t c(s)\, ds\right),\ \bar{w}_2(0) = \bar{w}_2(1) = 0\,. \tag{3.27}$$

Next we recall, by (2.20), that $\lambda_1(c), \lambda_2(c)$ are solutions of an affine system with $\Delta(c) > 0$ and coefficients which are Gâteaux differentiable, by (3.26), (3.27). Then, $\lambda_1(c), \lambda_2(c)$ are as well Gâteaux differentiable from $L^1(0,1)$ into $\mathbb{R}$. Moreover, (3.12), (3.13) imply the Gâteaux differentiability of $p, q : L^1(0,1) \to L^2(0,1)$, for instance. It follows immediately that $u : L^1(0,1) \to L^2(0,1)$ and $z : L^1(0,1) \to W^{2,2}(0,1)$ are Gâteaux differentiable. Finally, applying arguments similar to (3.23)–(3.25) to (3.7)–(3.9), we obtain that $v_1, v_2 : L^1(0,1) \to L^2(0,1)$ are also Gâteaux differentiable.

We denote by $\bar{v}_1, \bar{v}_2, \dots$ the directional derivatives of the mappings defined by (3.7)–(3.16) with respect to $c \in L^1(0,1)$ and in the direction $d \in L^1(0,1)$.

We thus have established the following result.

THEOREM 3.7. *The mappings defined in* (3.7)–(3.16) *are Gâteaux differentiable, and the directional derivatives satisfy the system*

$$\bar{v}_1' - c\,\bar{v}_2 = d\, v_2(c) + \bar{u} + \bar{g}_1\,, \tag{3.28}$$

$$\bar{v}_2' + c\,\bar{v}_1 = -\,d\, v_1(c) + \bar{z} + \bar{g}_2\,, \tag{3.29}$$

$$\bar{v}_1(0) = \bar{v}_2(0) = 0\,, \tag{3.30}$$

$$\bar{v}_1(1) = \bar{v}_2(1) = 0\,, \tag{3.31}$$

$$\bar{p}' - c\,\bar{q} = d\, q(c)\,, \tag{3.32}$$

$$\bar{q}' + c\,\bar{p} = -\,d\, p(c)\,, \tag{3.33}$$

$$\begin{aligned} \bar{p}(0) &= \bar{\lambda}_1 \cos\theta(0) + \bar{\lambda}_2 \sin\theta(0) + \eta(0)\left[\lambda_2 \cos\theta(0) - \lambda_1 \sin\theta(0)\right], \\ \bar{q}(0) &= -\,\bar{\lambda}_1 \sin\theta(0) + \bar{\lambda}_2 \cos\theta(0) - \eta(0)\left[\lambda_1 \cos\theta(0) + \lambda_2 \sin\theta(0)\right], \end{aligned} \tag{3.34}$$

$$\bar{u} = \varepsilon\, \bar{p}\,, \tag{3.35}$$

$$\bar{z}'' = -\,\bar{q}\,, \tag{3.36}$$

$$\bar{z}(0) = \bar{z}(1) = 0\,. \tag{3.37}$$

Remark 3.8. The system (3.28)–(3.37) admits a unique solution, since its homogeneous variant may be reformulated in the language of the control problem (P_ε). Here, homogeneous means that $\bar g_1 = 0, \bar g_2 = 0, d = 0, \eta(0) = 0$, and the corresponding solution of (P_ε) is in this situation clearly identically zero in $[0,1]$. Consequently, the limits defining $\bar v_1, \bar v_2, \dots$ are valid without taking subsequences; we have convergence of the entire sequences.

Next, we introduce the adjoint system associated with (3.28)–(3.37):

$$P_1' - c\,P_2 = 0\,, \tag{3.38}$$

$$P_2' + c\,P_1 = -\,v_2(c)\,, \tag{3.39}$$

$$P_3' - c\,P_4 = R\,, \tag{3.40}$$

$$P_4' + c\,P_3 = Q\,, \tag{3.41}$$

$$Q'' = -\,P_2\,, \tag{3.42}$$

$$R = \varepsilon\,P_1\,, \tag{3.43}$$

$$Q(0) = Q(1) = P_3(0) = P_3(1) = P_4(0) = P_4(1) = 0\,. \tag{3.44}$$

Proposition 3.9. *The system* (3.38)–(3.44) *has a unique solution such that* $P_1, P_2,\ P_3, P_4, R \in W^{1,1}(0,1)$ *and* $Q \in W^{2,\infty}(0,1)$.

Proof. Let $\mu_1, \mu_2 \in \mathbb{R}^2$ be some arbitrary initial conditions for (3.38), (3.39). Then

$$\begin{bmatrix} P_1 \\ P_2 \end{bmatrix}(t) = W_c(t) \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \gamma_1(t) \\ \gamma_2(t) \end{bmatrix},$$

where

$$\begin{bmatrix} \gamma_1(t) \\ \gamma_2(t) \end{bmatrix} = \int_0^t W_c(t)\, W_c^{-1}(s) \begin{bmatrix} 0 \\ -\,v_2(c) \end{bmatrix}(s)\, ds,$$

and $P_1, P_2 \in W^{1,1}(0,1)$ if $c \in L^1(0,1)$. Here, W_c is a new notation for the matrix W that puts into evidence its dependence on c.

Consequently, $R(t) = \varepsilon P_1$ and $Q(t)$ depend in an affine manner on μ_1, μ_2 and belong to $W^{1,1}(0,1)$ and $W^{2,\infty}(0,1)$, respectively. Then,

$$\begin{bmatrix} P_3 \\ P_4 \end{bmatrix}(t) = -\int_t^1 W_c(t) W_c^{-1}(s) \begin{bmatrix} R(s) \\ Q(s) \end{bmatrix} ds$$

belongs to $(W^{1,1}(0,1))^2$. We have used the final null conditions. Notice that the constraint

$$\int_0^1 W_c^{-1}(s) \begin{bmatrix} R(s) \\ Q(s) \end{bmatrix} ds = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{3.45}$$

should be fulfilled to obtain the initial null conditions (3.44) for P_3, P_4. By writing (3.45) explicitly, we obtain a linear system like (2.20) for μ_1, μ_2. Since its determinant is positive, it has a unique solution, and the proof is finished. ∎

Theorem 3.10. *The directional derivative of the cost functional in the problem* (Q) *at the point* $c \in L^1(0,1)$ *and in the direction* $d \in L^1(0,1)$ *is given by*

$$\int_0^1 d\Big(P_1\, v_2(c) - P_2\, v_1(c) + g_1'(c)^* P_1 + g_2'(c)^* P_2 - P_3\, q(c) + P_4\, p(c)\Big)\, ds\,. \tag{3.46}$$

Here, $g_i'(c)$, $i = 1, 2$, denote the Gâteaux derivative of g_i at $c \in L^1(0,1)$, and $g_i'(c)^* : L^2(0,1) \to L^\infty(0,1)$ is the adjoint operator.

Proof. We have (by (3.38), (3.39), partial integration, etc.) that

$$\begin{aligned}
&\lim_{\delta \to 0} \frac{1}{2\delta} \left[\int_0^1 (v_2(c + \delta\, d))^2 \, ds - \int_0^1 (v_2(c))^2 \, ds \right] = \int_0^1 v_2(c)\, \bar v_2 \, ds \\
&= - \int_0^1 (P_2' + c\, P_1) \bar v_2 \, ds - \int_0^1 (P_1' - c\, P_2)\, \bar v_1 \, ds \\
&= \int_0^1 P_1(\bar v_1' - c\, \bar v_2)\, ds + \int_0^1 P_2(\bar v_2' + c\, \bar v_1)\, ds \\
&= \int_0^1 d(P_1\, v_2(c) - P_2\, v_1(c))\, ds + \int_0^1 P_1(\bar u + \bar g_1)\, ds + \int_0^1 P_2(\bar z + \bar g_2)\, ds\,,
\end{aligned}$$

owing to (3.28), (3.29). Now recall that

$$\bar g_1 = g_1'(c)\, d\,, \quad \bar g_2 = g_2'(c)\, d\,.$$

Hence, using (3.40) and (3.41), we can write

$$\begin{aligned}
&\int_0^1 v_2(c)\, \bar v_2 \, ds = \int_0^1 d(P_1\, v_2(c) - P_2\, v_1(c) + g_1'(c)^* P_1 + g_2'(c)^* P_2)\, ds \\
&+ \int_0^1 \varepsilon^{-1}\, R\, \bar u \, ds - \int_0^1 Q''\, \bar z \, ds = \int_0^1 d(\dots)\, ds + \int_0^1 R\, \bar p \, ds + \int_0^1 Q\, \bar q \, ds \\
&= \int_0^1 d(\dots)\, ds + \int_0^1 \bar p (P_3' - c\, P_4)\, ds + \int_0^1 \bar q (P_4' + c\, P_3)\, ds\,.
\end{aligned}$$

From this, again using partial integration together with (3.32), (3.33), we obtain (3.46), and the proof is finished. ∎

Next, we shall study the differentiability properties of (Q) in the general case $\theta \in L^\infty(0,1)$. We consider variations of the form $\theta + \sigma\, \eta, \eta \in L^\infty(0,1), \sigma \in \mathbb{R}$ small. We assume that $f_i : L^\infty(0,1) \to L^2(0,1)$, $i = 1, 2$, depend directly on θ and are Gâteaux differentiable. A direct calculus starting from (2.3) and taking into account the dependence of $W(t)$ on θ leads to

$$\begin{aligned}
\begin{bmatrix} \bar l \\ \bar h \end{bmatrix}(t) = &- \int_0^t W_\theta(t)\, W_\theta^{-1}(s) \begin{bmatrix} \bar f_1(s) \\ \bar f_2(s) \end{bmatrix} ds - \begin{pmatrix} 0 & \eta(t) \\ -\eta(t) & 0 \end{pmatrix} \begin{bmatrix} l(\theta) \\ h(\theta) \end{bmatrix}(t) \\
&+ \int_0^t \begin{pmatrix} 0 & \eta(s) \\ -\eta(s) & 0 \end{pmatrix} W_\theta(t)\, W_\theta^{-1}(s) \begin{bmatrix} f_1(\theta) \\ f_2(\theta) \end{bmatrix}(s)\, ds\,.
\end{aligned} \tag{3.47}$$

By (2.2), it holds that

$$\bar g_1 = \varepsilon\, \bar l\,, \quad -\bar g_2'' = \bar h\,, \quad \bar g_2(0) = \bar g_2(1) = 0\,. \tag{3.48}$$

Comparing (3.47) with (3.20)–(3.22), we see that the integral formulation is more difficult to handle since it involves more products which generate more terms via differentiation.

For the auxiliary mappings w_1, w_2 defined in (2.15), (2.16), we write directly the increment ratios corresponding to θ and $\theta + \sigma\eta$, and we compute the limit corresponding to $\sigma \to 0$ to obtain that

$$\begin{aligned}
&\bar w_1'' = \eta \cos\theta\,, \quad \bar w_2'' = \eta \sin\theta\,, \\
&\bar w_i(0) = \bar w_i(1) = 0\,, \quad i = 1, 2\,.
\end{aligned} \tag{3.49}$$

Relations (3.47)–(3.49) also show the continuous dependence in $L^2(0,1)$ of $\bar g_i, \bar w_i$, $i = 1, 2$, and $\bar l, \bar h$ with respect to regularizations of η and θ, if the same is assumed for $f_i, \bar f_i, i = 1, 2$. For $\bar w_i, i = 1, 2$, and $\bar g_2$, this is valid even in $H^2(0,1)$. An elementary calculus, starting from (2.20), shows that the same continuity property remains valid for $\bar\lambda_1, \bar\lambda_2$.

From relation (3.3), we obtain that

$$\begin{bmatrix} \bar p \\ \bar q \end{bmatrix}(t) = \begin{pmatrix} 0 & \eta(t) \\ -\eta(t) & 0 \end{pmatrix} W_\theta(t) \begin{bmatrix} \lambda_1(\theta) \\ \lambda_2(\theta) \end{bmatrix} + W_\theta(t) \begin{bmatrix} \bar\lambda_1 \\ \bar\lambda_2 \end{bmatrix}, \tag{3.50}$$

with the same continuity property in $(L^2(0,1))^2$ with respect to regularizations of η and θ. By (3.4), (3.5), this property is preserved by $\bar u, \bar z$, and we have

$$\bar u = \varepsilon\, \bar p\,, \quad \bar z'' = -\,\bar q\,, \quad \bar z(0) = \bar z(1) = 0\,. \tag{3.51}$$

Finally, (3.1) gives

$$\begin{aligned} \begin{bmatrix} \bar v_1 \\ \bar v_2 \end{bmatrix}(t) = \int_0^t W_\theta(t)\, W_\theta^{-1}(s) \begin{bmatrix} \bar u(s) + \bar g_1(s) \\ \bar z(s) + \bar g_2(s) \end{bmatrix} ds + \begin{pmatrix} 0 & \eta(t) \\ -\eta(t) & 0 \end{pmatrix} \begin{bmatrix} v_1(\theta) \\ v_2(\theta) \end{bmatrix}(t) \\ - \int_0^t \begin{pmatrix} 0 & \eta(s) \\ -\eta(s) & 0 \end{pmatrix} W_\theta(t)\, W_\theta^{-1}(s) \begin{bmatrix} u(\theta) + g_1(\theta) \\ z(\theta) + g_2(\theta) \end{bmatrix}(s)\, ds \end{aligned} \tag{3.52}$$

with the same conclusion on the continuous dependence on η, θ. Let us now explicitly introduce the regularizations of θ and η,

$$\theta_\delta(t) = \int_R \theta(t - \delta\, y)\, \rho(y)\, dy\,, \quad \eta_\delta\,(t) = \int_R \eta(t - \delta\, y)\, \rho(y)\, dy\,, \tag{3.53}$$

where θ and η are extended by 0 outside the interval $[0,1]$, $\delta > 0$, and where $\rho \in C_0^\infty(\mathbb{R})$ is a Friedrichs mollifier. We also define $d_\delta = \eta_\delta', c_\delta = \theta_\delta'$ which exist in $L^1(0,1)$ but have no good convergence properties for $\delta \to 0$. Then, the systems (3.7)–(3.16), (3.28)–(3.37), and (3.38)–(3.44) can be solved for the data c_δ, d_δ. Let us denote the corresponding solutions with an index or an exponent δ. Then we can prove the following result.

THEOREM 3.11. *The gradient of the cost functional of the problem* (Q) *at the point* $\theta \in L^\infty(0,1)$ *and in the direction* $\eta \in L^\infty(0,1)$ *is given by*

$$\begin{aligned} \int_0^1 v_2(\theta)\, \bar v_2\, ds = \int_0^1 \eta \Big[g_1'(\theta)^* P_1 + g_2'(\theta)^* P_2 - v_1(\theta)\, v_2(\theta) \\ - P_1(\theta)\, (z(\theta) + g_2(\theta)) + P_2(\theta)\, (u(\theta) + g_1(\theta)) + q(\theta)\, R(\theta) - p(\theta)\, Q(\theta) \Big]\, ds\,. \end{aligned} \tag{3.54}$$

Here, $v_1(\theta), v_2(\theta), u(\theta), z(\theta), p(\theta), q(\theta)$ are obtained by (3.1)–(3.5) with $g_1(\theta), g_2(\theta)$ given by (2.2), (2.3), and P_1, P_2, P_3, P_4, R, Q are computed via (3.38)–(3.44) rewritten in integral form (which is obvious).

Proof. By (3.52), (3.53), we can write

$$\int_0^1 v_2(\theta)\, \bar v_2\, ds = \lim_{\delta \to 0} \int_0^1 v_2^\delta\, \bar v_2^\delta\, ds\,. \tag{3.55}$$

From Theorem 3.10, we obtain that

$$\int_0^1 v_2^\delta\, \bar v_2^\delta\, ds = \int_0^1 d_\delta \left(P_1^\delta\, v_2^\delta - P_2^\delta\, v_1^\delta - P_3^\delta\, q^\delta + P_4^\delta\, p^\delta + P_1^\delta\, \bar g_1^\delta + P_2^\delta\, \bar g_2^\delta \right) ds.$$

Using the boundary conditions and the differentiability properties, we first compute

$$\begin{aligned}
&\int_0^1 d_\delta \left(P_1^\delta\, v_2^\delta \,-\, P_2^\delta\, v_1^\delta \,-\, P_3^\delta\, q^\delta \,+\, P_4^\delta\, p^\delta\right) ds \\
&= -\int_0^1 \eta_\delta \left((P_1^\delta)'\, v_2^\delta \,+\, P_1^\delta\, (v_2^\delta)' \,+\, \cdots \,+\, (P_4^\delta)'\, p^\delta \,+\, P_4^\delta\, (p^\delta)'\right) ds \\
&= -\int_0^1 \eta_\delta \left(v_1^\delta\, v_2^\delta \,+\, P_1^\delta\, (z^\delta \,+\, g_2^\delta) \,-\, P_2^\delta (u^\delta \,+\, g_1^\delta) \,-\, q^\delta\, R^\delta \,+\, p^\delta\, Q^\delta\right) ds.
\end{aligned} \tag{3.56}$$

We indicate only a partial calculation on how the last equality in (3.56) is established:

$$\begin{aligned}
&(P_4^\delta)'\, p^\delta \,+\, P_4^\delta\, (p^\delta)' \,-\, (P_3^\delta)'\, q^\delta \,-\, P_3^\delta\, (q^\delta)' \\
&= (P_4^\delta)'\, p^\delta \,+\, P_4^\delta\, c_\delta\, q^\delta \,-\, (P_3^\delta)'\, q^\delta \,+\, P_3^\delta\, c_\delta\, p^\delta \\
&= q^\delta(-\,(P_3^\delta)' \,+\, c_\delta\, P_4^\delta) \,+\, p^\delta((P_4^\delta)' \,+\, c_\delta\, P_3^\delta) \\
&= -\,q^\delta\, R^\delta \,+\, p^\delta\, Q^\delta
\end{aligned}$$

by (3.11), (3.12) and (3.40), (3.41).

We also consider the term

$$\begin{aligned}
&\int_0^1 \left(P_1^\delta\, \bar{g}_1^\delta \,+\, P_2^\delta\, \bar{g}_2^\delta\right) ds \,=\, \int_0^1 \left(P_1^\delta\, g_1'(\theta_\delta)\, \eta_\delta \,+\, P_2^\delta\, g_2'(\theta_\delta)\eta_\delta\right) ds \\
&= \int_0^1 \eta_\delta \left[(g_1^\delta)'(\theta_\delta)^* P_1^\delta \,+\, (g_2^\delta)'(\theta_\delta)^* P_2^\delta\right] ds.
\end{aligned} \tag{3.57}$$

The derivatives of g_1, g_2 may be taken directly with respect to θ. This can be clearly seen from (3.23)–(3.25), where f_i may depend on θ, without modifying the argument.

We combine (3.55)–(3.57), and we pass to the limit as $\delta \to 0$. The continuity properties with respect to both η_δ and θ_δ have been explained in (3.47)–(3.52). We remark that the continuous dependence on $\delta \to 0$ is valid for $P_1^\delta, P_2^\delta, P_3^\delta, P_4^\delta, R^\delta, Q^\delta$ since the system (3.38)–(3.44) can be put into integral (mild) form as well. □

Remark 3.12. The gradient provided by Theorem 3.11 will be used in section 4 in the computation of numerical examples of shape optimization. It is also possible to write the first order optimality conditions for problem (Q) by requiring (3.54) to be positive in the admissible directions of variation.

4. Numerical experiments. We have computed several examples of arches, including their shape optimization, using the methods developed in this paper. Numerical examples concerning plates and beams have been reported in the works of Arnăutu, Langmach, Sprekels, and Tiba [2] and Sprekels and Tiba [23], where different (but related) approaches were used.

In Figures 1–4, deformations of various arches (Roman, gothic, closed) with different thicknesses $\varepsilon > 0$ and under certain square integrable loads $[f_1, f_2]$ are shown. The algorithm is based on Theorem 2.6 with explicit solutions of (2.20) obtained via MAPLE. The integrals appearing in the coefficients of (2.20) and elsewhere can be computed explicitly in the case of simple arches and simple forces (purely tangential or purely normal, etc.). Otherwise, standard numerical integration procedures on the real line should be applied.

The parametric representation of an arch associated to some function θ on a prescribed interval is given by $[\varphi_1, \varphi_2]$ with $\varphi_1' = \cos\theta, \varphi_2' = \sin\theta$, and with null

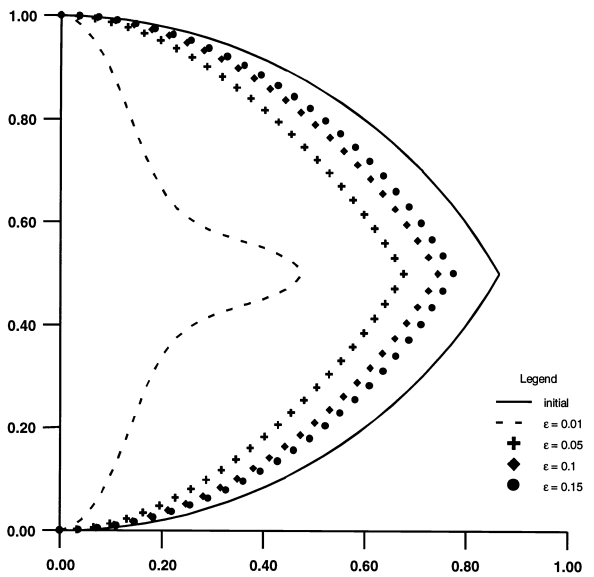


FIG. 1. $\theta(t) = t,\ t \in [0, \pi/3],\ \theta(t) = t + \pi/3,\ t \in [\pi/3, 2\pi/3],\ f_1(t) = 0,\ f_2(t) = 1(SE),\ E = 10.$

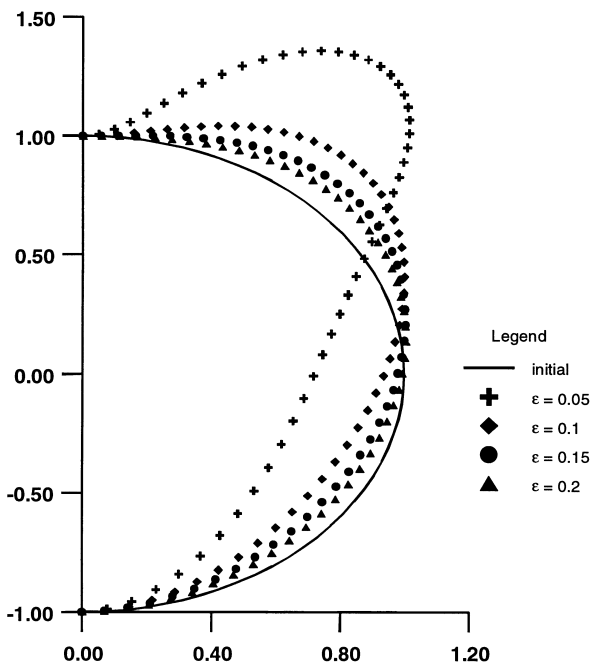


FIG. 2. $\theta(t) = t,\ t \in [0, \pi],\ f_1(t) = \sin(t)/S,\ f_2(t) = \cos(t)/S.$

initial conditions. Notice that in Figure 1, θ is discontinuous and $\varphi = [\varphi_1\,, \varphi_2]$ is just Lipschitz, which shows the importance of relaxing the regularity assumptions in (1.1) as is done in problem (P_ε) in section 2. Figures 2 and 4 show the same type of

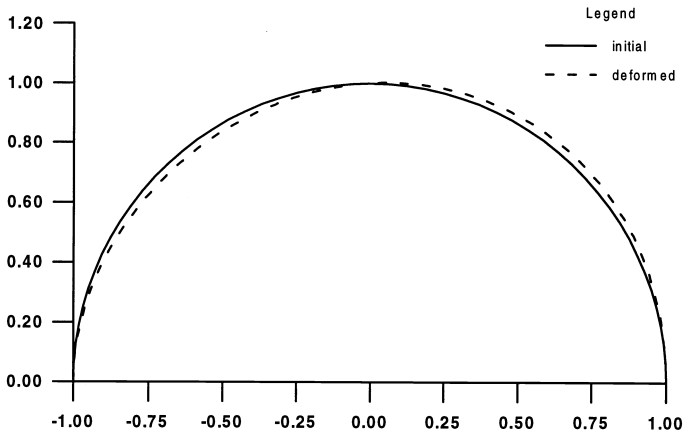


FIG. 3. $\theta(t) = t,\ f_1(t) = \sin(t),\ f_2(t) = 2\cos(t),\ t \in [0, \pi]$.

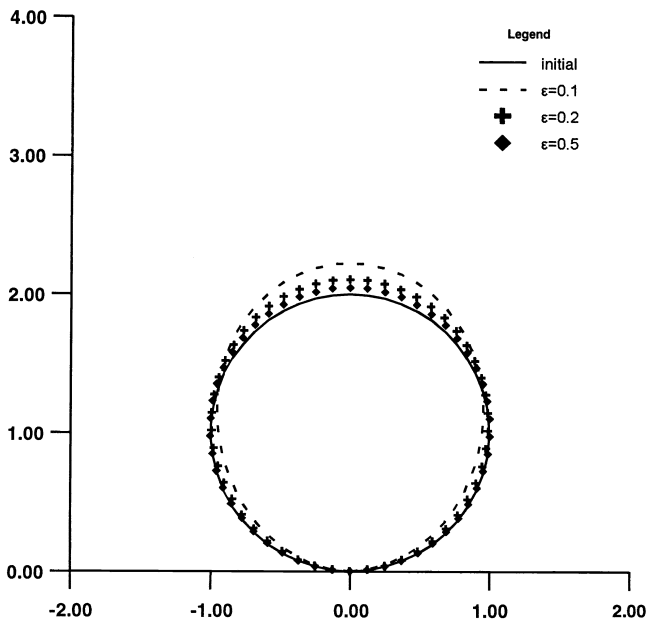


FIG. 4. $\theta(t) = t,\ f_1(t) = \sin(t)/(SE),\ f_2(t) = \cos(t)/(SE),\ t \in [0, 2\pi],\ E = 100$.

arch with similar loading. The difference in the shape of the obtained deformations is due to the fact that the first arch is clamped at both ends, while the closed arch is clamped only in the point $(0, 0)$. Figure 3 refers to the "flexural" model briefly explained in Theorem 2.11 and Remark 2.8. The constant E is the Young modulus of the material, while $S = \varepsilon^{3/2}$ gives the influence of the thickness $\varepsilon > 0$. We indicate, as a short example, the explicit form of the deformation $[v_1, v_2]$ corresponding to the situation described in Figure 2:

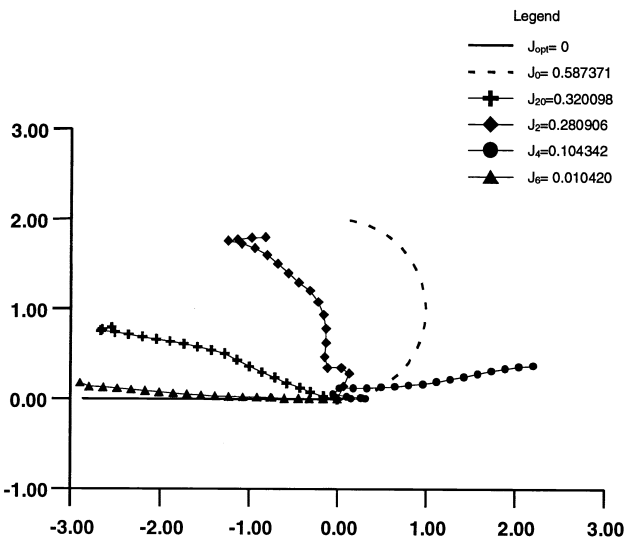


FIG. 5. $\theta(t) \in [0,\pi],\ f_1(t) = 1/S,\ f_2(t) = 0,\ \theta_0(t) = t,\ t \in [0,\pi]$.

$$v_1(t) = (6\,\varepsilon \sin t + 4 \sin t + 2\,\pi\,\varepsilon \sin t + \pi\,\varepsilon^2\, t \sin t - 4\,\varepsilon\, t \cos t - 2\,\varepsilon^2 \sin t$$
$$- 2\,\varepsilon\, t^2 \sin t - \varepsilon^2\, t^2 \sin t + \pi\, t \sin t - 4\, t \cos t - t^2 \sin t - 2\,\pi - 2\,\varepsilon\,\pi$$
$$+ 2\,\pi \cos t + 2\,\pi\,\varepsilon \cos t)/4\,\varepsilon^{3/2}(\varepsilon + 1)\,,$$
$$v_2(t) = (\varepsilon + 1)\,(2\, t \sin t + \pi\, t \cos t - \pi \sin t - t^2 \cos t)/4\,\varepsilon^{3/2}\,.$$

Figures 5–9 and Tables 1 and 2 concern optimization procedures for arches, according to the theory developed in section 3. For the computation of the gradient of the cost functional, as given in (3.54), it is necessary to obtain the numerical solution of the state system (3.1)–(3.5), of the adjoint system (3.38)–(3.44), and the approximation of the mappings $[g_1'(\theta)]^* P_1$ and $[g_2'(\theta)]^* P_2$. It is obvious that by the nature of the data an explicit calculation is not possible in the optimization routine.

We have considered an equidistant division of the interval of definition, denoted here by $[0, L]$, into N_0 (a natural number) subintervals $[t_i, t_{i+1}]$, with $t_i = i\,h, h = \frac{L}{N_0}$. The mapping $\theta \in L^\infty(0, L)$ is approximated, in different examples, by piecewise linear splines or by piecewise constant functions. The integrals are computed accordingly by standard quadrature formulas, and the solution of the ordinary differential system is obtained via linear finite elements. The scalars $\lambda_1^\varepsilon, \lambda_2^\varepsilon$ from (3.3) are found from the algebraic system (2.20). Similarly, the unknown initial conditions μ_1, μ_2 for (3.38), (3.39) satisfy a system of the same type as (2.20) with the mappings l, g_2 replaced by γ_1, γ with $\gamma'' = -\gamma_2, \gamma(0) = \gamma(L) = 0$ (see Proposition 3.9 and its proof). The functions $[g_1'(\theta)]^* P_1$ and $[g_2'(\theta)]^* P_2$ have been approximated in the following way:

$$[g_k'(\theta)]^* P_k(t_i) \simeq \frac{1}{h} \int_{t_i}^{t_{i+1}} [g_k'(\theta)]^*\, P_k(s)\, ds$$
$$= \frac{1}{h} \int_0^L P_k(s)\, \left(\bar{g}_k\, \chi_{[t_i, t_{i+1}]}\right)(s)\, ds\,, \quad k = 1, 2\,, \quad i = \overline{0, N-1}\,,$$
$$[g_k'(\theta)]^* P_k(L) \simeq 0\,, \quad k = 1, 2\,.$$

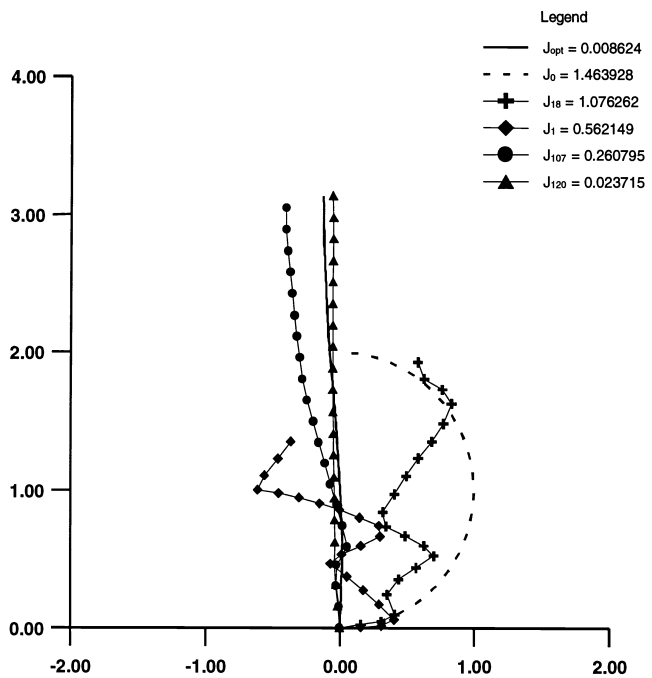


FIG. 6. $\theta(t) \in [0, \pi]$, $f_1(t) = \sin(\theta(t))/S$, $f_2(t) = \cos(\theta(t))/S$, $\theta_0(t) = t$, $t \in [0, \pi]$.

For the determination of $\bar{g}_k$ the relation (3.48) is used, and $\chi_{[t_i, t_{i+1}]}$ is the characteristic function of $[t_i, t_{i+1}]$.

Although the studied optimization problems are nonconvex, adaptations of Rosen's and Uzawa's gradient algorithms with projection (Gruver and Sachs [15], Arnăutu [1]), have been used. A maximal number of iterations (between 200 and 300) has been prescribed, and the solution has been chosen as the one which gives the best value of the cost. The algorithm stops as well if the value of the gradient or of the cost is zero.

For a given example, several tests have been performed with various values of the parameters N_0, α (the parameter from the Rosen algorithm) and with both algorithms. In general, the Rosen algorithm gives better results than the Uzawa algorithm. In the optimization problems, we have fixed $\varepsilon = 0.1$. A typical line search procedure is to subdivide the open-closed interval $]0, 1]$ into N_1 equal parts and to give the line search parameter the values $\frac{i}{N_1}, i = \overline{1, N_1}$. The one which gives the best cost will generate the next iteration. We have avoided, with good numerical results, the usual computation of the line search parameter by a one-dimensional optimization problem, which may be very time-consuming. The procedure used combines in an ad hoc manner the gradient algorithm principle and a global search. A projection on the admissible set has been performed in each iteration. The optimization problem (Q) looks for the shape of the arch which ensures the minimal normal deformation (in some integral sense) under the action of a prescribed force. We have examined purely tangential ($f_2 = 0$) or normal ($f_1 = 0$) forces (since they give the basis in the local system of axes), as well as forces not depending on the unknown arch. This last case is described in the local system of coordinates by $f_1(t) = \sin(\theta(t))/S$ and $f_2(t) = \cos(\theta(t))/S$ (for the force of modulus one and parallel to the vertical axis),

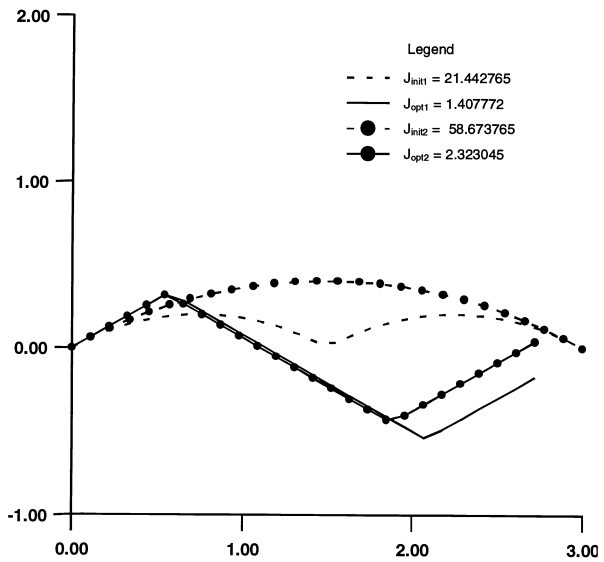


FIG. 7. $\theta(t) \in [\pi/3, 2\pi/3]$, $f_1(t) = \cos(\theta(t))/S$, $f_2(t) = \sin(\theta(t))/S$, $t \in [0, \pi]$, $\theta_{01}(t) = (2t + \pi)/3$, $t \in [0, \pi/2)$, $\theta_{01}(t) = 2t/3$, $t \in [\pi/2, \pi]$, $\theta_{02}(t) = (t + \pi)/3$, $t \in [0, \pi]$.

and in converse order for forces parallel to the horizontal axis. It should be noticed that the force is independent of the arch, but its local representation is dependent via θ.

The constraints for θ were given by subintervals of $[0, \pi]$ as indicated in the figures. This suffices for many applications and avoids the self-intersection of arches. However, some degenerate case is still possible, according to Figure 9.

In Figure 5, under the action of a tangential force, and starting with the initial iteration given by the Roman arch, it is seen that the global solution is the beam, which clearly has no normal deflection under such a load. In our representation, two global solutions (beams) are put into evidence, associated to $\theta = 0$ and to $\theta = \pi$. The figure shows some iterations produced by the algorithm and the corresponding values of the cost. In this experiment, we have used $N_0 = 200, n_1 = 10, \alpha = 0,75$, and the arch close to the beam was obtained in iteration $I = 24$. We underline that in this example, an infinity of global solutions (beams of any slope) exists, and this shows the difficulty of the numerical computations.

In Figure 6, the initial iteration is again the Roman arch, but the force is of constant modulus one and parallel to the vertical axis. The iterations that are represented show how the routine again finds the (unique if θ is constrained in $[0, \pi]$) global solution which is given by a vertical beam characterized by $\theta = \frac{\pi}{2}$. In this configuration, the prescribed force becomes purely tangential to the arch, and the global solution is a special case of the previous example (but not the problem as a whole). We have used $N_0 = 200, N_1 = 10, \alpha = 1$, and the global optimum was obtained at iteration $I = 139$.

The numerical results from Figures 5 and 6 match perfectly with the physical interpretation. This gives a strong validation of the notion of weak solutions that we are using and shows the stability of our methods.

In Figures 8 and 9, the case of a purely normal load is discussed, the difference

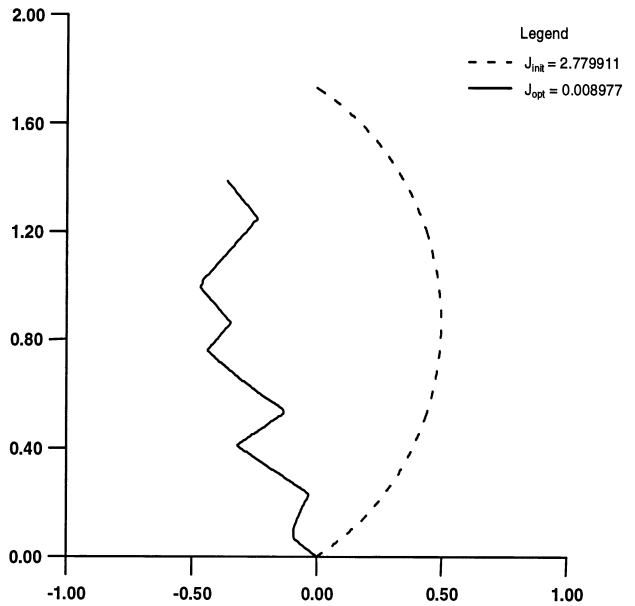


FIG. 8. $\theta(t) \in [\pi/6, 5\pi/6]$, $f_1(t) = 0$, $f_2(t) = 1/S$, $\theta_0(t) = t + \pi/3$, $t \in [0, 2\pi/3]$, $J_{init} = 2.779911$, $J_{opt} = 0.008977$.

being given by the constraints imposed on $\theta : \left[\frac{\pi}{6}, \frac{5\pi}{6}\right]$, respectively $[0, \pi]$. In Figure 9, the "optimal" found θ is represented, not the arch as usual. As the solution is bang-bang, $\theta \in \{0, \pi\}$ a.e. $t \in [0, \pi]$; then the arch degenerates and cannot be graphically represented. Suggested by the bang-bang structure of the obtained solution (the computations were made with $N_0 = 200, N_1 = 20, \alpha = 1, 5, I = 27$), we have simply generated a sequence θ^N, by giving to the new parameter N the values listed in Table 2 and to θ^N the values 0 and π, alternatively on subsequent subintervals. We have directly computed the costs $J(\theta^N)$ associated with such oscillating arches and listed them in Table 2. The conclusion is that the sequence θ^N is a very efficient minimizing sequence for this problem, ensuring for $N \geq 50$ lower values of the cost than the one computed by the complete numerical procedure (although this provides a satisfactory result as well). We stress that the oscillatory nature of the minimizing sequence $\{\theta^N\}$ is related to the noncompactness of the constraint set $\{\theta \in L^\infty(\Omega)\,;\, \theta(t) \in [0, \pi]$ for almost every $t \in (0, \pi)\}$ in $L^\infty(0, \pi)$. This set is only bounded and closed, which is not enough to ensure the existence of the optimal θ as discussed in Theorem 3.2 and Corollary 3.3. This numerical example can be interpreted as showing that the assumptions of Corollary 3.3 are sharp. We also underline that such compactness comments apply to Figures 5 and 6 as well, although global minimum points exist in these examples.

Figure 8 represents the initial (Roman) arch and the obtained solution, in the same problem as in Figure 9, with the constraints given by the set $[\frac{\pi}{6}, \frac{5\pi}{6}]$ in order to avoid degeneracy. The numerical test used $N_0 = 300, N_1 = 10, \alpha = 1, 5$, and the obtained optimum corresponded to the iteration $I = 160$. The bang-bang structure of the solution is again clear (recall that θ is the angle between the tangent to the arch and the horizontal axis). However, Table 1 shows that the simple sequence $\{\theta^N\}$, constructed as in the previous example but with the values $\pi/6, 5\pi/6$, is no

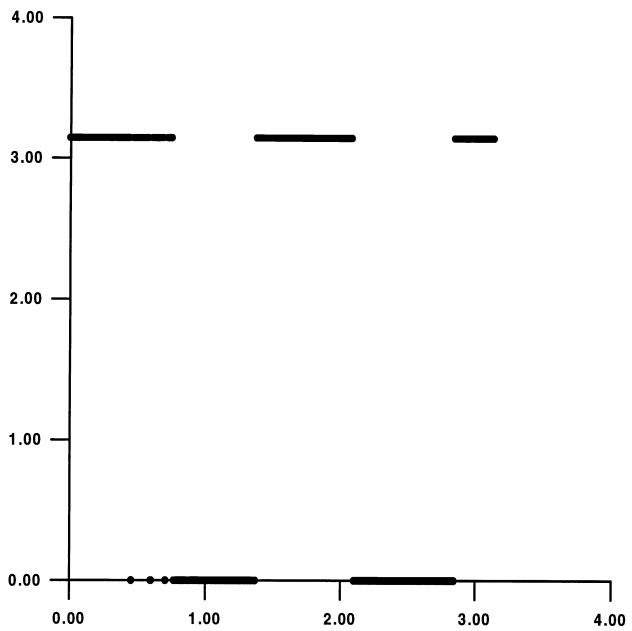


FIG. 9. $\theta(t) \in [0,\pi],\ f_1(t) = 0,\ f_2(t) = 1/S,\ \theta_0(t) = t,\ t \in [0,\pi],\ J_{init} = 82.922993,\ J_{opt} = 0.0024772.$

TABLE 1
$\theta(t) \in \{\pi/6, 5\pi/6\}, f_1(t) = 0,\ f_2(t) = 1/S, t \in [0, 2\pi/3].$

N	$J(\theta^N)$
30	0.0141367792
50	0.0247750769
100	0.0303698330
200	0.0318697376
300	0.0321519172
500	0.0322969269
800	0.0323467156
1000	0.0323582113

longer a minimizing sequence for this problem. The commuting points for the bang-bang solution are not equidistant in this example. Finally, in Figure 7, a "realistic" example is studied: the construction of a most resistant roof subject to a vertical constant load of modulus one. The reader should note that in this figure we have interchanged the axes to make the representation look more "physical." To perform a more precise calculation, we have fixed $N_0 = 500, N_1 = 100, \alpha = 10$. Two experiments are reported in Figure 7, one with the initial iteration given by a fragment of Roman arch, and another with the initial iteration given by two coupled fragments of Roman arch. In both cases, the numerical solutions were obtained in the first iteration, $I = 1$, and are very similar. In this example (as in Figure 8), the theoretical optimal value is "far" from zero, and the computed values are very good.

We close this presentation by underlining that working with low regularity assumptions was essential for the optimization applications in view of the bang-bang structure of the optimal θ, as found in many examples. However, in Figure 6 the

TABLE 2
$\theta(t) \in \{0, \pi\}, f_1(t) = 0,\ f_2(t) = 1/S, t \in [0, \pi]$.

N	$J(\theta^N)$
30	0.0095834975
50	0.0012420426
100	0.0000776279
200	0.0000048517
300	0.0000009584
500	0.0000001242
800	0.0000000190
1000	0.0000000078

global solution is not bang-bang, and this property seems to be related just to the applied force. That is why we did not study bang-bang properties in section 3, although such properties are known for plates, according to Sprekels and Tiba [19, 20]. We also underline the nonlocal optimization character of our numerical experiments as this is obvious from the reported results.

REFERENCES

[1] V. ARNĂUTU, *Numerical methods for variational problems*, Lecture Notes 8, Department of Mathematics, University of Jyväskylä, Finland, 2001.

[2] V. ARNĂUTU, H. LANGMACH, J. SPREKELS, AND D. TIBA, *On the approximation and the optimization of plates*, Numer. Funct. Anal. Optim., 21 (2000), pp. 337–354.

[3] V. BARBU, *Analysis and Control of Nonlinear Infinite Dimensional Systems*, Academic Press, Boston, 1993.

[4] PH. BÉNILAN, *Solutions intégrales d'équations d'évolution dans un espace de Banach*, C. R. Acad. Sci. Paris Sér. A-B, 274 (1972), pp. A47–50.

[5] A. BLOUZA AND H. LE DRET, *Existence et unicité pour le modèle de Koiter pour une coque peu régulière*, C. R. Acad. Sci. Paris Sér. A-B, 319 (1994), pp. 1127–1132.

[6] H. BREZIS, *Analyse fonctionelle, théorie et applications*, Masson, Paris, 1983.

[7] D. CHAPELLE, *A locking-free approximation of curved rods by straight beam elements*, Numer. Math., 77 (1997), pp. 299–322.

[8] D. CHENAIS AND J. C. PAUMIER, *On the locking phenomenon for a class of elliptic problems*, Numer. Math., 67 (1994), pp. 427–440.

[9] D. CHENAIS AND B. ROUSSELET, *Dependence of the buckling load of a nonshallow arch with respect to the shape of its midcurve*, RAIRO Modél. Math. Anal. Numér., 24 (1990), pp. 307–341.

[10] D. CHENAIS, B. ROUSSELET, AND R. BENEDICT, *Design sensitivity for arch structures with respect to midsurface shape under static loading*, J. Optim. Theory Appl., 58 (1988), pp. 225–239.

[11] PH. CIARLET, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[12] PH. CIARLET, *Introduction to Linear Shell Theory*, Gauthier–Villars, Paris, 1998.

[13] G. GEYMONAT AND G. GILARDI, *Contre-examples à l'inégalité de Korn et au lemme de Lions dans les domaines irréguliers*, in Equations aux dérivées partielles et applications. Articles dédiés à J. L. Lions, Gauthier–Villars, Paris, 1998, pp. 541–548.

[14] G. GEYMONAT AND E. SANCHEZ-PALENCIA, *On the rigidity of certain surfaces with folds and applications to shell theory*, Arch. Ration. Mech. Anal., 129 (1995), pp. 11–45.

[15] W. A. GRUVER AND E. SACHS, *Algorithmic Methods in Optimal Control*, Res. Notes Math. 47, Pitman, Boston, London, 1981.

[16] L. S. PONTRYAGIN, *Gewöhnliche Differentialgleichungen*, VEB Deutscher Verlag der Wissenschaften, Berlin, 1965.

[17] B. ROUSSELET, J. PIEKARSKI, AND A. MYSLINSKI, *Design sensitivity for a hyperelastic rod in large displacements with respect to its midcurve shape*, J. Optim. Theory Appl., 96 (1998), pp. 683–708.

[18] J. Sprekels and D. Tiba, *On the approximation and optimization of fourth order elliptic systems*, in Optimal Control of Partial Differential Equations, Internat. Ser. Numer. Math. 133, Birkhäuser-Verlag, Basel, 1999, pp. 277–286.

[19] J. Sprekels and D. Tiba, *Propriétés de bang-bang généralisées dans l'optimisation des plaques*, C. R. Acad. Sci. Paris Sér. I Math., 327 (1998), pp. 705–710.

[20] J. Sprekels and D. Tiba, *A duality approach in the optimization of beams and plates*, SIAM J. Control Optim., 37 (1998–1999), pp. 486–501.

[21] J. Sprekels and D. Tiba, *A duality-type method for the design of beams*, Adv. Math. Sci. Appl., 9 (1999), pp. 84–102.

[22] J. Sprekels and D. Tiba, *Sur les arches lipschitziennes*, C.R. Acad. Sci. Paris Sér. I Math., 331 (2000), pp. 179–184.

[23] J. Sprekels and D. Tiba, *Optimization of Clamped Plates with Discontinuous Thickness*, submitted.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 4, pp. 1134–1158

SENSITIVITY TO INFINITESIMAL DELAYS IN NEUTRAL EQUATIONS*

W. MICHIELS†, K. ENGELBORGHS†, D. ROOSE†, AND D. DOCHAIN‡

Abstract. The stability of a steady state solution of a neutral functional differential equation can be sensitive to infinitesimal changes in the delays. This phenomenon is caused by the behavior of the essential spectrum and is determined by the roots of an exponential polynomial. Avellar and Hale [*J. Math. Anal. Appl.*, 73 (1980), pp. 434–452] have considered the case of multiple fixed and nonzero delays. In the first part of this paper their results are illustrated by means of spectral plots. In the second part we extend the theory of Avellar and Hale to the limit case whereby some of the delays are brought to zero, which may lead to characteristic roots with arbitrarily large real part. Necessary and sufficient conditions are provided. Using these results we show that the ratio of the delays plays a crucial role when several delays tend to zero simultaneously. As an illustration of the theory, we analyze the robustness of a boundary controlled PDE in the presence of a small feedback delay.

Key words. neutral equation, sensitivity, characteristic roots

AMS subject classifications. 34K40, 34K35

PII. S0363012999355071

1. Introduction.

In this paper we study the behavior of the roots of the exponential polynomial

$$H(\lambda) \triangleq 1 - \sum_{j=1}^{N} a_j e^{-\lambda \tau_j}, \quad \tau_j \in \mathbb{R}_0^+, \quad a_j \in \mathbb{R}, \quad j = 1, \ldots, N, \tag{1.1}$$

in the complex plane. $H(\lambda) = 0$ is the characteristic equation of the functional difference equation $x(t) = \sum_{j=1}^{N} a_j x(t - \tau_j)$, which determines the essential spectrum of the solution operator of the neutral functional differential equation (NFDE),

$$\frac{d}{dt}\left(x(t) - \sum_{j=1}^{N} a_j x(t - \tau_j) \right) = b_0 x(t) - \sum_{j=1}^{N} b_j x(t - \tau_j). \tag{1.2}$$

It is well known that the spectrum of (1.2) determines the stability of its zero solution, since equations of this form satisfy the spectrum determined growth condition; see [8, Corollary IX.3.1]. However, the stability of the zero solution may be sensitive to arbitrarily small changes in the delays τ_j. Since this sensitivity is caused by the essential spectrum, which is determined by (1.1) (see [9]), we perform a detailed study

*Received by the editors April 19, 1999; accepted for publication (in revised form) March 22, 2001; published electronically November 28, 2001. This paper presents research results of the Belgian programme on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture (IUAP P4/02). The scientific responsibility rests with its authors.

http://www.siam.org/journals/sicon/40-4/35507.html

†Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan 200A, B-3001 Leuven, Belgium (Wim.Michiels@cs.kuleuven.ac.be, Koen.Engelborghs@cs.kuleuven.ac.be, Dirk.Roose@cs.kuleuven.ac.be). The second author is a Postdoctoral Fellow of the Fund for Scientific Research–Flanders (Belgium).

‡Université Catholique de Louvain, Bâtiment Euler, 4 Av. Georges Lemaître, B-1348 Louvain-la-Neuve, Belgium (dochain@auto.ucl.ac.be).

of (1.1). In terms of the characteristic roots of (1.1), this sensitivity is caused by the occurrence of so-called infinite root chains, sequences of roots whose imaginary parts grow unbounded, yet whose real parts have a finite limit. It was shown by Avellar and Hale [1] that, as a consequence, the smallest upper bound $c = \sup\{\Re(\lambda) : H(\lambda) = 0\}$ is not continuous w.r.t. the delays τ_j. Hence it is possible that arbitrarily small changes in the delays destabilize the NFDE system. This is of importance in control problems, since such critical delay changes can be caused by a small delay in the application of the control action.

NFDEs arise, for example, in models of distributed networks [13, 12], combustion [20], and the control of structures through delayed forcing depending on the acceleration [2].

The lack of robustness w.r.t. small changes in the delays is also observed for boundary controlled (hyperbolic) PDEs [4, 5, 6, 7, 9, 10, 15, 19, 21] and feedback controlled descriptor systems [14]: a small delay in the application of the control action, which is inevitable in practice due to, e.g., measurements or AD-DA conversion, can lead to instability of the stable undelayed system. Hence it is very important to include all possible delays in the model.

In the literature, basically two approaches are used to analyze such problems. A first approach, which will be followed in this paper, studies directly the influence of delay perturbations on the characteristic roots of equations of the form (1.1). In [1] Avellar and Hale consider (1.1) and show that the rational (in)dependency of the delays plays a crucial role for the robustness properties of the given system. However, in [1] only perturbations of fixed, nonzero delays are considered. These results do not safely apply to the case discussed in this paper, where zero delays are perturbed to (small nonzero) delays. In the latter case, infinite root chains occur, whose real part may tend to $+\infty$, a phenomenon which does not occur in the presence of perturbations of (only) nonzero delays. As in [1], we allow a general dependency structure on the delay perturbations and derive sufficient and necessary conditions for the occurrence of characteristic roots with large real part for *vanishing* delays. Furthermore, we discuss the occurrence of roots with large positive real but small imaginary part. We prove that the occurrence of these phenomena is determined by what we call the "small delay" part of the characteristic equation (except for some degenerate cases).

On the other hand, for the analysis of a small time-delay in feedback systems, another approach can be followed. In [15, 16] Logemann and coworkers rewrite the closed loop system as an input-output mapping $H(s)$ with (delayed) unity feedback $e^{-\epsilon s}$, where ϵ represents a small feedback delay, and formulate conditions for robustness and nonrobustness of stability against the small time-delay on the open loop transfer function $H(s)$. They show that, depending on the properties of $H(s)$, a small time-delay may not only result in instability but may also cause characteristic roots with arbitrarily large imaginary and real part. While [16] assumes regularity of $H(s)$, some extensions are made in [15] to the case where $H(s)$ is non well-posed. In [17] Logemann and Townley use this approach to show that when a NFDE

$$\frac{d}{dt}\mathcal{D}x_t = \mathcal{L}x_t + Bu(t), \qquad \mathcal{D}x_t = x(t) - \sum_i D_i x(t - h_i), \tag{1.3}$$

with control input $u(t)$, has an exponentially unstable difference operator $\mathcal{D}$, any stabilizing state feedback law is not robust against a small perturbation of the feedback delays. Since the essential spectrum is determined by the difference equation inside the differentiation operator, such a feedback law should include velocity feedback

$u = -\frac{d}{dt}(\sum_j F_j x(t-k_j))$, $k_j > 0$, which leads to a closed loop system with the essential spectrum determined by

$$\det\left(I - \sum_i D_i e^{-\lambda h_i} + \sum_j F_j e^{-\lambda(k_j+\epsilon)}\right) = 0, \tag{1.4}$$

where ϵ represents a small perturbation in the feedback delays. Since there is only one perturbation ϵ on the (fixed and nonzero) delays h_i, k_j in (1.4), the analysis of this robustness problem can be recast in the framework of [16]. That procedure is not possible for the broader class of delay perturbations considered in this paper.

Note that for some problems both approaches described above can be used. For instance, in section 6 we apply the theory developed throughout this paper to a boundary controlled wave equation, which was also analyzed in [16].

In section 2 we repeat the main results of [1] and introduce necessary notation. In section 3 we visualize and interpret these results by means of plots of the characteristic roots, thereby explaining the nature of the instability mechanisms. We explain that the sensitivity of c to arbitrary small changes of the delays is necessarily caused by roots with large imaginary part. In section 4 we demonstrate with a simple example that the theory of [1, 17] is not sufficient to deal with vanishing delays. We prove under which conditions characteristic roots with large real part can occur and give a thorough discussion of these results. We conclude in section 6 with two illustrative examples.

2. Analysis with fixed delays. In this section we briefly describe the main results of [1], where the case of fixed and nonzero delays is considered.

2.1. Definitions and notation. Throughout this section we consider the roots of exponential polynomials of the form

$$H(\lambda) \triangleq 1 - \sum_{j=1}^{N} a_j e^{-\lambda\tau_j}, \tag{2.1}$$

where we assume that the delays τ_j are fixed and satisfy $0 < \tau_1 < \tau_2 < \cdots < \tau_N$.

Define the collection of the real parts of all the roots of (2.1) as Z,

$$Z = \{\Re(\lambda) : H(\lambda) = 0\},$$

and denote its closure by $\bar{Z}$. The smallest upper bound of $\bar{Z}$, which is important for stability considerations, is

$$c = \sup\{\Re(\lambda) : H(\lambda) = 0\}.$$

Assume that the N delays τ_j, $j = 1, \ldots, N$, depend on $M \leq N$ so-called independent delays $r_1, \ldots, r_M$:

$$\tau_j = \sum_{k=1}^{M} \gamma_{j,k} r_k = \gamma_j \cdot r, \tag{2.2}$$

whereby $\gamma_j = (\gamma_{j,1}, \ldots, \gamma_{j,M}) \in \mathbb{N}^M$ are nonzero vectors with nonnegative integer coefficients and $r \in (0, \infty)^M$. Dependency of the kind (2.2) often appears in difference

equations arising from practical applications, as, for example, in (delayed) boundary controlled wave equations (see section 6). The same holds when dealing with vector valued difference equations. Indeed, the characteristic equation of

$$x(t) = \sum_{k=1}^{M} A_k x(t - r_k), \quad x(t) \in \mathbb{R}^n, \quad A_k \in \mathbb{R}^{n\times n},\ l = 1, \dots, M,$$

is given by

$$\det\left(I - \sum_{k=1}^{M} A_k e^{-\lambda r_k}\right) = 0,$$

which is seen, using an explicit formula for the determinant, to be an exponential polynomial with dependent delays.

2.2. Rationally dependent and rationally independent delays. The numbers $r_1, r_2, \dots, r_M$ are rationally independent if and only if

$$\sum_{k=1}^{M} n_k r_k = 0, \quad n_k \in \mathbb{Z},$$

implies $n_k = 0$, $k = 1, \dots, M$. For example, two numbers are rationally independent if their ratio is an irrational number. An important property of rationally independent numbers which will be used throughout the paper is given by Kronecker's theorem [11, Theorem 444].

THEOREM 2.1. *Given $r = (r_1, r_2, \dots, r_M)$ with rationally independent components and $\theta = (\theta_1, \dots, \theta_M)$ arbitrary, there exists a sequence of real numbers $\{d_n\}_{n\geq 1}$ such that*

$$\lim_{n\to\infty} e^{i(d_n r_k - \theta_k)} \to 1, \quad k = 1, \dots, M.$$

We now provide a useful characterization of $\bar{Z}(r)$ and its dependence on r. First, consider the following definitions. For any two sets E and $F \subset \mathbb{R}$ and any $\rho \in \mathbb{R}$, let

$$\begin{gathered} d(\rho, E) = \inf_{t\in E} |\rho - t|, \\ \delta(E, F) = \sup_{\rho\in E} d(\rho, F), \quad \text{and} \\ D(E, F) = \max\left\{\delta(E, F), \delta(F, E)\right\}. \end{gathered}$$

The number $D(E, F)$ is called the Hausdorff distance between the sets E and F.

We will illustrate with examples that $\bar{Z}(r)$ is not continuous w.r.t. the delays $r \in (0, \infty)^M$ since arbitrarily small delay changes can change their rational (in)dependence. However, the following weaker property holds [1, Lemma 2.5].

THEOREM 2.2. *$\bar{Z}(r)$ is lower semicontinuous in r; that is, for each $r_0 \in (0, \infty)^M$,*

$$\lim_{r\to r_0} \delta(\bar{Z}(r_0), \bar{Z}(r)) = 0.$$

When the delays r are rationally independent, we have [1, Theorem 2.2].

THEOREM 2.3. *$\bar{Z}(r)$ is continuous in the Hausdorff metric for rationally independent delays r.*

This result is important because it implies the continuity of the supremum $c(r)$ of $\bar{Z}(r)$ at rationally independent r. In this case the set $\bar{Z}(r)$ is completely characterized by [1, Theorem 3.1 and Corollary 3.2].

THEOREM 2.4. *If the components of r are rationally independent, then the following statements are equivalent:*

$$\alpha \in \bar{Z}(r)$$
$$\Updownarrow$$
$$\exists \theta = (\theta_1, \ldots, \theta_M) \text{ with } \theta_k \in [0, 2\pi],\ k = 1, \ldots, M,$$
$$\text{such that } 1 - \sum_{j=1}^{N} a_j e^{-\alpha \gamma_j \cdot r} e^{-i \gamma_j \cdot \theta} = 0.$$

COROLLARY 2.5. *$\bar{Z}$ is the union of a finite number of intervals.*

The combination of Theorems 2.2 and 2.3 is very important for control problems because the right-most characteristic roots determine stability: suppose, for example, that r_0 is given with rationally dependent components and denote the maximum of $\bar{Z}(r_0)$ by $c(r_0)$. On the other hand, consider a sequence of rationally independent delays $\{r_n\}_{n \geq 1}$ with limit r_0 and denote by $c(r_n)$ the maximum of $\bar{Z}(r_n)$. Then from Theorems 2.2 and 2.3 it follows that

$$c(r_0) \leq \lim_{n \to \infty} c(r_n). \tag{2.3}$$

In other words, the supremum of $\bar{Z}$ is always higher when one considers the given delays as independent. In section 3 it will be shown that inequality (2.3) can be strict. This means that when the delays in the characteristic equation modelling a physical system are results of independent phenomena (for example, independent measurements), one has always to consider the delays as rationally independent in order to obtain a reliable upper bound on the real parts of the spectrum. In section 3 will be explained what happens with the individual characteristic roots when one deals with rationally independent delays close to rationally dependent delays.

2.3. Special cases.

Fully independent delays. This corresponds to the case where $M = N$, $\gamma_j = e_j$, the jth unity vector in $\mathbb{R}^N$, and the delays $\tau_1, \tau_2, \ldots, \tau_N$ are rationally independent. Theorem 2.4 can be rewritten as the following.

THEOREM 2.6. *When the delays are rationally independent,*

$$c = \sup \{\Re(\lambda) : H(\lambda) = 0\}$$

satisfies

$$1 - \sum_{j=1}^{N} |a_j| e^{-c\tau_j} = 0. \tag{2.4}$$

The solution c of (2.4) also serves as a (nonstrict) upper bound in the case of rationally dependent delays.

Commensurate delays. This is the case when $M = 1$. Thus delays $\tau_1, \ldots, \tau_n$ are commensurate if and only if there exists a real number r such that $\tau_j = n_j r$ with $n_j \in \mathbb{N}$, $j = 1, \ldots, N$; i.e., all the delays are integer multiples of a same number. In this case (2.1) can be rewritten as a polynomial in $e^{-\lambda r}$. As a consequence, $\bar{Z}(r)$ consists of a finite number of points, and the spectrum is vertically periodic with period $\frac{2\pi}{r} i$.

3. Visualization and interpretation. In the previous section the delays r were considered fixed. When one approaches rationally dependent delays r_0, the supremum of the real parts of the characteristic roots can have a discontinuity. First, we illustrate how this discontinuity is compatible with the continuous movement of individual roots as r approaches r_0. Second, we discuss the consequences for control applications.

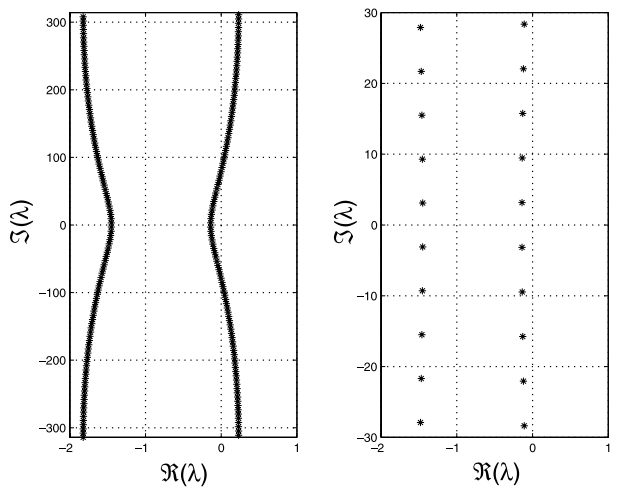


FIG. 3.1. *Part of the spectrum of* (3.1) *on two different scales for* $h = 0.01$.

3.1. Nonuniform convergence. Consider, as an example, the characteristic equation

$$H(\lambda, h) \triangleq 1 + 1.1e^{-\lambda} + 0.2e^{-\lambda(2+h)} = 0, \tag{3.1}$$

with delays 1 and $2 + h$. When h is zero, (3.1) is a quadratic equation in $e^{-\lambda}$ and the roots are $\lambda \approx -1.4704 + i(2l+1)\pi,\ l \in \mathbb{Z}$, and $\lambda \approx -0.1391 + i(2l+1)\pi,\ l \in \mathbb{Z}$. When $h > 0$ and irrational (and therefore the two delays are rationally independent), the supremum $c(h) = \sup\{\Re(\lambda) \,|\, H(\lambda) = 0\}$ satisfies

$$1 - 1.1e^{-c(h)} - 0.2e^{-c(h)(2+h)} = 0, \tag{3.2}$$

which yields $\lim_{h\to 0} c(h) \approx 0.2302 > c(0) \approx -0.1391$. Hence $c(h)$, and the corresponding stability of the associated essential spectrum changes discontinuously w.r.t. h. Individual (single) roots, however, move continuously w.r.t. the delays. From

$$1 + 1.1e^{-\lambda} + 0.2e^{-\lambda(2+h)} = 0$$

one derives

$$\frac{d\lambda}{dh} = \frac{-0.2\lambda e^{-\lambda(2+h)}}{1.1e^{-\lambda} + 0.2(2+h)e^{-\lambda(2+h)}}.$$

But this "sensitivity" of the individual roots increases to infinity as their modulus $|\lambda| \to \infty$. Figure 3.1 shows part of the spectrum of (3.1) on two different scales when $h = 0.01$. When h is reduced to zero, the spectrum converges *pointwise* and nonuniformly to the limit case $h = 0$, as shown in Figure 3.2.

3.2. Unstable difference equations cannot be stabilized. Consider the following control system with input $u(t)$:

$$\frac{d}{dt}(x(t) + 2x(t-1)) = ax(t) + bx(t-\tau) + u(t). \tag{3.3}$$

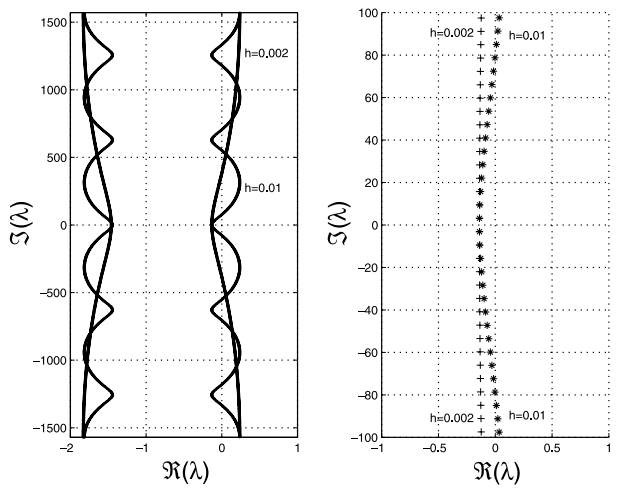


FIG. 3.2. *Part of the spectrum of* (3.1) *for* $h = 0.01$ *and* $h = 0.002$.

When $u(t) \equiv 0$, the difference equation

$$x(t) + 2x(t-1) = 0 \tag{3.4}$$

determines the essential spectrum of the semigroup associated with (3.3); see [8]. The zero solution of (3.4) is clearly unstable: all eigenvalues have real part $\log(2)$. In [17] it is shown that such an equation cannot be stabilized robustly in the presence of a small time-delay in the application of the feedback law.

When applying the velocity feedback $u(t) = \frac{3}{2}\dot{x}(t-1-h)$, the difference equation is modified to

$$x(t) + 2x(t-1) - \frac{3}{2}x(t-1-h) = 0, \tag{3.5}$$

where h models the estimation error of the delay. For $h = 0$ the difference equation is clearly stabilized: all roots have real part $-\log(2)$. However, for irrational h the supremum $c(h)$ of the real parts of the spectrum can be calculated from

$$1 - 2e^{-c(h)} - \frac{3}{2}e^{-c(h)(1+h)} = 0,$$

from which follows $\lim_{h\to 0} c(h) = \log(3.5) > \log(2)$. Thus a practical feedback destabilizes the original system even more. This is shown in Figure 3.3.

Because sensitivity to small changes of the delays is caused by roots of (2.1) with large modulus, and because the set of the real parts of the roots of (2.1) is contained in a finite number of intervals, such roots have large imaginary part. Thus sensitivity to infinitesimal changes in the delays is caused by modes of very high frequency. This is shown in Figure 3.4. Note that the control of (3.3) works for low frequency modes while it does not for high frequency modes (see Figure 3.3). We remark that the question arises whether the model used is a valid description of the modelled reality for such frequencies. In reality one (usually) expects larger damping for larger frequencies. Whether this damping occurs strongly and soon enough depends on

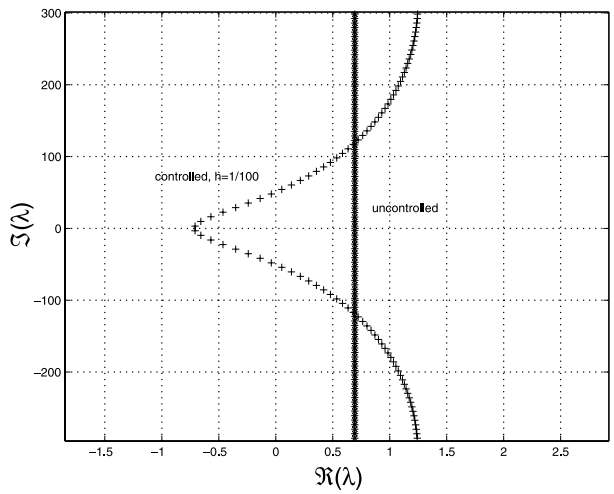


FIG. 3.3. *Part of the spectrum of the uncontrolled system* (3.4) *and the controlled system* (3.5) *for* $h = 0.01$.

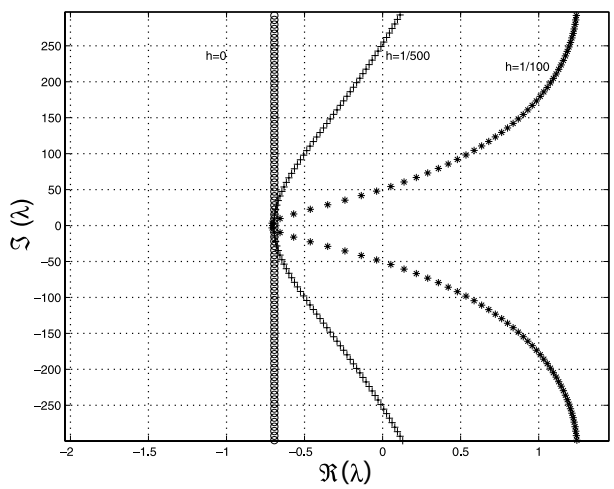


FIG. 3.4. *When* $h \to 0$, *the spectrum of* (3.5) *converges pointwise to the spectrum of* $x(t) + \frac{1}{2}x(t-1) = 0$.

the particular application. In section 4 we will see that our generalization leads to situations where sensitivity is *not necessarily* caused by high frequency modes.

4. Vanishing delays. The analysis in section 2 is valid under the assumption that all the delays are fixed, different, and nonzero. These assumptions can be relaxed to the requirements that, first, the smallest delays are not arbitrary close to zero and, second, that the largest delays are not arbitrary close to each other. In this section we explicitly deal with these limit cases. We show that vanishing delays can give rise to roots with unbounded positive real parts and that, in a similar way, coinciding largest delays can give rise to roots with unbounded negative real part. Since the latter is of less importance for applications, we mention only briefly the occurrence of this phenomenon.

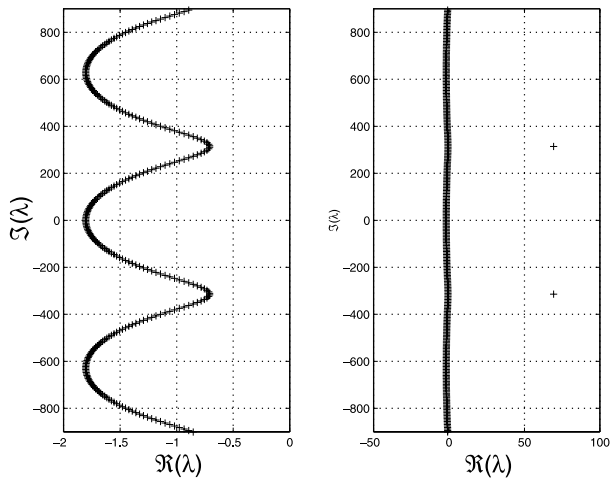


FIG. 4.1. *Zeros of* (4.1) *for* $h = 0.01$ *in two different regions of the complex plane.*

4.1. Introductory example. We investigate the zeros of

$$H(\lambda, h) = 1 + 2e^{-\lambda h} - \frac{1}{2}e^{-\lambda} \tag{4.1}$$

as $h \to 0+$.

If we set h to 0 in (4.1), all roots of $H(\lambda, 0)$ are of the form $\lambda = -\log(6) + i2\pi l$, $l \in \mathbb{Z}$, and the collection of real parts of the roots of $H(\lambda, 0)$ is $\bar{Z}(0) = \{-\log(6)\}$. However, from the analysis of section 2, we know that for h and 1 rationally independent (i.e., h irrational), $\bar{Z}(h)$ coincides with the α-components of all solutions $(\alpha, \theta_1, \theta_2)$ of

$$1 + 2e^{-\alpha h}e^{-i\theta_1} - \frac{1}{2}e^{-\alpha}e^{-i\theta_2} = 0. \tag{4.2}$$

If h goes to 0 in (4.2), we are led to the conclusion that

$$\lim_{h \to 0+} \bar{Z}(h) = [-\log(6), -\log(2)], \tag{4.3}$$

i.e., that, although the real part of each individual root of $H(\lambda, h)$ approaches $-\log(6)$ as h goes to zero, at the same time the collection of all the real parts of all roots converges to (4.3).

Figure 4.1 (left panel) shows part of the roots of (4.1) for $h = 0.01$. At first glance this confirms the above conclusions. However, if we look at a larger region in the complex plane (see the right panel of Figure 4.1) we see that there exist additional roots of $H(\lambda, h)$ with quite different behavior. When h is further reduced, the real parts of the roots at $\Re(\lambda) \approx 69.3$ move off to $+\infty$, approximately as the solutions of

$$1 + 2e^{-\lambda h} = 0. \tag{4.4}$$

Indeed, if the real part of λ is large, we cannot set λh to zero. Rather, we can neglect $\frac{1}{2}e^{-\lambda}$, leading to (4.4) and

$$\lambda \approx \frac{1}{h}\left(\log 2 + i(2l+1)\pi\right), \quad l \in \mathbb{Z}. \tag{4.5}$$

Formula (4.5) clearly illustrates that arbitrarily small delays ($0 < h \ll 1$) can lead to arbitrarily unstable characteristic roots ($\Re(\lambda) \gg 1$).

The situation can be summarized as follows. When h tends to zero, the spectrum consists partly of roots with bounded real part, which can be analyzed along the lines of section 2, and partly of "diverging" roots, whose real part grows without bound as the solutions of the "small-delay part" (4.4) of (4.1). In the rest of this section these properties will be generalized to the multiple delay case.

4.2. Notation. The general form of the exponential polynomial studied within this section is

$$H(\lambda, r, s) = 1 - \sum_{i \in I} a_i e^{-\lambda \tau_i} - \sum_{j \in J} b_j e^{-\lambda \tau_j}, \tag{4.6}$$

where

$$\forall i \in I : \tau_i = \gamma_i \cdot r, \quad \forall j \in J : \tau_j = \gamma_j \cdot r + \nu_j \cdot s,$$

and where

$$I = \{1, 2, \ldots, N_1\}, \quad J = \{N_1 + 1, N_1 + 2, \ldots, N_1 + N_2\}$$

are used for notational convenience. The components of $r \in [0, +\infty)^M$ and $s \in [0, +\infty)^L$ are the independent delays; $\gamma_i \in \mathbb{N}^M$, $\gamma_j \in \mathbb{N}^M$, and $\nu_j \in \mathbb{N}^L$ are vectors with nonnegative integer coefficients. γ_i and ν_j are nonzero vectors; that is, both have at least one nonzero element for all i and j. Splitting the independent delays into r and s opens the possibility of dealing with a combination of "normal" and arbitrarily small delays by letting $r \to 0$ combined with constant $s > 0$.

We also extend the definition of the inner product "$\cdot$" to the situation with $\gamma_i \in \mathbb{N}^M$ and $R \in [0, +\infty]^M$. Then

$$\gamma_i \cdot R = \sum_{j=1}^{M} \gamma_{i,j} R_j$$

has the usual meaning, except that $\gamma_{i,j} \times (R_j = +\infty)$ is taken to be 0 when $\gamma_{i,j} = 0$, and $+\infty$ otherwise. The underlying rationale for this is that $R_j = +\infty$ will be the result of some limit while the $\gamma_{i,j}$ are fixed.

4.3. Arbitrarily unstable characteristic roots. The "small-delay part" of (4.6) is

$$1 - \sum_{i \in I} a_i e^{-\lambda \tau_i}.$$

We now prove how its solutions determine when arbitrarily small delays can lead to arbitrarily unstable characteristic roots.

THEOREM 4.1. *The following statements are equivalent:*

$$\exists \theta \in [0, 2\pi]^M,\ \exists R \in [0, +\infty]^M \text{ such that } 1 - \textstyle\sum_{i \in I} a_i e^{-\gamma_i \cdot R} e^{-i \gamma_i \cdot \theta} = 0$$
$$\Updownarrow$$
$$\exists \{r_n\}_{n \geq 1},\ \{c_n\}_{n \geq 1},\ \{d_n\}_{n \geq 1},$$
$$\text{with } \lim_{n \to \infty} c_n = \infty,\ r_n \geq 0, \text{ and } \lim_{n \to \infty} \|r_n\| = 0 \text{ and such that}$$
$$\lim_{n \to \infty} H(c_n + i d_n, r_n, s) = 0 \text{ for fixed } s > 0.$$

Proof of $\Downarrow$. Consider a (re)ordered partition of $R = (R_1, \ldots, R_K, R_{K+1}, \ldots, R_M)$ such that $R_1, \ldots, R_K$ are finite and $R_{K+1}, \ldots, R_M$ are infinite. That is, let $R = (R^{[1]}, R^{[2]})$ with $R^{[1]} = (R_1, R_2, \ldots, R_K) \in \mathbb{R}^K$ and $R^{[2]} = (\infty, \infty, \ldots, \infty) = \infty^{M-K}$. Likewise, consider the corresponding partition for θ and γ_i: $\theta = (\theta^{[1]}, \theta^{[2]})$ and $\gamma_i = (\gamma_i^{[1]}, \gamma_i^{[2]})$, with $\theta^{[1]} = (\theta_1, \theta_2, \ldots, \theta_K)$ the first K components and $\theta^{[2]} = (\theta_{K+1}, \theta_{K+2}, \ldots, \theta_M)$ the remaining $M-K$ components of θ, and similarly for $\gamma_i^{[1]}$ and $\gamma_i^{[2]}$, $i \in I$. Define the set of indices $I_1 \subseteq I$, whereby for $i \in I_1$ the last $M-K$ components of γ_i are zero, that is, where $\gamma_i^{[2]} = 0^{M-K}$, and set $I_2 = I \setminus I_1$. Obviously

$$1 - \sum_{i \in I} a_i e^{-\gamma_i \cdot R} e^{-i\gamma_i \cdot \theta} = 0$$

can be written as

$$1 - \sum_{i \in I_1} a_i e^{-\gamma_i^{[1]} \cdot R^{[1]}} e^{-i\gamma_i^{[1]} \cdot \theta^{[1]}} = 0.$$

Because the components of $R^{[1]}$ may be rationally dependent, consider a sequence $\{u_n^{[1]}\}_{n \geq 1}$ that converges to $R^{[1]}$ but whereby the components of $u_n^{[1]} \in (0, +\infty)^K$ are rationally independent for each n. Choose a (strictly positive) sequence of real numbers $\{\epsilon_n\}_{n \geq 1}$ with $\lim_{n \to \infty} \epsilon_n = 0$ such that

$$\|u_n^{[1]} - R^{[1]}\| < \epsilon_n.$$

Because $u_n^{[1]}$ has rationally independent coefficients, due to Theorem 2.1, there exists, for each n, a sequence of real numbers $\{v_{n,m}\}_{m \geq 1}$ such that

$$\lim_{m \to \infty} e^{i\gamma_i^{[1]} \cdot (v_{n,m} u_n^{[1]} - \theta^{[1]})} = 1 \ \forall i \in I_1;$$

hence $\exists m^*(n)$ such that $|e^{i\gamma_i^{[1]} \cdot (v_{n,m^*(n)} u_n^{[1]} - \theta^{[1]})} - 1| < \epsilon_n \ \forall i \in I_1$. Set $v_n = v_{n,m^*(n)}$. We have created $\{u_n^{[1]}\}_{n \geq 1}$ and $\{v_n\}_{n \geq 1}$ with

$$\begin{gathered} \|u_n^{[1]} - R^{[1]}\| < \epsilon_n, \\ |e^{i\gamma_i^{[1]} \cdot (v_n u_n^{[1]} - \theta^{[1]})} - 1| < \epsilon_n \ \forall i \in I_1, \text{ and} \\ \lim_{n \to \infty} \epsilon_n = 0. \end{gathered}$$

Choose further $\{u_n^{[2]}\}_{n \geq 1}$, with $u_n^{[2]} \in (0, +\infty)^{M-K}$ and with $\lim_{n \to \infty} u_n^{[2]} = \infty^{M-K}$, and define $\{u_n\}_{n \geq 1}$ as $u_n = (u_n^{[1]}, u_n^{[2]}) \in (0, +\infty)^M$.

We are now in a position to choose a sequence of real parts c_n. Choose $\{c_n\}_{n \geq 1}$ with $c_n \in (0, +\infty)$ such that c_n goes to infinity faster than every component of u_n, that is, such that $\lim_{n \to \infty} c_n = +\infty$ and $\lim_{n \to \infty} \frac{1}{c_n} u_n = 0^M$. Second, define a sequence of imaginary parts $\{d_n\}_{n \geq 1}$ as $d_n = c_n v_n$, and a sequence of vanishing delays $\{r_n\}_{n \geq 1}$ as $r_n = \frac{1}{c_n} u_n$.

We now have

$$\begin{aligned} &H(c_n + id_n, r_n, s) \\ &\quad = 1 - \sum_{i \in I} a_i e^{-c_n \gamma_i \cdot r_n} e^{-id_n \gamma_i \cdot r_n} - \sum_{j \in J} b_j e^{-c_n(\gamma_j \cdot r_n + \nu_j \cdot s)} e^{-id_n(\gamma_j \cdot r_n + \nu_j \cdot s)} \\ &\quad = 1 - \sum_{i \in I} a_i e^{-\gamma_i \cdot u_n} e^{-i\gamma_i \cdot v_n u_n} - \sum_{j \in J} b_j e^{-c_n(\gamma_j \cdot r_n + \nu_j \cdot s)} e^{-id_n(\gamma_j \cdot r_n + \nu_j \cdot s)}. \end{aligned}$$

The second term can be split into, first, $\sum_{i\in I_1} a_i e^{-\gamma_i\cdot u_n} e^{-i\gamma_i\cdot v_n u_n}$, whereby the last $M-K$ components of γ_i are zero and thus $\gamma_i\cdot u_n = \gamma_i^{[1]}\cdot u_n^{[1]}$, and, second, $\sum_{i\in I_2} a_i e^{-\gamma_i\cdot u_n} e^{-i\gamma_i\cdot v_n u_n}$, whereby $\lim_{n\to\infty}\gamma_i\cdot u_n = +\infty$.

Hence

$$
\begin{aligned}
&H(c_n + id_n, r_n, s)\\
&= 1 - \sum_{i\in I_1} a_i e^{-\gamma_i^{[1]}\cdot R^{[1]}} e^{-i\gamma_i^{[1]}\cdot\theta^{[1]}} e^{-\gamma_i^{[1]}\cdot(u_n^{[1]}-R^{[1]})} e^{-i\gamma_i^{[1]}\cdot(v_n u_n^{[1]}-\theta^{[1]})}\\
&\quad - \sum_{i\in I_2} a_i e^{-\gamma_i\cdot u_n} e^{-i\gamma_i\cdot v_n u_n} - \sum_{j\in J} b_j e^{-c_n(\gamma_j\cdot r_n+\nu_j\cdot s)} e^{-id_n(\gamma_j\cdot r_n+\nu_j\cdot s)}\\
&= 1 - \sum_{i\in I_1} a_i e^{-\gamma_i^{[1]}\cdot R^{[1]}} e^{-i\gamma_i^{[1]}\cdot\theta^{[1]}} \underbrace{e^{-\gamma_i^{[1]}\cdot(u_n^{[1]}-R^{[1]})}}_{\to 1} \underbrace{e^{-i\gamma_i^{[1]}\cdot(v_n u_n^{[1]}-\theta^{[1]})}}_{\to 1}\\
&\quad - \sum_{i\in I_2} a_i e^{-\overbrace{\gamma_i\cdot u_n}^{\to\infty}} e^{-i\gamma_i\cdot v_n\bar{u}_n} - \sum_{j\in J} b_j e^{-c_n\overbrace{\nu_j\cdot s}^{\neq 0}} e^{-\overbrace{c_n\gamma_j\cdot r_n}^{\to\infty}} e^{-id_n(\gamma_j\cdot r_n+\nu_j\cdot s)}
\end{aligned}
$$

tends to zero as n approaches infinity, which completes this part of the proof.

Proof of $\Uparrow$. The $\lim_{n\to\infty} H(c_n + id_n, r_n, s) = 0$ implies

$$\lim_{n\to\infty} 1 - \sum_{i\in I} a_i e^{-\gamma_i\cdot c_n r_n} e^{-i\gamma_i\cdot d_n r_n} = 0, \tag{4.7}$$

because the other term vanishes at infinity.

Consider the sequence $\{c_n r_n\}_{n\geq 1}$ with elements $c_n r_n = (c_n r_{n,1}, \ldots, c_n r_{n,M})$. For each sequence of the kth component, $\{c_n r_{n,k}\}_{n\geq 1}$, there are two possibilities as n tends to infinity: either the sequence is unbounded (with or without limit) or the sequence is bounded (with or without limit). Suppose that $\{c_n r_{n,k}\}_{n\geq 1}$ is unbounded. Then there exists a subsequence with limit infinity. When, on the other hand, $\{c_n r_{n,k}\}_{n\geq 1}$ is bounded, a subsequence with finite limit exists.

This way one can repeatedly construct subsequences to obtain an infinite set of indices S_1 such that for each $k\in\{1, 2\ldots, M\}$, $\lim_{n\to\infty, n\in S_1} c_n r_{n,k}$ exists in $[0, +\infty]$. We still have

$$\lim_{n\to\infty, n\in S_1} 1 - \sum_{i\in I} a_i e^{-\gamma_i\cdot c_n r_n} e^{-i\gamma_i\cdot d_n r_n} = 0, \tag{4.8}$$

and because each γ_i is a vector of integer coefficients, $e^{-i\gamma_i\cdot d_n r_n}$ equals $e^{-i\gamma_i\cdot((d_n r_n) \bmod 2\pi)}$ and thus

$$\lim_{n\to\infty, n\in S_1} 1 - \sum_{i\in I} a_i e^{-\gamma_i\cdot c_n r_n} e^{-i\gamma_i\cdot((d_n r_n) \bmod 2\pi)} = 0.$$

Consider the sequence of vectors $\{(d_n r_n) \bmod 2\pi\}_{n\geq 1, n\in S_1}$, which is bounded and thus contained in a compact set (w.r.t. some norm). Consequently, it must have a converging subsequence, denoted by the indices $S_2 \subset S_1$. Finally,

$$\lim_{n\to\infty, n\in S_2} 1 - \sum_{i\in I} a_i e^{-\gamma_i\cdot c_n r_n} e^{-i\gamma_i\cdot((d_n r_n) \bmod 2\pi)} = 0,$$

and we define

$$R_k = \lim_{n\to\infty, n\in S_2} c_n r_{n,k} \in [0,+\infty], \quad k = 1,2,\ldots,M,$$

and

$$\theta_k = \lim_{n\to\infty, n\in S_2} (d_n r_{n,k}) \bmod 2\pi \in [0,2\pi], \quad k = 1,2,\ldots,M.$$

Hence we have defined every component of R and θ in such a way that

$$1 - \sum_{i\in I} a_i e^{-\gamma_i\cdot R} e^{-i\gamma_i\cdot\theta} = 0,$$

which completes the proof. □

We now strengthen the results of the previous theorem: we prove the existence of a sequence of *roots* λ_n of $H(\lambda, r_n, s) = 0$ with real part tending to $+\infty$, in other words, that arbitrarily small delays cause roots with arbitrarily large real part.

THEOREM 4.2. *Consider the statements*

(a) $\exists\theta \in [0,2\pi]^M$, $\exists R \in [0,+\infty]^M$ *such that* $1 - \sum_{i\in I} a_i e^{-\gamma_i\cdot R} e^{-i\gamma_i\cdot\theta} = 0$;

(b) $\exists i \in I$ *such that* $\gamma_i \cdot R \neq 0$ *and* $\gamma_i \cdot R \neq +\infty$;

(c) $\exists \{r_n\}_{n\geq 1}$, $\{\lambda_n = e_n + if_n\}_{n\geq 1}$ *with* $\lim_{n\to\infty} e_n = +\infty$, $r_n \geq 0$, *and* $\lim_{n\to\infty} r_n = 0^M$ *such that* $H(\lambda, r_n, s) = 0$.

Then the following hold:

(1) (a) *and* (b) $\Rightarrow$ (c),

(2) (c) $\Rightarrow$ (a).

The special case where (a) holds but (b) is violated is treated separately in subsection 4.5. This corresponds to a degenerate case where the "small delay part" of the characteristic equation doesn't provide enough information about the existence of roots with large real part for vanishing delays.

Proof of (2). This proof is trivial (Theorem 4.1).

Proof of (1). Following Theorem 4.1 there exist sequences $\{r_n\}_{n\geq 1}$, $\{c_n\}_{n\geq 1}$, and $\{d_n\}_{n\geq 1}$, with $\lim_{n\to\infty} c_n = +\infty$, $r_n \geq 0$, and $\lim_{n\to\infty} r_n = 0^M$, such that

$$\lim_{n\to\infty} H(c_n + id_n, r_n, s) = 0,$$

and the proof has provided us with a way to construct such sequences.

Choose $\{\epsilon_n\}_{n\geq 1}$, $\{u_n^{[1]}\}_{n\geq 1}$, and $\{v_n\}_{n\geq 1}$ as in Theorem 4.1 and such that

$$\left| 1 - \sum_{i\in I_1} a_i e^{-\gamma_i^{[1]}\cdot R^{[1]}} e^{-i\gamma_i^{[1]}\cdot\theta^{[1]}} e^{-\gamma_i^{[1]}\cdot(u_n^{[1]} - R^{[1]})} e^{-i\gamma_i^{[1]}\cdot(v_n u_n^{[1]} - \theta^{[1]})} \right| < e^{-n}.$$

Further requirements on the decay-rate, which can be chosen arbitrarily fast, will be given later in the proof (see formulae (4.9) and (4.10)). Choose $\{c_n\}_{n\geq 1}$ as $c_n = n$, $\{u_n^{[2]}\}_{n\geq 1}$ as $u_n^{[2]} = \sqrt{n}(1,1,\ldots,1)$, $\{d_n\}_{n\geq 1}$ as $d_n = v_n c_n$, $\{u_n\}_{n\geq 1}$ as $u_n = (u_n^{[1]}, u_n^{[2]})$, and $\{r_n\}_{n\geq 1}$ as $r_n = \frac{1}{c_n} u_n$.

Consider the functions $\bar{H}_n(\lambda) = c_n H(\lambda, r_n, s)$. Using the particular choices above it is straightforward to show that

$$\lim_{n\to\infty} \bar{H}_n(c_n + id_n) = 0, \qquad \lim_{n\to\infty} \frac{d\bar{H}_n}{d\lambda}(c_n + id_n) = Q,$$

where

$$Q = \sum_{i \in I_1} a_i \gamma_i^{[1]} \cdot R^{[1]} e^{-\gamma_i^{[1]} \cdot R^{[1]}} e^{-i\gamma_i^{[1]} \cdot \theta^{[1]}} \in \mathbb{C}$$

and

$$\lim_{n \to \infty} \frac{d^k \bar{H}_n}{d\lambda^k}(c_n + id_n) = 0, \quad k \geq 2.$$

Around $c_n + id_n$, the analytic function $\bar{H}_n$ can be expanded as

$$\bar{H}_n(\lambda) = \sum_{k=0}^{\infty} \frac{d^k \bar{H}_n}{d\lambda^k}\bigg|_{c_n + id_n} (\lambda - (c_n + id_n))^k,$$

and when defining $\mu = \lambda - (c_n + id_n)$,

$$\bar{H}_n(\mu) = \sum_{k=0}^{\infty} \frac{d^k \bar{H}_n}{d\lambda^k}\bigg|_{c_n + id_n} \mu^k.$$

We will now show that $\bar{H}_n(\mu)$ converges uniformly on $|\mu| \leq 1$ to the function $\bar{H}(\mu) = Q\mu$ or, equivalently, that the function

$$D_n(\mu) = \bar{H}_n(\mu) - Q\mu$$

converges uniformly to zero in $|\mu| \leq 1$. In the appendix it is shown (Lemma A.1) that this implies that $\bar{H}_n(\mu)$ has a root near $\mu = 0$ for sufficiently large n when $Q \neq 0$. One should emphasize that μ is in fact a local coordinate depending on n: we compare the local behavior of each function $\bar{H}_n(\lambda)$ near $c_n + id_n$ with $\bar{H}(\mu) = Q\mu$.

To construct an upper bound on $|D_n(\mu)|$ for $|\mu| \leq 1$, first define strictly positive numbers Δ_i, Δ_j, β_i and integer P such that, for $n \geq P$,

$$\begin{gathered} \gamma_j \cdot r_n \leq \Delta_j, \quad j \in J, \\ \gamma_i \cdot c_n r_n \leq \gamma_i^{[1]} \cdot R^{[1]} + \Delta_i, \quad i \in I_1, \\ \gamma_i \cdot r_n \leq 1, \quad i \in I = I_1 \cup I_2, \text{ and} \\ \gamma_i \cdot c_n r_n \geq \beta_i \sqrt{n}, \quad i \in I_2. \end{gathered}$$

Second, the decay rate of $\{\epsilon_n\}_{n \geq 1}$ should be chosen in such a way that for $n \geq P$

$$n \left| 1 - \sum_{i \in I_1} a_i e^{-\gamma_i \cdot c_n r_n} e^{-i\gamma_i \cdot d_n r_n} \right| \leq \frac{1}{n} \sum_{i \in I_1} |a_i| \tag{4.9}$$

and

$$\left| \sum_{i \in I_1} a_i \gamma_i \cdot c_n r_n e^{-\gamma_i \cdot c_n r_n} e^{-i\gamma_i \cdot d_n r_n} - Q \right| \leq \frac{1}{n} \sum_{i \in I_1} |a_i| (\gamma_i^{[1]} \cdot R^{[1]} + \Delta_i). \tag{4.10}$$

In this way we can compute bounds for $D_n(0)$ and its derivatives in $\mu = 0$, $n \geq P$:

$$\begin{aligned} |D_n(0)| &\leq n \left| 1 - \sum_{i \in I_1} a_i e^{-\gamma_i \cdot c_n r_n} e^{-i\gamma_i \cdot d_n r_n} \right| + n \left| \sum_{i \in I_2} a_i e^{-\gamma_i \cdot c_n r_n} e^{-i\gamma_i \cdot d_n r_n} \right| \\ &\quad + n \sum_{j \in J} \left| b_j e^{-\nu_j \cdot s c_n} e^{-\gamma_j \cdot c_n r_n} e^{-i\gamma_j \cdot d_n r_n} \right| \\ &\leq \frac{1}{n} \sum_{i \in I_1} |a_i| + \sum_{i \in I_2} n e^{-\beta_i \sqrt{n}} |a_i| + \sum_{j \in J} n e^{-n\nu_j \cdot s} |b_j|, \end{aligned}$$

and, concerning the derivatives,

$$\left|\frac{dD_n}{d\mu}(0)\right| \leq \frac{1}{n}\sum_{i\in I_1}|a_i|(\gamma_i^{[1]}\cdot R+\Delta_i)+\sum_{i\in I_2} ne^{-\beta_i\sqrt{n}}|a_i| + \sum_{j\in J} n(\Delta_j+\nu_j\cdot s)e^{-n\nu_j\cdot s}|b_j|,$$

and for $k\geq 2$,

$$\left|\frac{d^kD_n}{d\mu^k}(0)\right| \leq \frac{1}{n}\sum_{i\in I_1}|a_i|(\gamma_i^{[1]}\cdot R^{[1]}+\Delta_i)+\sum_{i\in I_2} ne^{-\beta_i\sqrt{n}}|a_i| + \sum_{j\in J} n(\Delta_j+\nu_j\cdot s)^k e^{-n\nu_j\cdot s}|b_j|.$$

Finally, for each μ with $|\mu|\leq 1$,

$$\begin{aligned}
|D_n(\mu)| &\leq \sum_{k=0}^{\infty}\frac{1}{k!}\left|\frac{d^kD}{d\mu^k}\right| \\
&\leq \sum_{i\in I_1}\frac{1}{n}|a_i|(\gamma_i^{[1]}\cdot R^{[1]}+\Delta_i)\sum_{k=0}^{\infty}\frac{1}{k!}+\sum_{i\in I_2} ne^{-\beta_i\sqrt{n}}\sum_{k=0}^{\infty}|a_i|\frac{1}{k!} \\
&\quad+\sum_{j\in J} ne^{-n\nu_j\cdot s}|b_j|\sum_{k=0}^{\infty}\frac{(\Delta_j+\nu_j\cdot s)^k}{k!} \\
&\leq \sum_{i\in I_1}\frac{1}{n}|a_i|(\gamma^{[1]}\cdot R^{[1]}+\Delta_i)e+\sum_{i\in I_2} ne^{-\beta_i\sqrt{n}}|a_i|e \\
&\quad+\sum_{j\in J} ne^{-n\nu_j\cdot s}|b_j|e^{(\Delta_j+\nu_j\cdot s)}
\end{aligned}$$

can thus be bounded uniformly, whereby the upper bound tends to zero as $n\to\infty$.

Because the series $\{\bar{H}_n(\mu)\}_{n\geq 1}$ converges uniformly to a function $\bar{H}(\mu)=Q\mu$, due to Lemma A.1, there must be a number $N\in\mathbb{N}$ such that for $\forall n\geq N$, $\bar{H}_n(\mu)$ has a root which converges to zero as $n\to\infty$ whenever $Q\neq 0$. Returning to the original problem, this means that c_n+id_n asymptotically coincides with a root e_n+if_n of $\bar{H}_n(\lambda)$, which is of course a root of $H(\lambda,r_n,s)$. Thus $\forall n\geq N$, $\exists e_n+if_n : H(e_n+if_n,r_n,s)=0$ and $\lim_{n\to\infty}(c_n-e_n)+i(d_n-f_n)=0$. By renumbering the sequences starting from $n=N$, the proof is complete in the case $Q\neq 0$.

We will now show that when $Q=0$ there always exists an integer k such that $\sum_{i\in I_1} a_i(\gamma_i^{[1]}\cdot R^{[1]})^l e^{-\gamma_i^{[1]}\cdot R^{[1]}}e^{-i\gamma_i^{[1]}\cdot\theta^{[1]}}=0$ for $l<k$ and $Q'=\sum_{i\in I_1} a_i(\gamma_i^{[1]}\cdot R^{[1]})^k e^{-\gamma_i^{[1]}\cdot R^{[1]}}e^{-i\gamma_i^{[1]}\cdot\theta^{[1]}}\neq 0$. In this case, $c_n^kH(\lambda,r_n,s)$ will converge locally to $Q'(\lambda-(c_n+id_n))^k$, and the proof proceeds as in the case $Q\neq 0$.

We prove that if Q' does not exist, (b) is necessarily violated. For notational convenience, define $\hat{a}_i = a_ie^{-\gamma_i^{[1]}\cdot R^{[1]}}e^{-i\gamma_i^{[1]}\cdot\theta^{[1]}}$. Given $\sum_{i\in I_1}\hat{a}_i=1$, $\sum_{i\in I_1}(\gamma_i^{[1]}\cdot R^{[1]})\hat{a}_i=0$, and $\sum_{i\in I_1}(\gamma_i^{[1]}\cdot R^{[1]})^2\hat{a}_i=0,\ldots$ (because Q' does not exist), multiply these equations with scalars and add them together to obtain

$$p(0)=\sum_{i\in I_1}\hat{a}_ip(\gamma_i^{[1]}\cdot R^{[1]}) \tag{4.11}$$

for all polynomials $p(\lambda)$. In (4.11) it is possible that for some indices $k, l \in I_1$, $\gamma_k^{[1]} \cdot R^{[1]} = \gamma_l^{[1]} \cdot R^{[1]}$ or $\gamma_k^{[1]} \cdot R^{[1]} = 0$. We group these factors:

$$g_0 p(0) + \sum g_j p(\gamma_j^{[1]} \cdot R^{[1]}) = 0.$$

If we choose for $p(\lambda)$ an interpolating (complex) polynomial such that $p(0) = \bar{g}_0$ and $p(\gamma_j^{[1]} \cdot R^{[1]}) = \bar{g}_j \ \forall j$, this leads to

$$|g_0|^2 + \sum |g_j|^2 = 0 \Rightarrow g_0 = 0,\ g_j = 0 \quad \forall j \Rightarrow g_0 + \sum g_j e^{\gamma_j^{[1]} \cdot R^{[1]}} = 0$$

or

$$1 - \sum_{i \in I_1} a_i e^{-i\gamma_i^{[1]} \cdot \theta^{[1]}} = 0.$$

This means that in $1 - \sum_{i \in I} a_i e^{-\gamma_i \cdot R} e^{-i\gamma_i \cdot \theta} = 0 \ \forall i \in I$, $\gamma_i \cdot R$ is zero or infinity, which contradicts statement (b). □

For control applications the following theorem is important. It provides sufficient and necessary conditions for (4.6) to have roots with arbitrarily large real part but small imaginary part in the presence of vanishing delays.

THEOREM 4.3. *Consider the statements*

(a) $\exists R \in [0, +\infty]^M$ *such that* $1 - \sum_{i \in I} a_i e^{-\gamma_i \cdot R} = 0$;

(b) $\exists i \in I$ *such that* $\gamma_i \cdot R \neq 0$ *and* $\gamma_i \cdot R \neq +\infty$;

(c) $\exists \{r_n\}_{n \geq 1}$, $\{\lambda_n\}_{n \geq 1}$ *with* $\lambda_n \in \mathbb{C}$, $\lim_{n \to \infty} \Re(\lambda_n) = +\infty$, $\lim_{n \to \infty} \Im(\lambda_n) = 0$, $r_n \geq 0$, *and* $\lim_{n \to \infty} \|r_n\| = 0$ *and such that* $H(\lambda_n, r_n, s) = 0$.

Then the following hold:

(1) (a) *and* (b) $\Rightarrow$ (c),

(2) (c) $\Rightarrow$ (a).

The above results can easily be shown by following the lines of the proofs of Theorems 4.1 and 4.2. Note that the occurrence of arbitrarily large roots does not depend on whether the small delays are rationally (in)dependent.

4.4. Interpretation and illustration. When the delays r in (4.6) approach zero, the real part of each characteristic root remains bounded or moves to $+\infty$. We call the set of these roots the *finite* (respectively, the *infinite*) part of the spectrum.

The supremum of the real parts of the finite part of the spectrum, c_{f}, can be calculated as the right-most solution α of

$$1 - \sum_{i \in I} a_i e^{-i\gamma_i \cdot \theta_1} - \sum_{j \in J} b_j e^{-\alpha \nu_j \cdot s} e^{-i(\gamma_j \cdot \theta_1 + \nu_j \cdot \theta_2)} = 0, \tag{4.12}$$

which corresponds to applying Theorem 2.4 and putting αr to zero. This last step is allowed because the delays r are arbitrarily small and α is infinite.

The infinite part of the spectrum is empty if

$$1 - \sum_{i \in I} a_i e^{-\gamma_i \cdot R} e^{-i\gamma_i \cdot \theta} = 0 \tag{4.13}$$

has no solution with $R \in [0, +\infty]^M$. Otherwise, we can conclude that arbitrarily unstable roots exist, except in the degenerate case where all the solutions of (4.13) with $R \geq 0$ satisfy that $\gamma_i \cdot R$ is zero or infinite for all $i \in I$. For this degenerate case, the

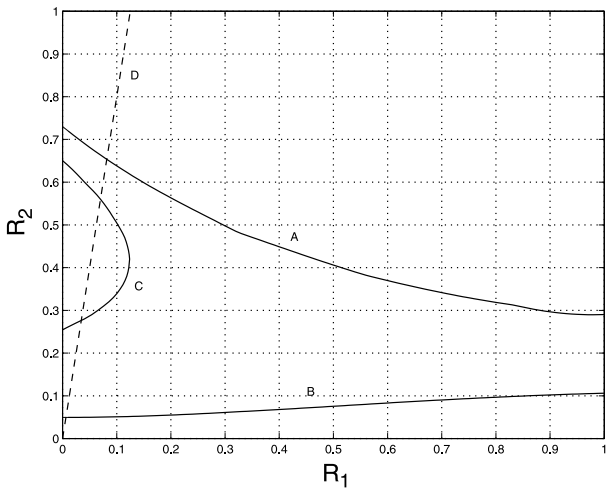


FIG. 4.2. *Projection of solutions of* (4.16).

small-delay part of the exponential polynomial does not provide enough information about the existence of the infinite spectrum, as will be shown in section 4.5.

We now illustrate how the solutions of (4.13) determine the structure of the infinite part of the spectrum and raise and partially answer an important open question which naturally arises from our analysis.

In both Theorems 4.2 and 4.3 we have allowed complete freedom in the way the independent delays approach zero. In practical applications it may be of interest to know whether additional dependencies between the delays (like, e.g., $r_1 < r_2 \to 0$) can influence the stability results. For this consider the following example:

$$1 - 2e^{-\lambda(r_1+r_2)} + 4e^{-\lambda 8 r_2} - \frac{1}{2}e^{-\lambda 2} = 0, \tag{4.14}$$

which has as small-delay part, when r_1 and $r_2 \to 0+$,

$$1 - 2e^{-\lambda(r_1+r_2)} + 4e^{-\lambda 8 r_2} = 0. \tag{4.15}$$

The supremum of the real parts of the finite part of the spectrum is $c_{\mathrm{f}} = -\frac{1}{2}\log 2 < 0$. The infinite part of the spectrum is nonempty if and only if

$$1 - 2e^{-(R_1+R_2)}e^{-i(\theta_1+\theta_2)} + 4e^{-8R_2}e^{-i8\theta_2} = 0 \tag{4.16}$$

has solutions with R_1 and $R_2 \in [0, +\infty]$. In Figure 4.2 the *area* between curves A, B, and C is the projection on the (R_1, R_2)-plane of the solutions of (4.16). Using Theorem 4.3, (4.14) has roots with real part tending to $+\infty$ but small imaginary part if and only if (4.16) has a solution with R_1 and $R_2 \in [0, +\infty]$ and $\theta_1 = \theta_2 = 0$. These solutions form the *curve* C in Figure 4.2.

Following Theorem 4.2 the solutions of $1 - \sum_{i\in I} a_i e^{-\gamma_i \cdot R} e^{-i\gamma_i \cdot \theta} = 0$ determine the existence of roots with large real part.

When the delays approach zero with a fixed ratio these solutions further provide a complete characterization of the structure of the diverging characteristic roots. This will be illustrated for the example discussed above. We will consider only the solutions of (4.15), since one can show that these solutions asymptotically coincide with solutions of (4.14) as $(r_1, r_2) \to (0, 0)$ (proof similar to that for Theorem 4.2).

Suppose (4.16) has the following interval of solutions parameterized by α:

$$(R_1, R_2) = \alpha(R_1^*, R_2^*), \quad 0 < \alpha_1 \leq \alpha \leq \alpha_2,$$

whereby we assume at the moment that R_1^* and R_2^* are finite and rationally independent. Then for each α we have

$$1 - 2e^{-\alpha k \frac{R_1^* + R_2^*}{k}} e^{-(\theta_1+\theta_2)} + 4e^{-8\alpha k \frac{R_2^*}{k}} e^{-8\theta_2} = 0,$$

whereby θ_1 and θ_2 depend on α. Applying Theorem 2.4 to this formula, it is clear that for the vanishing delays $(R_1, R_2) = (\frac{R_1^*}{k}, \frac{R_2^*}{k}), k \to \infty$, (4.15) has solutions with real part arbitrarily close to $k\alpha$, $\alpha_1 \leq \alpha \leq \alpha_2$. This leads to the following conclusion: the intervals in Figure 4.2 which form the intersection of a line through the origin and the projected solution surface of (4.16) correspond to intervals of $\bar{Z}$ which move to infinity ($k \to \infty$) when the delays approach zero in the ratio determined by the slope of the line. Second, the number and lengths of these intervals, when existing, depend on the way (with what ratio) the delays approach zero.

Using the same kind of arguments, the intersection of a line through the origin with curve C in Figure 4.2 corresponds to a real root going to infinity. As can be seen from the picture, the *occurrence* of this phenomenon depends on additional dependencies between the delays such as their having a fixed ratio. Whether the same is true for the occurrence of roots with large real part but arbitrary imaginary part remains an open question.

Note that when $\frac{R_1^*}{R_2^*}$ would be rational, the delays are commensurate and $\bar{Z}$ consists of a number of points. However, when the ratio is close[1] to an irrational number (due to the lower semicontinuity of $\bar{Z}$ w.r.t. the delays), these points will fill up the intervals (rationally independent case) quite well. Figure 4.3 shows some of the roots of (4.15) for the delays $(R_1, R_2) = (\frac{1}{k}, \frac{8}{k})$, which correspond to line D in Figure 4.2. For computational convenience these delays are chosen to be rationally dependent, but the resonance is relatively weak: the two intervals predicted by Figure 4.2 are already visible. The two real roots correspond to the intersection of curves D and C in Figure 4.2.

One can remark that when the delays approach zero the imaginary parts of the nonreal roots increase to infinity. As already mentioned, when dealing with practical control problems the question arises whether the damping at very high frequencies is underestimated in the model (and the corresponding exponential polynomial) or not. In any case the large real roots are important. When one can estimate the ratio of the delays in the real system, one can predict whether such unstable behavior occurs, using diagrams like Figure 4.2.

4.5. Degenerate case. We call the exponential polynomial (4.6) degenerate if and only if all solutions with $R \geq 0$ of

$$1 - \sum_{i \in I} a_i e^{-\gamma_i \cdot R} e^{-i\gamma_i \cdot \theta} = 0 \tag{4.17}$$

satisfy

$$\gamma_i \cdot R = 0 \text{ or } \gamma_i \cdot R = +\infty \quad \forall i \in I.$$

[1] A (positive) rational number a is close to an irrational number if and only if $a = \frac{N_1}{N_2}$, whereby N_1 and $N_2 \in \mathbb{N}$ are large and coprime.

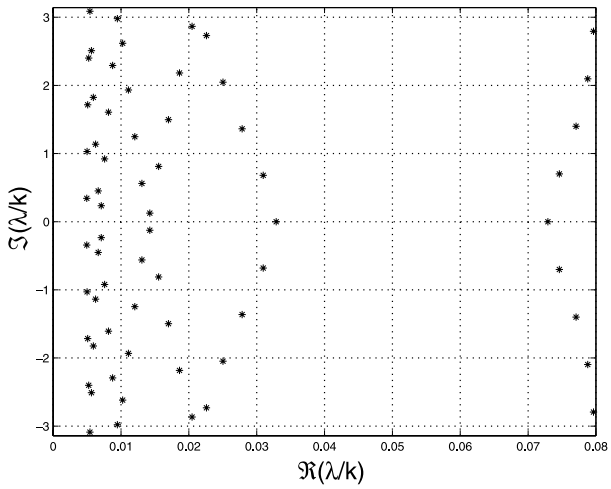


FIG. 4.3. *Part of the spectrum of* (4.15) *for the delays* $(r_1, r_2) = (\frac{1}{k}, \frac{8}{k})$*, corresponding to the dashed line* D *in Figure* 4.2.

For instance,

$$1 - \sum_{i=1}^{N} a_i \left(e^{-\lambda . r}\right)^i - \sum_{j \in J} b_j e^{-\lambda(\gamma_j r + \nu_j \cdot s)} = 0$$

is degenerate if and only if the polynomial $1 - \sum_{i=1}^{N} a_i x^i$ has some roots on, and all other roots outside, the unit circle. For $N = 2$, the equation $1 - a_1 e^{-\lambda r_1} - a_2 e^{-\lambda r_2} - \sum_{j \in J} b_j e^{-\lambda(\gamma_j \cdot r + \nu_j \cdot s)} = 0$ is degenerate if and only if $|a_1| + |a_2| = 1$.

Note that we are now in the case where statement (b) is not satisfied in Theorems 4.2 and 4.3. By means of an example, we show that in such a situation the large-delay part plays a role in the existence of arbitrarily large characteristic roots for vanishing delays. Therefore, consider first

$$H(\lambda, h) \triangleq 1 + e^{-h\lambda} + e^{-2\lambda} - e^{-(2+h)\lambda} = 0, \tag{4.18}$$

whose "small-delay" part is degenerate. With

$$h_n = \tfrac{1}{n+\frac{1}{2}}, \qquad d_n = (n + \tfrac{1}{2})\pi,$$

we have

$$H(c + id_n, h_n) = 0 \Leftrightarrow e^{-ch_n} = \tanh(c);$$

hence there are roots with arbitrarily large real part as $n \to \infty$. However, when one modifies (4.18) to

$$1 + e^{-h\lambda} + e^{-2\lambda} + e^{-(2+h)\lambda} = 0, \tag{4.19}$$

degeneration is preserved while there are *no* roots with arbitrarily large real part since (4.19) can be factored into $(1 + e^{-h\lambda})(1 + e^{-2\lambda})$.

5. Largest delays coincide. In the previous section we showed that in the presence of both normal and arbitrarily small delays there may exist roots with arbitrarily large real part. An analogous phenomenon occurs when the largest delays are arbitrarily close together. Then roots may exist with real part moving to $-\infty$. From a control point of view this phenomenon is less important (no stability problem), and therefore we give only an illustrative example. Sufficient and necessary conditions can be found in [18].

In the characteristic equation,

$$1 + e^{-\lambda} - 3e^{-\lambda 2} + 2e^{-\lambda(2+h)} = 0, \tag{5.1}$$

the largest delays 2 and $2 + h$ coincide as $h \to 0$. Multiplying this equation by $e^{\lambda 2}$ yields

$$e^{\lambda 2} + e^{\lambda} - 3 + 2e^{-\lambda h} = 0,$$

and the roots of $-3 + 2e^{-\lambda h}$,

$$\lambda = \frac{-\log(3/2) + i2\pi l}{h}, \quad l \in \mathbb{N},$$

approximate corresponding roots of (5.1) as $h \to 0+$.

6. Applications and examples. We illustrate the results obtained in this paper by means of two examples.

6.1. Boundary controlled PDE. This example and the phenomena which occur are also described in [5, 16]. Consider

$$w_{xx} = w_{tt}, \quad 0 \le t < \infty, \quad 0 \le x \le 1, \tag{6.1}$$

subject to the boundary conditions

$$\begin{cases} w(0,t) = 0, \\ w_x(1,t) = -kw_t(1,t-h), \end{cases} \tag{6.2}$$

where $h \ge 0$, $k > 0$. Formulae (6.1) and (6.2) describe the transversal movement of a beam clamped at one side and stabilized by applying a force at the other side; h represents a small delay in the velocity feedback. Substituting a solution of the form $w(x,t) = e^{\lambda t}z(x)$ into (6.1) and taking the boundary conditions into account, the following characteristic equation is obtained:

$$e^{\lambda h} + k\tanh(\lambda) = 0, \tag{6.3}$$

which can be rewritten as

$$1 + e^{-\lambda 2} + ke^{-\lambda h} - ke^{-\lambda(2+h)} = 0. \tag{6.4}$$

Remark. The damped wave equation $\bar{w}_{tt} - \bar{w}_{xx} + 2a\bar{w}_t + a^2\bar{w} = 0$ with boundary conditions $\bar{w}(0,t) = 0$ and $\bar{w}_x(1,t) = -\bar{k}\bar{w}_t(1,t-h)$, which was analyzed in [16], can be transformed into (6.1)–(6.2) using the relations $\bar{k} = e^{-ah}k$, $\bar{w} = e^{-at}w$, the latter introducing a shift a of the spectrum.

6.1.1. Analysis of the undelayed case. When $h = 0$, the characteristic roots are

$$\lambda = -\frac{1}{2}\mathrm{Log}\left(\frac{1+k}{1-k}\right) + i\pi l, \quad l \in \mathbb{Z},$$

where Log denotes the principal value of the logarithm. Because $c(0) = \Re(\mathrm{Log}(\frac{1+k}{1-k})) < 0$ for $k > 0$, the undelayed system is stable, i.e., all roots lie in the left half plane. When k approaches 1, the real part of the characteristic roots moves to $-\infty$, which indicates superstability. This can be explained as follows. The general solution of (6.1) can be written as a combination of two travelling waves, a solution $\phi(x-t)$ moving to the right and a solution $\psi(x+t)$ moving to the left. When $k = 1$, $\phi(x-t)$ satisfies the second boundary condition, and thus the reflection coefficient at $x = 1$ is zero; at $x = 0$ the wave $\phi(x+t)$ is reflected completely. Consequently, all perturbations disappear in a finite time (at most 2 time-units).

6.1.2. Analysis for arbitrarily small delays. Equation (6.4) is very interesting from a theoretical point of view. The three delays h, 2, and $2+h$ are functions of only two independent delays, 2 and h. When $h \to 0$ there is an arbitrarily small delay, and the largest delays asymptotically coincide.

When the delays 2 and h are rationally independent, c_{f} is calculated as the rightmost solution α of

$$1 + e^{-2\alpha}e^{-i\theta_1} + ke^{-i\theta_2} - ke^{-2\alpha}e^{-i(\theta_1+\theta_2)} = 0, \tag{6.5}$$

which, after multiplication with $e^{2\alpha}e^{i(\theta_1+\theta_2)}$, can be rewritten as

$$e^{2\alpha}e^{i\theta_1} = \frac{k - e^{i\theta_2}}{k + e^{i\theta_2}}$$

and can be interpreted for each α as the intersection points of two circles in the complex plane. Note that in this case the obtained upper bound will equal the upper bound calculated when all delays are considered independent:

$$|1-k| - e^{-2c_{\mathrm{f}}} - ke^{-2c_{\mathrm{f}}} = 0. \tag{6.6}$$

Indeed, (6.5) is transformed into (6.6) when choosing $\theta_1 = \theta_2 = \pi$ if $k < 1$ and $\theta_1 = 0$, $\theta_2 = \pi$ if $k > 1$. Thus the upper bound c_{f} in the case of rationally independent delays satisfies

$$c_{\mathrm{f}} = \frac{1}{2}\mathrm{Log}\left(\frac{1+k}{1-k}\right). \tag{6.7}$$

Following Theorem 4.2 there are characteristic roots with arbitrarily large real part (for arbitrarily small delays) if $\exists R \in (0, +\infty]$ and $\theta \in [0, 2\pi]$ such that $1 + ke^{-R}e^{-i\theta} = 0$. This is the case when $k > 1$. These roots have large imaginary part (no solution with $\theta = 0$). When $k = 1$ we have a degenerate case: (6.4) becomes (4.18), for which the existence of an infinite spectrum is shown in subsection 4.5.

One can prove that there are roots whose real part moves off to $-\infty$ for vanishing h when $k < 1$ [18].

The obtained results are summarized in Figure 6.1.

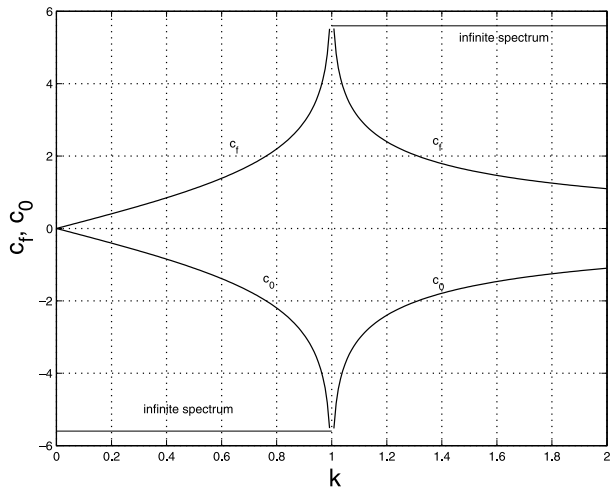


FIG. 6.1. *Upper bounds on the spectrum of* (6.1): $c(0) = c_0$ *is the real part of all characteristic roots in the undelayed case,* c_f *the supremum of the finite part of the spectrum for an arbitrarily small delay,* $h \to 0+$.

6.2. Bifurcation diagram: Parameter dependence. As a last example, we discuss the position of the roots of the characteristic equation

$$1 + ae^{-\lambda h} + be^{-\lambda 2h} \tag{6.8}$$

as a function of the parameters a and b. Note that (3.1) is a special case of (6.8).

First we fix h. The delays h and $2h$ are clearly dependent, and the roots of (6.8) can be calculated from

$$e^{-\lambda h} = \frac{-a \pm \sqrt{a^2 - 4b}}{2b}. \tag{6.9}$$

Hence when $a^2 - 4b < 0$ (> 0) the spectrum is situated on one (two) vertical lines in the complex plane. The positions of these lines depend on the parameters. For $a^2 - 4b > 0$, one can show that such a line crosses the imaginary axis when $a = b + 1$ and when $a = -b - 1$, indicating a change of stability of the line under consideration. When $a^2 - 4b < 0$, the single line crosses the imaginary axis when $b = 1$. The corresponding curves in two-parameter space (a, b) are shown in Figure 6.2.

When h and $2h$ should be considered to be independent, i.e., when these delays are perturbed in such a way that their rational dependency is lost, the upper bound $c(h)$ can be calculated from

$$1 - |a|e^{-c(h)h} - |b|e^{-2c(h)h} = 0,$$

and thus the system is stable when $|a| + |b| \leq 1$. Comparing the dependent with the independent case, one can see that the dangerous area in the parameter space, i.e., the parameter values for which small changes in the delays destroy stability, is enclosed by the triangles $(0, 1), (1, 0), (1, 2)$ and $(0, -1), (1, 0), (1, -2)$ (see Figure 6.2).

Equation (6.8) corresponds to the small-delay part of

$$1 + ae^{-\lambda h} + be^{-\lambda 2h} + de^{-\lambda} = 0 \tag{6.10}$$

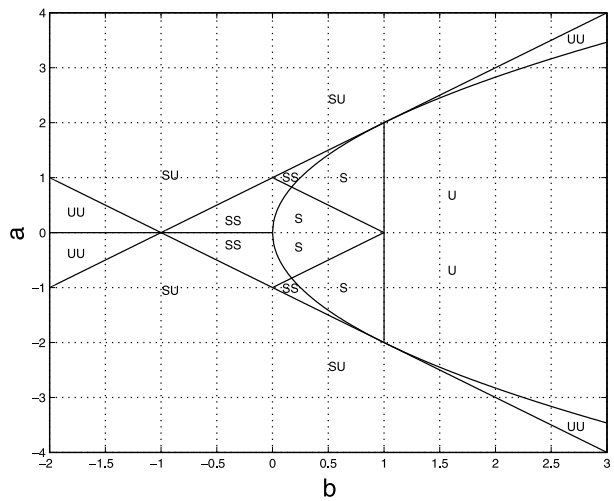


FIG. 6.2. *Stability areas of* (6.8). *When the delays are dependent, "S" denotes a stable vertical line of characteristic roots in the complex plane, and "U" an unstable line. When the delays are independent, the system is stable for parameter values inside the curve* $|a| + |b| = 1$.

as $h \to 0+$. One can show that when the three delays are considered to be independent, (6.10) has roots with real part $\to \infty$ for parameter values a and b outside the curve $|a| + |b| = 1$, and that there are arbitrarily unstable roots with small imaginary part when $b < \frac{a^2}{4}$ if $a \leq -2$ and $b < -a - 1$ if $a \geq -2$. All roots of (6.10) are in the open left half plane when $|a| + |b| < 1$ and $|d| < 1 - |a| - |b|$.

7. Conclusions. Sensitivity of NFDEs to infinitesimal changes of the delays is caused by the behavior of the essential spectrum which is determined by the roots of an exponential polynomial. A remarkable conclusion of the theory developed in [1, 9], concerning the roots of exponential polynomials, is that the supremum of the real parts of the spectrum can change discontinuously w.r.t. the delays, whereas the individual roots move continuously. In the first part of this paper the underlying mechanisms are interpreted and explained by means of spectral plots. For example, when rationally independent delays approach rationally dependent delays, this gives rise to a pointwise but nonuniform convergence of the spectrum, whereby the sensitivity of an individual root increases as its modulus increases. In a second part, we extend the theory developed in [1] to the case where some of the delays are arbitrarily small, which can result in characteristic roots with arbitrarily large real part. Sufficient and necessary conditions are provided. Thereby we also treat the special case of roots with large real part but small imaginary part. We further show that when the small delays are brought to zero in a fixed ratio, the structure of the set of the roots with large real part depends strongly on this ratio. The paper ends with two illustrative examples.

Appendix. We formulate a lemma which is used in the proof of Theorem 4.2. The lemma is a modification of Hurwitz's theorem; see, e.g., [3].

LEMMA A.1. *Let* $f(\lambda)$ *and the sequence* $\{f_n(\lambda)\}_{n\geq 1}$ *be analytic functions on an (open) domain* D. *Suppose that* $\{f_n(\lambda)\}_{n\geq 1}$ *converges uniformly to* $f(\lambda)$ *on the disc* $\{\lambda : |\lambda| \leq R\} \subset D$ *for some* $R > 0$, *and that on this disc* $f(\lambda)$ *only has a zero in* $\lambda = 0$ *with multiplicity* k. *Then there exists a number* $N \in \mathbb{N}$ *such that,* $\forall n \geq N$, $f_n(\lambda)$ *has exactly* k *zeros* $\lambda_{n,1}, \ldots, \lambda_{n,k}$ *in* $|\lambda| \leq R$, *whereby* $\lim_{n\to\infty} \lambda_{n,j} = 0$ $\forall j \in \{1, \ldots, k\}$.

Proof. Consider for some $0 < r \leq R$ the curve $\Gamma : [0, 2\pi] \to \mathbb{C} : t \to \Gamma(t) = re^{it}$. The function $|f(\lambda)|$ attains a minimum M on Γ whereby $M > 0$. Because the sequence of functions $f_n(\lambda)$ is uniformly converging,

$$\exists N \text{ such that } \forall n \geq N : |f_n(\lambda) - f(\lambda)| < M \leq |f(\lambda)| \text{ on } \Gamma.$$

Consequently,

$$\left| 1 - \frac{f_n(\lambda)}{f(\lambda)} \right| < 1$$

on Γ.

For each $n \geq N$ the curve $\gamma(t) = \frac{f_n(re^{it})}{f(re^{it})}, \quad t \in [0, 2\pi]$, satisfies

$$|1 - \gamma_n(t)| < 1,$$

and, because it can be embedded in a closed disc not containing the origin, the winding number of the curve w.r.t. the origin, $n(\gamma_n, 0)$, is zero:

$$n(\gamma_n, 0) = \frac{1}{2\pi i} \int_{\gamma_n} \frac{d\lambda}{\lambda} = \frac{1}{2\pi i} \int_0^{2\pi} \frac{\gamma_n'(t)}{\gamma_n(t)} dt = 0.$$

Using the definition of $\gamma_n(t)$, $\frac{1}{2\pi i} \int_0^{2\pi} \frac{\gamma_n'(t)}{\gamma_n(t)} dt$ can be written as $\frac{1}{2\pi i} \int_\Gamma (\frac{f_n'(\lambda)}{f_n(\lambda)} - \frac{f'(\lambda)}{f(\lambda)}) d\lambda$. This integral is well defined because from the previous it follows that neither f_n nor f are zero in $\{\lambda : r - \epsilon_n < |\lambda| < r + \epsilon_n\}$ for some $\epsilon_n \in \mathbb{R}_0^+$.

But this implies

$$\frac{1}{2\pi i} \int_\Gamma \frac{f'}{f} d\lambda = \frac{1}{2\pi i} \int_\Gamma \frac{f_n'}{f_n} d\lambda \quad \forall n \geq N.$$

Following the theorem of the principle of the argument (see, for example, [22]), these integrals correspond to the zero-pole excess of f and f_n (number of zeros − number of poles) inside Γ and in this case to the number of zeros. When taking $r = R$ the first statement of the lemma is proven. The second statement follows from the fact that r can be chosen arbitrarily small. ∎

REFERENCES

[1] C. Avellar and J. Hale, *On the zeros of exponential polynomials*, J. Math. Anal. Appl., 73 (1980), pp. 434–452.

[2] S. A. Campbell, *Resonant codimension two bifurcation in a neutral functional-differential equation*, Nonlinear Anal., 30 (1997), pp. 4577–4584.

[3] J. Conway, *Functions of One Complex Variable*, Springer-Verlag, New York, 1995.

[4] R. Datko, *Not all feedback stabilized hyperbolic systems are robust with respect to small time delays in their feedbacks*, SIAM J. Control Optim., 26 (1988), pp. 697–713.

[5] R. Datko, *Two examples of ill-posedness with respect to time delays revisited*, IEEE Trans. Automat. Control, 42 (1997), pp. 434–452.

[6] R. Datko, J. Lagnese, and M. Polis, *An example on the effect of time delays in boundary feedback stabilization of wave equations*, SIAM J. Control Optim., 24 (1986), pp. 152–156.

[7] R. Datko and Y. You, *Some second-order vibrating systems cannot tolerate small time delays in their damping*, J. Optim. Theory Appl., 70 (1991), pp. 521–537.

[8] J. Hale and S. Verduyn Lunel, *Introduction to Functional Differential Equations*, Appl. Math. Sci. 99, Springer-Verlag, New York, 1993.

[9] J. K. Hale, *Effects of delays on dynamics*, in Topological Methods in Differential Equations and Inclusions, A. Granas, M. Frigon, and G. Sabidussi, eds., Kluwer Academic Publishers, Norwell, MA, 1995, pp. 191–238.

[10] K. Hannsgen, Y. Renardy, and R. Wheeler, *Effectiveness and robustness with respect to time delays of boundary feedback stabilization in one-dimensional viscoelasticity*, SIAM J. Control Optim., 26 (1988), pp. 1200–1234.

[11] G. Hardy and E. Wright, *An Introduction to the Theory of Numbers*, Oxford University Press, London, UK, 1968.

[12] V. Kolmanovskii and V. Nosov, *Stability of Functional Differential Equations*, Math. Sci. Engrg. 180, Academic Press, San Diego, CA, 1986.

[13] V. B. Kolmanovskii and A. Myshkis, *Introduction to the Theory and Application of Functional Differential Equations*, Math. Appl. 463, Kluwer Acadamic Publishers, Dordrecht, The Netherlands, 1999.

[14] H. Logemann, *Destabilizing effects of small time-delays on feedback-controlled descriptor systems*, Linear Algebra Appl., 272 (1998), pp. 131–153.

[15] H. Logemann and R. Rebarber, *The effect of small time-delays on the closed-loop stability of boundary control systems*, Math. Control Signals Systems, 9 (1996), pp. 123–151.

[16] H. Logemann, R. Rebarber, and G. Weiss, *Conditions for robustness and nonrobustness of the stability of feedback control systems with respect to small delays in the feedback loop*, SIAM J. Control Optim., 34 (1996), pp. 572–600.

[17] H. Logemann and S. Townley, *The effect of small delays in the feedback loop on the stability of neutral systems*, Systems Control Lett., 27 (1996), pp. 267–274.

[18] W. Michiels, K. Engelborghs, and D. Roose, *Sensitivity to Infinitesimal Delays in Neutral Equations*, TW report 286, Department of Computer Science, Katholieke Universiteit Leuven, Leuven, Belgium, 1998.

[19] Ö. Morgül, *On the stabilization and stability robustness against small delays of some damped wave equations*, IEEE Trans. Automat. Control, 40 (1995), pp. 1626–1630.

[20] R. Murray, C. Jacobson, R. Casas, A. Khibnik, C. Johnson, Jr., R. Bitmead, A. Peracchio, and W. Proscia, *System identification for limit cycling systems: a case study for combustion instabilities*, in Proceedings of the American Control Conference, Philadelphia, PA, 1998.

[21] R. Rebarber and S. Townley, *Robustness with respect to delays for exponential stability of distributed parameter systems*, SIAM J. Control Optim., 37 (1999), pp. 230–244.

[22] W. Rudin, *Functional Analysis*, McGraw-Hill, New York, 1973.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 4, pp. 1159–1188

VISCOSITY SOLUTIONS METHODS FOR SINGULAR PERTURBATIONS IN DETERMINISTIC AND STOCHASTIC CONTROL*

OLIVIER ALVAREZ† AND MARTINO BARDI‡

Abstract. Viscosity solutions methods are used to pass to the limit in some penalization problems for first order and second order, degenerate parabolic, Hamilton–Jacobi–Bellman equations. This characterizes the limit of the value functions of singularly perturbed optimal control problems for deterministic systems and for controlled degenerate diffusions. The results apply to cases where the usual order reduction method does not give the correct limit, and to systems with fast state variables depending nonlinearly on the control. Some connections with ergodic control and periodic homogenization are discussed.

Key words. singular perturbations, deterministic optimal control, stochastic optimal control, nonlinear systems, order reduction, viscosity solutions, Hamilton–Jacobi–Bellman equations, penalization, periodic homogenization, ergodic control, state constraints

AMS subject classifications. 49L25, 35B25, 93C73, 93E20, 35B27

PII. S0363012900366741

Introduction. In this paper we study three penalization problems for fully nonlinear partial differential equations motivated by the optimal control theory for systems with different time scales. In all the problems the limit PDE is of lower dimension, and the limit operator is not obvious to guess. Problems of this kind were first studied by Jensen and Lions [29] for classic solutions of quasilinear uniformly elliptic PDEs. Here we study a first order Hamilton–Jacobi (H–J) equation and a degenerate parabolic, fully nonlinear, Hamilton–Jacobi–Bellman (H–J–B) equation in the framework of viscosity solutions.

The first problem we consider is the limit as $\varepsilon \to 0+$ of

$$-\partial_t u_\varepsilon + H\left(x, y, D_x u_\varepsilon, \frac{D_y u_\varepsilon}{\varepsilon}\right) = 0 \quad \text{in } (0,T)\times\mathbb{R}^n \times Y \tag{1}$$

for the Hamiltonian

$$H(x,y,p,q) = \max_{\alpha\in A}\{-(p, f(x,y,\alpha)) - (q, g(x,y,\alpha)) - l(x,y,\alpha)\},$$

where $(\cdot,\cdot)$ denotes the scalar product, and $Y \subseteq \mathbb{R}^m$ is open, bounded, connected, and smooth. This is the H–J–B equation associated via dynamic programming with the minimization of the functional

$$J(t,x,y,\alpha.) := \int_t^T l(x_s, y_s, \alpha_s)\, ds + h(x_T, y_T)$$

*Received by the editors February 2, 2000; accepted for publication (in revised form) March 26, 2001; published electronically November 28, 2001. This research was done within the TMR Project "Viscosity solutions and their applications" of the European Community.
http://www.siam.org/journals/sicon/40-4/36674.html
†UMR 60-85, Université de Rouen, 76821 Mont-Saint Aignan Cedex, France (alvarez@univ-rouen.fr).
‡Dipartimento di Matematica P. e A., Università di Padova, Via Belzoni 7, 35131 Padova, Italy (bardi@math.unipd.it). The research of the second author was partially supported by M.U.R.S.T. project "Analisi e controllo di equazioni di evoluzione deterministiche e stocastiche."

on the trajectories of the system

$$\dot{x}_s = f(x_s, y_s, \alpha_s), \qquad \dot{y}_s = \frac{1}{\varepsilon} g(x_s, y_s, \alpha_s), \tag{2}$$

with $x_t = x$, $y_t = y$, where $\alpha_\cdot$ is the control function subject to $\alpha_s \in A$. Singular perturbation problems for deterministic controlled systems were studied by many authors; see, e.g., the books by Kokotović, Khalil, and O'Reilly [32], Bensoussan [12], and Dontchev and Zolezzi [18], as well as the recent articles by Artstein and Gaitsgory [22, 4, 3, 5], Veliov [40], Subbotina [39], Bagagiolo and Bardi [6], and the references therein. We recall that the theory of singular perturbations has many important applications, in particular to the order reduction of large scale systems.

As in [6] (see also [9, 7]) we assume the Hamiltonian H to be coercive in the $q = D_y u$ variables, which amounts to the complete controllability of the fast variables y of the system, and consider the boundary condition on ∂Y corresponding to the state-space constraint on the fast variables

$$y_s \in \overline{Y} \quad \text{for all } t \le s \le T.$$

In [6] a separability assumption on the controls acting on the fast and the slow variables yielded a simple explicit formula for the Hamiltonian $\overline{H}$ of the limit PDE. Here we show that in general the limit Hamiltonian $\overline{H} = \overline{H}(x,p)$ is the unique constant such that the boundary value problem

$$H(x, y, p, D_y\chi) \ge \overline{H} \quad \text{in } \overline{Y}, \qquad H(x, y, p, D_y\chi) \le \overline{H} \quad \text{in } Y$$

has a viscosity solution $\chi = \chi(y)$ for fixed (x,p). The existence and uniqueness of the *effective Hamiltonian* $\overline{H}$ was proved by Capuzzo-Dolcetta and Lions [16] in connection with ergodic control problems. We prove that the viscosity solution $u_\varepsilon(t,x,y)$ of (1) with constrained boundary conditions on ∂Y and the terminal condition

$$u_\varepsilon(T, x, y) = h(x, y)$$

converges uniformly as $\varepsilon \to 0+$ to the viscosity solution $u = u(t,x)$ of

$$-\partial_t u + \overline{H}(x, Du) = 0 \text{ in } (0,T) \times \mathbb{R}^n \quad \text{and} \quad u(T,x) = \underline{h}(x) := \inf_y h(x,y) \text{ for } x \in \mathbb{R}^n.$$

The effective Hamiltonian $\overline{H}$ admits a representation as the long time limit of the value function of a control problem in $\overline{Y} \subseteq \mathbb{R}^m$; see [16, 9]. This formula shows the strong connection between our result and the recent work of Artstein and Gaitsgory [5], even if they do not consider state constraints, make somewhat different assumptions, and use completely different methods. We also give a new representation of $\overline{H}$ as the Bellman Hamiltonian associated to a suitable set of "limiting" relaxed controls. This provides an interpretation of the limit u as the value function of an optimal control problem with n-dimensional state space, which is therefore the appropriate limit of the previous problem for the $(n+m)$-dimensional system (2) as $\varepsilon \to 0+$. One might guess from (2) that in the limit the fast variables satisfy $g(x_s, y_s, \alpha_s) \equiv 0$ and the Hamiltonian becomes

$$H_0(x,p) := \sup_{\{(\alpha,y):\, g(x,y,\alpha)=0\}} \{-(p, f(x,y,\alpha)) - l(x,y,\alpha)\}.$$

This is indeed the case in many classic problems [32, 12], and we give some examples where $\overline{H} = H_0$. In general, however, $H_0(x,p) \le \overline{H}(x,p)$ and the inequality can be

strict when the fast variables oscillate very rapidly in the limit. In this case we can pass to the limit because an averaging phenomenon occurs; this was studied, for instance, in [22, 4, 5, 41], and see the references therein for earlier literature on averaging in ordinary differential equations. Our main new contribution is a PDE approach to the problem, where the theory of viscosity solutions provides many useful tools: the characterization of the effective Hamiltonian, the perturbed test function method of Evans [19, 20], and the relaxed semilimits of Barles and Perthame [11, 9] that we slightly modify here.

The viscosity solutions methods allow us to treat our second and third problem in a very similar way. They are the limits as $\varepsilon \to 0+$ of

$$-\partial_t u_\varepsilon + H\left(x, y, D_x u_\varepsilon, \frac{D_y u_\varepsilon}{\varepsilon}, D^2_{xx} u_\varepsilon, \frac{D^2_{yy} u_\varepsilon}{\varepsilon^{2\gamma}}, \frac{D^2_{xy} u_\varepsilon}{\varepsilon^{\gamma}}\right) = 0 \quad \text{in } (0,T)\times\mathbb{R}^n \times Y$$

for $\gamma = \frac{1}{2}$ and $\gamma = 1$, respectively, where

$$H(x,y,p,q,X,Y,Z) = \sup_{\alpha\in A}\left\{-\frac{1}{2}[\operatorname{tr}(\sigma\sigma^T X) + \operatorname{tr}(\tau\tau^T Y) + \operatorname{tr}(\tau\sigma^T Z) + \operatorname{tr}(Z\tau\sigma^T)]\right.$$
$$\left. -(p, f(x,y,\alpha)) - (q, g(x,y,\alpha)) - l(x,y,\alpha)\right\},$$

and the coefficients σ, τ, f, g, l are functions of (x, y, α). This is the fully nonlinear, degenerate parabolic, H–J–B equation arising in the minimization of the expectation $E\,J(t,x,y,\alpha)$ for the singularly perturbed controlled degenerate diffusion process

$$\begin{aligned} dx_s &= f(x_s, y_s, \alpha_s)\,ds + \sigma(x_s, y_s, \alpha_s)\,dW_s, \\ dy_s &= \frac{1}{\varepsilon} g(x_s, y_s, \alpha_s)\,ds + \frac{1}{\varepsilon^\gamma}\tau(x_s, y_s, \alpha_s)\,dW_s. \end{aligned} \tag{3}$$

Problems of this nature with $\gamma = 1/2$ can be found in [12] for dispersion matrices σ, τ that are not degenerate and are independent of α, and in the book by Kushner [33] for possibly degenerate diffusions that are still independent of α and for uncontrolled fast drift g. The more recent papers by Kabanov and colleagues [31, 30], for $\gamma < 1/2$ and $\gamma = 1/2$, respectively, allow the fast drift to depend linearly on the control, whereas both dispersion matrices are uncontrolled. Here we allow all the terms of the fast dynamics to depend nonlinearly on the control. On the other hand, we limit ourselves to the case of fast variables constrained on an m-dimensional torus, that is, $Y = [0,1]^m$, all the data are 1-periodic in the y variable, and periodic boundary conditions are imposed on ∂Y. In this case the existence, uniqueness, and representation of the effective (second order) Hamiltonian can be taken from a recent article by Arisawa and Lions [2], and the simplicity of the boundary conditions reduces the technical difficulties of the proof. Moreover, we do not try here to represent the solution of the limit problem as a value function, nor to prove the convergence of nearly optimal controls. Among other technical conditions, we suppose that the terminal cost h is independent of y and make some mild restrictions on the slow dynamics. Concerning assumptions about fast dynamics, we shall make the one introduced by Arisawa and Lions [2] that guarantees some averaging behavior (ergodicity) in the fast dynamics.

Section 2 is devoted to the classic scaling of (3) corresponding to $\gamma = 1/2$. Our results include the following three cases as well as several combinations.

- The diffusions (in the fast and slow variables) are uniformly nondegenerate.
- The problem is deterministic and the dynamical system in the fast variable is controllable.
- The system in the fast variable is independent of x and y and satisfies the nonresonance condition of [2] (see section 2 for a precise statement).

The second case is the one covered by the first section in the state constraint setting. We remark that we have to make some nontrivial modifications to the perturbed test function method of Evans [20] in order to avoid the crucial assumption that the Hamiltonian be uniformly continuous in all variables. For control problems this corresponds to the restrictive condition that the dynamics do not depend on the state variables.

Section 3 discusses the less usual case $\gamma = 1$. Our motivation for studying this scaling is homogenization. For this problem, the dynamical system is

$$dx_s = f\left(x_s, \frac{x_s}{\varepsilon}, \alpha_s\right) ds + \sigma\left(x_s, \frac{x_s}{\varepsilon}, \alpha_s\right) dW_s.$$

This corresponds to the singular perturbation problem for the artificial fast variable $y_s = x_s/\varepsilon$ with $\gamma = 1$, $g = f$, and $\tau = \sigma$. For homogenization, our assumption is one of the following:

- The diffusion is uniformly nondegenerate.
- The problem is deterministic and controllable.
- The system is purely stochastic ($f \equiv 0$), is independent of x and y, and satisfies the nonresonance condition.

In this generality, the results for the stochastic problems seem to be new. They extend in various ways previous work on periodic homogenization for uniformly elliptic, nondivergence form, quasilinear equations by Bensoussan, Boccardo, and Murat [13] and Evans [19, 20], and for first order H–J equations by Lions, Papanicolaou, and Varadhan [35] and Evans [20]. See also the additional references on the viscosity solutions approach to homogenization in section 3.

The main goal of this paper is to illustrate a unified PDE approach to singular perturbations for deterministic and stochastic systems, and we do not pursue the minimal assumptions. We believe our method works for several other problems such as, for instance, deterministic systems under weaker controllability assumptions or with state constraints on the slow variables x as well, and stochastic systems with fast variables subject to more general state constraints or governed by a diffusion reflected on ∂Y (giving raise to Neumann boundary conditions). We will come back to some of these problems in future papers.

The first application of viscosity solutions methods to singular perturbation problems in control goes back to Lions [34], and more references can be found in [6]. To our knowledge the present paper is the first using these methods for the second order PDEs associated with controlled diffusion processes.

1. Deterministic control with state constraints on the fast variables.

1.1. The ε-problem. Let $T > 0$ be fixed. For every $\varepsilon > 0$, we consider the control problem in $(0, T] \times \mathbb{R}^n \times \mathbb{R}^m$ with dynamics

$$\dot{x}_s = f(x_s, y_s, \alpha_s), \qquad \dot{y}_s = \frac{1}{\varepsilon} g(x_s, y_s, \alpha_s)$$

for $s \geq t$, with $x_t = x$, $y_t = y$, and the constraint on the fast variables

$$y_s \in \overline{Y} \quad \text{for all } t \leq s \leq T,$$

where $\overline{Y} \subseteq \mathbb{R}^m$ is a given compact set. The control functions are measurable $\alpha : (0,T] \to A$, where A is a compact set, such that the corresponding trajectory satisfies the state constraint; we denote this set with $\mathcal{A}_{(x,y)}$. The value function is defined on $(0,T] \times \mathbb{R}^n \times \overline{Y}$ by

$$u_\varepsilon(t,x,y) = \inf_{\alpha \in \mathcal{A}_{(x,y)}} \left\{ \int_t^T l(x_s, y_s, \alpha_s)\, ds + h(x_T, y_T) \right\}.$$

Now we begin to list the hypotheses of section 1. The list will end at the beginning of the next subsection.

— A is a compact metric space.

— $Y \subseteq \mathbb{R}^m$ is a bounded connected open set with Lipschitz boundary in the following sense: there exist $\eta : \overline{Y} \to \mathbb{R}^m$ bounded and uniformly continuous and $c > 0$ such that

$$B(y + t\eta(y), ct) \subseteq Y \quad \text{for all } y \in \overline{Y},\ 0 < t \le c. \tag{4}$$

— The functions f, g, l, and h are continuous and bounded.

— The functions f and g are Lipschitz continuous in (x,y) uniformly in α; the functions l and h are uniformly continuous in (x,y) uniformly in α.

We set

$$\underline{h}(x) = \inf_y h(x,y).$$

It is easy to see that under the preceding hypotheses $\underline{h}$ is uniformly continuous and bounded.

We continue the list of hypotheses.

— The problem is controllable in y, i.e., there exists $r > 0$ such that

$$B(0,r) \subset \overline{\text{conv}}\{g(x,y,\alpha) \mid \alpha \in A\}.$$

— $\mathcal{A}_{(x,y)} \neq \emptyset$ for all $y \in \overline{Y}$, and u_ε is continuous in $(0,T] \times \mathbb{R}^n \times \overline{Y}$.

Remark. The last assumption is not a consequence of the previous hypotheses on the data if the boundary of Y has corners and $g(x,y,A)$ is not convex, as it is easy to see on simple examples. However, it is automatically satisfied if the boundary is smooth, say C^2, by a result of Soner (see [37] or sect. IV.5 of [9]), based on the "interior field condition"

$$\min_{a \in A} g(x,y,a) \cdot n(y) < 0 \quad \text{for all } y \in \partial Y,\ x \in \mathbb{R}^n,$$

where $n(y)$ is the exterior normal to Y at y, which holds in our case because of the controllability assumption on the fast variables. A more general sufficient condition for the continuity of u_ε that allows for piecewise smooth ∂Y is the following:

$$\overline{Y} = \{y \in \mathbb{R}^m \mid g_i(y) \le 0 \quad \text{for all } i = 1, \dots, p\}$$

for some $g_i \in C^{1,1}(\mathbb{R}^m)$ with $|Dg_i| > 0$, $i = 1, \dots, p$, and

$$\min_{a \in A} \max_{\{i \mid g_i(y) = 0\}} g(x,y,a) \cdot Dg_i(y) < 0 \quad \text{for all } y \in \partial Y,\ x \in \mathbb{R}^n. \tag{5}$$

This is proved in Theorem A.1 of [6]. Note that (5) is automatically satisfied if $g(x,y,A)$ is convex, in addition to the controllability assumption. In the general case

of merely Lipschitz ∂Y, a suitable formulation of the interior field condition can be found in the paper of Ishii and Koike [27], where the continuity of the value function is proved for the infinite horizon problem.

THEOREM 1. *The value function u_ε is the unique viscosity solution in $BC((0,T]\times\mathbb{R}^n\times\overline{Y})$ of the terminal-boundary value problem*

$$\begin{cases} -\partial_t u_\varepsilon + H(x,y,D_x u_\varepsilon, \frac{D_y u_\varepsilon}{\varepsilon}) \geq 0 & \text{in } (0,T)\times\mathbb{R}^n\times\overline{Y}, \\ -\partial_t u_\varepsilon + H(x,y,D_x u_\varepsilon, \frac{D_y u_\varepsilon}{\varepsilon}) \leq 0 & \text{in } (0,T)\times\mathbb{R}^n\times Y, \\ u(T,\cdot) = h \quad \text{in } \mathbb{R}^n\times\overline{Y}, \end{cases} \tag{6}$$

for the Hamiltonian

$$H(x,y,p,q) = \max_{\alpha\in A}\{-(p,f(x,y,\alpha)) - (q,g(x,y,\alpha)) - l(x,y,\alpha)\}.$$

Proof. The fact that a continuous value function satisfies the appropriate H–J–B equation in the viscosity sense is a standard consequence of the dynamic programming principle; for the boundary condition on ∂Y due to the state constraint, see [37] or [9]. The uniqueness can be proved by combining the proof of Theorem III.3.7 in [9] with Soner's argument on ∂Y (see [37] or sect. IV.5 of [9]). □

1.2. The effective Hamiltonian and the limit problem. We introduce the auxiliary m-dimensional system with the same vector field as the fast variables but with $\varepsilon = 1$ and frozen x

$$\dot{y}_s = g(x,y_s,\alpha_s), \qquad y_0 = y,$$

and denote with $\mathcal{A}_y = \mathcal{A}_y^x$ the set of measurable $\alpha : (0,T] \to A$ such that $y_s \in \overline{Y}$ for all $s \geq 0$. In this subsection the notation y_s will always be used for trajectories of this system for some admissible $\alpha \in \mathcal{A}_y$ and fixed x. Our last hypothesis is the following.

— For all $y \in \partial Y$, $x \in \mathbb{R}^n$, and $\varepsilon > 0$ there are controls in $\mathcal{A}_y$ such that the corresponding trajectory satisfies $y_\varepsilon \in Y$ and $y_{-\varepsilon} \in Y$, respectively.

Remark. This assumption follows immediately from the controllability hypothesis if ∂Y is smooth or if $g(x,y,A)$ is convex.

THEOREM 2. *For fixed $(\overline{x},\overline{p})$ there exists a unique constant $\lambda = \overline{H}(\overline{x},\overline{p})$ such that the problem*

$$H(\overline{x},y,\overline{p},D_y\chi) \geq \lambda \quad \text{in } \overline{Y}, \qquad H(\overline{x},y,\overline{p},D_y\chi) \leq \lambda \quad \text{in } Y \tag{7}$$

has a Lipschitz continuous solution χ. Moreover, $\overline{H}(\overline{x},\overline{p}) = \lim_{\delta\to 0+}\delta w_{\delta,\overline{x},\overline{p}}(y)$ uniformly in y, where

$$w_{\delta,x,p}(y) := \sup_{\alpha\in\mathcal{A}_y}\left\{\int_0^{+\infty} e^{-\delta s}\big(-(p,f(x,y_s,\alpha_s)) - l(x,y_s,\alpha_s)\big)\,ds \;\middle|\; \dot{y}_s = g(x,y_s,\alpha_s), y_0 = y\right\},$$

and $Lip(\chi) \leq Cr^{-1}(||l||_\infty + |\overline{p}|\,||f||_\infty)$, where C depends only on the set Y and r is the radius appearing in the controllability assumption.

For the proof we need the following property of sets with Lipschitz boundary.

LEMMA 3. *Let Y be a bounded connected open set with Lipschitz boundary, i.e.,* (4) *holds, and define*

$$d(y,\overline{y}) := \inf\{length(\gamma)|\,\gamma \subseteq Y \text{ polygonal with endpoints } y \text{ and } \overline{y}\}.$$

Then

(i) *there exist $\rho, C > 0$ such that for all $z \in \overline{Y}$*

$$d(y,\overline{y}) \le C|y-\overline{y}| \qquad \text{for all } y,\overline{y} \in Y \cap B(z,\rho);$$

(ii) *there exists $M > 0$ such that all $y,\overline{y} \in \overline{Y}$ can be joined by a polygonal $\gamma \subset \overline{Y}$ with $length(\gamma) \le M$.*

Proof. Since Y is open and connected, d is finite for all $y,\overline{y} \in Y$. Let us first consider the case $z \in \partial Y$. By Proposition A.2 and Remark A.3 in [10] there exist $r', L > 0$ independent of z such that, in $B(z,2r')$, Y is the epigraph of a Lipschitz function defined on the hyperplane orthogonal to a vector ξ and with Lipschitz constant bounded by L. Then there is k independent of z such that

$$B(y+t\xi, kt) \subseteq Y \qquad \text{for all } 0 < t < k,\ y \in \overline{Y} \cap B(z,r').$$

If we set $v = y-\overline{y}/|y-\overline{y}|$, the segments $\{y+t\xi-tkv|\,0<t<k\}$ and $\{\overline{y}+t\xi+tkv|\,0<t<k\}$ lie in Y, and they intersect for $t = |y-\overline{y}|/2k$, which is acceptable for $|y-\overline{y}| < 2k^2$. Thus, for $\rho = r' \wedge 2k^2$, two points $y,\overline{y} \in \overline{Y} \cap B(z,\rho)$ can be joined by a polygonal of length $|y-\overline{y}|/k$ and lying in Y, except possibly for the endpoints. This proves (i) in a neighborhood of ∂Y, and (i) is trivial in the complement of this neighborhood.

To prove (ii), we redefine d by allowing polygonals $\gamma \subseteq \overline{Y}$. Since ∂Y is Lipschitz, $d(y,\overline{y})$ is finite for all $y,\overline{y} \in \overline{Y}$, and it is a metric on $\overline{Y}$. Moreover, it is locally uniformly equivalent to the Euclidean metric, because it satisfies (i) for all $y,\overline{y} \in \overline{Y} \cap B(z,\rho)$, whereas the inequality $|y-\overline{y}| \le d(y,\overline{y})$ is trivial. Then the topology induced by d is equivalent to the usual one, so $\overline{Y}$ is compact for this topology and therefore it is bounded for the metric d, which gives the desired conclusion. □

Proof of Theorem 2. The proof is essentially the same as that of Theorem VIII.1 in [16] or Theorem VII.1.1 in [9], once we prove that $w_\delta(\cdot) := w_{\delta,x,p}(\cdot)$ is Lipschitz with constant independent of δ. In fact, this implies that δw_δ converges uniformly to a constant λ as $\delta \to 0+$ and, for a fixed $y^* \in \overline{Y}$, $w_\delta(y) - w_\delta(y^*) \to \chi(y)$ uniformly, at least along a subsequence. By a standard viscosity solutions argument the pair (χ,λ) satisfies (7). An argument based on the comparison principle for constrained viscosity solutions shows the uniqueness of the constant λ in (7), so $\lambda = \lim_{\delta\to 0+} w_\delta(y)$.

To prove the Lipschitz estimate for w_δ we claim that for some $\rho > 0$, for all $z \in \overline{Y}$, $y,\overline{y} \in \overline{Y} \cap B(z,\rho)$, $\varepsilon > 0$, there exists a control in $\mathcal{A}_y$ such that

$$y_s = \overline{y}, \qquad s \le \frac{C}{r}|y-\overline{y}| + \varepsilon,$$

where C depends only on the set Y and r is the radius appearing in the controllability assumption. Then a simple argument yields

$$|w_\delta(y) - w_\delta(\overline{y})| \le \frac{C}{r}(||l||_\infty + |\overline{p}|\,||f||_\infty)|y-\overline{y}|, \tag{8}$$

as in Proposition III.2.3 of [9].

To prove the claim we first consider $y, \overline{y} \in Y \cap B(z,\rho)$, where ρ comes from Lemma 3(i), so that there is a polygonal $\gamma \subseteq Y$ of length $\sigma \le C|y-\overline{y}| + \varepsilon/2$ joining y and $\overline{y}$. By the controllability assumption and standard results in control theory, γ can be approximated uniformly by trajectories $y_\cdot$ in Y with speed r and $y_0 = y$. Then for any $\tau > 0$ there is a control in $\mathcal{A}_y$ such that $|y_{\sigma/r} - \overline{y}| \le \tau$, and by using the controllability again we build a trajectory in Y joining y and $\overline{y}$ in a time $s \le \sigma/r + \tau C'/r \le C|y-\overline{y}|/r + \varepsilon$, which proves the claim in this case.

If either y or $\overline{y}$ belongs to ∂Y, we use the assumption of this subsection to move y forward to some $y' \in Y$ and $\overline{y}$ backward to $\overline{y}' \in Y$, spending a time 2ε, then piece together these trajectories with the one previously built joining y' and $\overline{y}'$ and complete the proof of the claim.

Finally, the estimate on the Lipschitz constant of χ is obtained by letting $\delta \to 0+$ in (8). □

Remark. The effective Hamiltonian $-\overline{H}(x,p)$ is the optimal average cost of an ergodic control problem in the y variable, i.e., it satisfies the formula

$$\overline{H}(x,p) = \sup_{\alpha\in\mathcal{A}_y} \limsup_{t\to+\infty} \left\{ -\frac{1}{t}\int_0^t \big((p, f(x,y_s,\alpha_s)) + l(x,y_s,\alpha_s)\big)\, ds \right. \tag{9}$$
$$\left. \mid \dot{y}_s = g(x,y_s,\alpha_s), y_0 = y \right\}$$

for all $y \in \overline{Y}$, by Prop. VII.1.3 in [9]. It is also the rescaled limit of the value functions of finite horizon problems as the horizon goes to infinity:

$$\overline{H}(x,p) = \lim_{t\to+\infty} \sup_{\alpha\in\mathcal{A}_y} \left\{ -\frac{1}{t}\int_0^t \big((p, f(x,y_s,\alpha_s)) + l(x,y_s,\alpha_s)\big)\, ds \right. \tag{10}$$
$$\left. \mid \dot{y}_s = g(x,y_s,\alpha_s), y_0 = y \right\}$$

for all $y \in \overline{Y}$, and the convergence is uniform in y as $t \to +\infty$. To see this, consider the value function

$$v(t,y) = \inf_{\alpha\in\mathcal{A}_y} \left\{ \int_0^t \big((p, f(x,y_s,\alpha_s)) + l(x,y_s,\alpha_s)\big)\, ds \mid \dot{y}_s = g(x,y_s,\alpha_s), y_0 = y \right\}.$$

It solves the H–J equation

$$\begin{aligned} \partial_t v + H(\overline{x}, y, \overline{p}, D_y v) &\ge 0 \quad \text{in } (0,+\infty)\times\overline{Y}, \\ \partial_t v + H(\overline{x}, y, \overline{p}, D_y v) &\le 0 \quad \text{in } (0,+\infty)\times Y, \end{aligned}$$

with the initial condition $v(0,\cdot) \equiv 0$. By Theorem 2, the function $\chi(y) - t\overline{H}(\overline{x},\overline{p})$ is a solution of the same Cauchy problem but with a different initial condition. The comparison principle implies that $v(t,y) - \chi(y) + t\overline{H}(\overline{x},\overline{p})$ is bounded, so that $v(t,y)/t \to -\overline{H}(\overline{x},\overline{p})$ as $t \to +\infty$, uniformly in y. See also Theorem VIII.1 in [16] and Exercise VII.1.1 in [9].

Our next result gives the regularity of the effective Hamiltonian.

PROPOSITION 4. *The effective Hamiltonian $\overline{H}$ has the following properties:*

(i) *For all x, p*

$$\inf_y \inf_q H(x,y,p,q) \le \overline{H}(x,p) \le \sup_y H(x,y,p,0);$$

(ii) $\overline{H}$ *is convex in* p;
(iii) *for all* x, p, p'

$$|\overline{H}(x,p) - \overline{H}(x,p')| \le ||f||_\infty |p - p'|;$$

(iv) *for all* x, x', p

$$|\overline{H}(x',p) - \overline{H}(x,p)| \le Lip(f)|p|\,|x'-x| + \omega_l(|x'-x|) + Lip(g)C\frac{||l||_\infty + |p|\,||f||_\infty}{r}|x'-x|,$$

where ω_l *is the modulus of continuity of* l *with respect to* x *and* C *depends only on* Y;
(v) $\overline{H}$ *is uniformly continuous on* $\mathbb{R}^n \times B(0,R)$ *for all* $R > 0$.

The proof is essentially the same as that of Proposition 3 in [1]. The bound for the Lipschitz continuity in x follows easily from the bound on $Lip(\chi)$ of Theorem 2.

The last result of this subsection gives the solution of the limit problem.

PROPOSITION 5. *There exists a unique viscosity solution in* $BC((0,T] \times \mathbb{R}^n)$ *of*

$$-\partial_t u + \overline{H}(x, Du) = 0 \text{ in } (0,T) \times \mathbb{R}^n \quad \text{and} \quad u(T,\cdot) = \underline{h} \text{ on } \mathbb{R}^n. \tag{11}$$

Proof. In view of the regularity of $\overline{H}$, in particular of

$$|\overline{H}(x',p) - \overline{H}(x,p)| \le C|p|\,|x'-x| + \omega(|x'-x|),$$

where ω is a modulus, a comparison theorem holds for (11); see, e.g., Theorem III.3.7 and Exercise V.1.5 in [9]. The existence can be proved by the Perron–Ishii method (see, e.g., section V.2.2 of [9]), or by representing the effective Hamiltonian $\overline{H}$ as the Bellman Hamiltonian of a control problem and then proving that the value function of such a problem solves (11). □

1.3. An example: Separated controls. In [6] the controls acting on the slow and the fast variables are separated, in the sense that $\alpha = (\beta, \gamma) \in B \times C$ with

$$f = f(x,y,\beta), \quad l = l(x,y,\beta), \quad g = g(x,y,\gamma).$$

In this case the Hamiltonian of the ε-problem is

$$H(x,y,p,q) = H_1(x,y,p) + H_2(x,y,q),$$

$$H_1(x,y,p) := \max_{\beta\in B}\{-(p, f(x,y,\beta)) - l(x,y,\beta)\}, \quad H_2(x,y,q) = \sup_{\gamma\in C}\{-(q, g(x,y,\gamma))\},$$

and we expect that the effective Hamiltonian will be

$$\overline{H}(x,p) = \max_{y\in\overline{Y}} H_1(x,y,p). \tag{12}$$

In fact, from the representation (10) of $\overline{H}$ we get, for all $y \in \overline{Y}$,

$$\overline{H}(x,p) = \lim_{t\to+\infty} \sup_{\gamma_\cdot} \left\{ \frac{1}{t}\int_0^t H_1(x, y_s, p)\, ds \mid \dot{y}_s = g(x, y_s, \gamma_s),\ y_0 = y,\ y_s \in \overline{Y} \right\}. \tag{13}$$

This gives immediately $\overline{H}(x,p) \le \max_{y\in\overline{Y}} H_1(x,y,p)$. For the opposite inequality we fix $\overline{y}$ such that $H_1(x,\overline{y},p) = \max_y H_1(x,y,p)$. If there is a control $\overline{\gamma}$ such that $g(x,\overline{y},\overline{\gamma}) = 0$, we choose $y = \overline{y}$ in the right-hand side of (13) and the sup is attained

by $\gamma = \overline{\gamma}$ because the average cost is $H_1(x, \overline{y}, p)$. If such a control $\overline{\gamma}$ does not exist, but $\overline{y} \in Y$, we deduce from the controllability assumption on the fast variables that for any $\varepsilon > 0$ there is a control γ such that $y_0 = \overline{y}$ implies $|y_s - \overline{y}| \le \varepsilon$ for all $s > 0$. The average cost associated with this control is bounded below by $\inf_{B(\overline{y},\varepsilon)\cap\overline{Y}} H_1(x, y, p)$. By taking the limit as $\varepsilon \to 0$, the continuity of H_1 gives $\overline{H}(x, p) \ge H_1(x, \overline{y}, p)$. In the remaining case of $\overline{y} \in \partial Y$ we use the assumption of subsection 1.2 to move the system from $y_0 = \overline{y}$ to $\widetilde{y} \in Y \cap B(\overline{y}, \varepsilon/2)$ in a short time, and then we can use the controllability assumption as before to keep the trajectory in $B(\overline{y}, \varepsilon)$ forever. Therefore we reach the desired inequality as in the previous case.

In conclusion, the representation (12) of the effective Hamiltonian holds under the current hypotheses, and we recover the main result of [6] as a special case of the convergence theorem proved later in this section (under slightly different assumptions).

Remark. In this case the limit problem (11) has a simple control interpretation. In fact its solution is the value function of the problem of minimizing the functional

$$\underline{J}(t, x, \alpha_\cdot, y_\cdot) := \int_t^T l(x_s, y_s, \alpha_s)\, ds + \underline{h}(x_T) \tag{14}$$

on the trajectories of the system

$$\dot{x}_s = f(x_s, y_s, \alpha_s), \qquad x_t = x,$$

with measurable controls $\alpha_\cdot : (0, T] \to A$ and $y_\cdot : (0, T] \to \overline{Y}$. Therefore the fast variables become controls in the limit problem.

1.4. Connections with the order reduction method. If we try to follow the classical Levinson–Tichonov approach to singularly perturbed ordinary differential equations, we have to set formally $\varepsilon = 0$ in the dynamical system and get $g(x_s, y_s, \alpha_s) \equiv 0$. This leads to the conjecture that the limit dynamics are governed by the differential inclusion

$$\dot{x}_s = f(x_s, y_s, \alpha_s), \qquad (y_s, \alpha_s) \in Z(x_s), \tag{15}$$

where

$$Z(x) := \{(y, \alpha) \in \overline{Y} \times A |\, g(x, y, \alpha) = 0\},$$

and of course (15) makes sense if $Z(x_s) \ne \emptyset$ for almost every s. The conjecture turns out to be true in many important problems that can be put in the reduced order form; see, e.g., [32, 12] and the references therein. In this case the limit Hamiltonian is

$$H_0(x, p) := \sup_{(y,\alpha)\in Z(x)} F(x, y, \alpha, p),$$

where

$$F(x, y, \alpha, p) := -(p, f(x, y, \alpha)) - l(x, y, \alpha).$$

LEMMA 6. *Assume in addition that $Z(x) \ne \emptyset$. Then $H_0(x, p) \le \overline{H}(x, p)$ for all $x, p \in \mathbb{R}^n$.*

Proof. Fix (x, p) and $(\overline{\alpha}, \overline{y}) \in Z(x)$. Since $y_s \equiv \overline{y}$ solves $\dot{y}_s = f(x, y_s, \alpha_s)$ for $\alpha_s \equiv \overline{\alpha}$,

$$F(x, \overline{y}, \overline{\alpha}, p) \le \sup_{\alpha\in\mathcal{A}} \left\{ -\frac{1}{t} \int_0^t \big((p, f(x, y_s, \alpha_s)) + l(x, y_s, \alpha_s) \big)\, ds \right\}$$

for all $t > 0$ and all solutions of $\dot{y}_s = g(x, y_s, \alpha_s)$, $y_0 = y$, $y_s \in \overline{Y}$. Then (10) implies $F(x, \overline{y}, \overline{\alpha}, p) \le \overline{H}(x, p)$, and we conclude by the arbitrariness of $(\overline{y}, \overline{\alpha})$. □

If the multifunction $Z(\cdot)$ is regular enough, say Lipschitz continuous in the Hausdorff metrics, then the value function $v(t, x)$ of the problem with dynamics (15) and cost functional $\underline{J}$ defined by (14) is the viscosity solution of

$$-\partial_t v + H_0(x, Dv) = 0 \text{ in } (0, T) \times \mathbb{R}^n \quad \text{and} \quad v(T, \cdot) = \underline{h} \text{ on } \mathbb{R}^n;$$

see, e.g., [15]. Then a comparison theorem gives $v \ge u$, where u is the solution of the limit problem (11).

Next we give three examples where $H_0 = \overline{H}$, and therefore $v = u$. In the first two we make assumptions on the m-dimensional control problem of minimizing $\int_0^t F(x, y_s, \alpha_s, p)\, ds$ for fixed (x, p), that is connected to $\overline{H}$ by the formulas (9) and (10).

Example 1: *The affine-convex case.* Suppose that A and Y are convex and, for all fixed x, f and g are affine and l is convex with respect to (y, a). Note that $g(x, y, A)$ is convex, so it contains 0 by the controllability assumption and $Z(x) \ne \emptyset$. We define

$$G(x, \alpha, p) := \limsup_{t \to +\infty} \frac{1}{t} \int_0^t F(x, y_s, \alpha_s, p)\, ds, \qquad \alpha \in \mathcal{A}_{y_0}, \tag{16}$$

where y_0 is fixed and $\dot{y}_s = g(x, y_s, \alpha_s)$. By the representation formula (9),

$$\overline{H}(x, p) = \sup_{\alpha \in \mathcal{A}_{y_0}} G(x, \alpha, p)$$

for any choice of y_0. We fix $\alpha \in \mathcal{A}_{y_0}$ and choose $t_n \to +\infty$ such that

$$G(x, \alpha, p) = \lim_n \frac{1}{t_n} \int_0^{t_n} F(x, y_s, \alpha_s, p)\, ds. \tag{17}$$

By the convexity and compactness of A and $\overline{Y}$ we can extract a subsequence such that

$$\lim_n \frac{1}{t_n} \int_0^{t_n} y_s\, ds = \overline{y} \in \overline{Y}, \qquad \lim_n \frac{1}{t_n} \int_0^{t_n} \alpha_s\, ds = \overline{\alpha} \in A.$$

The assumptions of this example imply the concavity of $F(x, \cdot, \cdot, p)$ for all fixed (x, p), so

$$G(x, \alpha, p) \le \lim_n F\left(x, \frac{1}{t_n} \int_0^{t_n} y_s\, ds, \frac{1}{t_n} \int_0^{t_n} \alpha_s\, ds, p\right) = F(x, \overline{y}, \overline{\alpha}, p).$$

Moreover,

$$g\left(x, \frac{1}{t} \int_0^t y_s\, ds, \frac{1}{t} \int_0^t \alpha_s\, ds\right) = \frac{1}{t} \int_0^t g(x, y_s, \alpha_s)\, ds = \frac{1}{t}(y_t - y_0).$$

Here we set $t = t_n$ and pass to the limit to get, by the boundedness of $\overline{Y}$, $g(x, \overline{y}, \overline{\alpha}) = 0$. Then

$$G(x, \alpha, p) \le \sup_{(\overline{y}, \overline{\alpha}) \in Z(x)} F(x, \overline{y}, \overline{\alpha}, p) = H_0(x, p).$$

Since $\alpha \in \mathcal{A}$ is arbitrary we obtain $\overline{H}(x,p) \le H_0(x,p)$, and we can conclude that $\overline{H} = H_0$ by the preceding lemma.

Example 2: *The case of an asymptotically stable optimal trajectory.* Here we follow Example 7.4 in [5] and suppose for all (x,p) that for some y_0 there is a control $\alpha^* \in \mathcal{A}$ such that

$$\int_0^t F(x, y_s^*, \alpha_s^*, p)\, ds = \sup_{\alpha \in \mathcal{A}} \int_0^t F(x, y_s, \alpha_s, p)\, ds \qquad \text{for all } t > 0,$$

where $\dot{y}_s^* = g(x, y_s^*, \alpha_s^*)$, $y_0^* = y_0$, and

$$\lim_{s \to +\infty} \alpha_s^* = \alpha^*, \qquad \lim_{s \to +\infty} y_s^* = y^*$$

for some $\alpha^* = \alpha^*(x,p) \in A$, $y^* = y^*(x,p) \in \overline{Y}$. Then $g(x, y^*(x,p), \alpha^*(x,p)) = 0$ and

$$\lim_{t \to +\infty} \frac{1}{t} \int_0^t F(x, y_s^*, \alpha_s^*, p)\, ds = F(x, y^*(x,p), \alpha^*(x,p), p).$$

By the representation formula (10) we get

$$\overline{H}(x,p) = F(x, y^*(x,p), \alpha^*(x,p), p) \le \sup_{(y,\alpha) \in Z(x)} F(x,y,\alpha,p) = H_0(x,p),$$

and the equality $\overline{H} = H_0$ follows from the preceding lemma.

Example 3: *Separated controls.* In the case of subsection 1.3, formula (12) implies $\overline{H} \le H_0$, and therefore $\overline{H} = H_0$, if for all x and y there exists $\gamma^* \in C$ such that $g(x,y,\gamma^*) = 0$. Note that this condition follows from the controllability assumption on the fast variables if in addition $g(x,y,C)$ is a convex set for all x, y; this is the case, for instance, if one uses relaxed controls $\gamma_\cdot$.

It is obvious that the equality $H_0 = \overline{H}$ cannot hold at points where $Z(x) = \emptyset$, but it is known that the equality may also fail at points where $Z(x) \ne \emptyset$; see, e.g., [4]. We end this subsection with a simple example that exhibits this phenomenon and satisfies our assumptions.

Example 4: $-\infty < H_0 < \overline{H}$. Consider $A = [-1,1]$, $Y =]-1/2, 1/2[$, $g(y,\alpha) = \alpha - y$, $l(x,y,\alpha) = l_1(x) + |y|^2 - |\alpha|^2$, with l_1 continuous and bounded, and any f so that the assumptions of subsection 1.1 are satisfied. Then $H_0(x,0) = -l_1(x)$. On the other hand, by switching fast enough from $\alpha = 1$ to $\alpha = -1$ we can keep the solution y_s of $\dot{y}_s = g(x, y_s, \alpha_s)$, $y_0 = 0$, in any neighborhood of 0, so $\sup_\alpha \frac{1}{t} \int_0^t F(x, y_s, \alpha_s, 0)\, ds = -l_1(x) + 1$ and then $\overline{H}(x,0) = -l_1(x) + 1$ by (10).

1.5. A control interpretation for the limit problem. Now we construct an optimal control problem whose Hamiltonian is $\overline{H}$. Let $(\overline{Y} \times A)^r$ be the set of Radon probability measures on $\overline{Y} \times A$, and extend $\varphi = f, l, g$ to functions f^r, l^r, g^r defined on $R^n \times (\overline{Y} \times A)^r$ as it is usually done for relaxed controls, namely,

$$\varphi^r(x,\mu) := \int_{\overline{Y} \times A} \varphi(x,y,a) d\mu(y,a), \qquad \mu \in (\overline{Y} \times A)^r.$$

We call a *limiting relaxed control* a measure $\mu \in (\overline{Y} \times A)^r$ such that, for some $\alpha \in \mathcal{A}$, $t_n \to +\infty$, and y_0,

$$\mu_n := \frac{1}{t_n} \int_0^{t_n} \delta_{(y_s, \alpha_s)} ds \to \mu \qquad \text{weak star},$$

where $\delta_{(y,\alpha)}$ is the Dirac's mass at (y,α) and $\dot{y}_s = g(x, y_s, \alpha_s)$. Note that, for any Borel $Q \subseteq Y \times A$, $\mu_n(Q)$ is the proportion of time spent by (y_s, α_s) in Q, that is,

$$\mu_n(Q) = \frac{1}{t_n} |\{s \in [0, t_n] \,:\, (y_s, \alpha_s) \in Q\}|, \tag{18}$$

where $|\cdot|$ denotes the Lebesgue measure, so μ_n is an *occupational probability measure* in the terminology of [23]. We denote with $Z_l(x)$ the set of limiting relaxed control. The reason for the notation is that

$$Z_l(x) \subseteq Z^r(x) := \{\mu \in (\overline{Y} \times A)^r \,|\, g^r(x,\mu) = 0\}.$$

In fact,

$$g^r\left(x, \frac{1}{t_n}\int_0^{t_n} \delta_{(y_s,\alpha_s)} ds\right) = \frac{1}{t_n}\int_0^{t_n} g^r(x, \delta_{(y_s,\alpha_s)}) ds = \frac{1}{t_n}\int_0^{t_n} g(x, y_s, \alpha_s) ds$$
$$= \frac{1}{t_n}(y_{t_n} - y_0),$$

and the limit as $n \to +\infty$ gives $g^r(x,\mu) = 0$ by definition of weak star convergence.

Now we define

$$H_l^r(x,p) := \sup_{\mu \in Z_l(x)} F^r(x,\mu,p), \qquad F^r(x,\mu,p) := -(p, f^r(x,\mu)) - l^r(x,\mu).$$

THEOREM 7. *For all $x, p \in \mathbb{R}^n$, $\overline{H}(x,p) = H_l^r(x,p)$.*

Proof. The proof is similar to the affine-convex example of the previous subsection and we use the same notations. Let $\mu \in Z_l(x)$ be generated by $\alpha_\cdot, y_\cdot$, and the sequence $t_n \to +\infty$. Then

$$\frac{1}{t_n}\int_0^{t_n} F(x, y_s, \alpha_s, p) ds = \frac{1}{t_n}\int_0^{t_n} F^r(x, \delta_{(y_s,\alpha_s)}, p) ds = F^r\left(x, \frac{1}{t_n}\int_0^{t_n} \delta_{(y_s,\alpha_s)} ds, p\right),$$

and the right-hand side converges to $F^r(x,\mu,p)$ as $n \to +\infty$ by definition of weak star convergence. This proves

$$G(x,\alpha,p) = F^r(x,\mu,p), \tag{19}$$

where G is defined by (16). By taking the sup over $\mu \in Z_l(x)$ we get

$$H_l^r(x,p) \le \sup_{\alpha \in \mathcal{A}} G(x,\alpha,p) = \overline{H}(x,p).$$

To prove the opposite inequality we fix α, y_0, and $t_n \to +\infty$ such that

$$G(x,\alpha,p) = \lim_n \frac{1}{t_n}\int_0^{t_n} F(x, y_s, \alpha_s, p) ds.$$

By the compactness of $(\overline{Y} \times A)^r$ we can extract a subsequence, that we do not relabel, such that $\frac{1}{t_n}\int_0^{t_n} \delta_{(y_s,\alpha_s)} ds$ converges weak star to some μ, and $\mu \in Z_l(x)$. By taking the sup over $\alpha \in \mathcal{A}$ in (19) we then get, again by using (10),

$$\overline{H}(x,p) \le H_l^r(x,p),$$

which completes the proof. □

Remark. The control problem associated with the Hamiltonian H_l^r and the terminal cost $\underline{h}$ is the minimization of

$$\underline{J}^r(t,x,\mu) := \int_t^T l^r(x_s,\mu_s)\,ds + \underline{h}(x_T)$$

for the system

$$\dot{x}_s = f^r(x_s,\mu_s), \qquad \mu_s \in Z_l(x_s), \qquad x_t = x,$$

and measurable control functions $\mu : [0,T] \to (\overline{Y} \times A)^r$. If the multifunction $Z_l(\cdot)$ is regular enough, say it takes compact values and is Lipschitz continuous with respect to the Hausdorff metrics [15], then the value function of this control problem is continuous and it is the solution of the limit problem (11). We postpone to a future paper the investigation of the properties of Z_l and the connections with Artstein's invariant measures [3] and the related limit control problems of Vigodner [41].

After this paper was completed, Gaitsgory pointed out to us that under the current assumptions Z_l coincides with the *limit occupational measures set* constructed in his paper with Leizarowitz [23]; indeed, it is easy to deduce from (18) that any limiting relaxed control is a limit occupational measure, while the converse statement is based on the controllability of the fast variables and the results and methods of [23].

Remark. In connection with the reduced order method, we note that

$$H_0 \le \overline{H} = H_l^r \le H_0^r := \sup_{\mu \in Z^r(x)} F^r(x,\mu,p).$$

1.6. Convergence.

THEOREM 8. *As $\varepsilon \to 0+$ the functions $\{u_\varepsilon\}$ converge uniformly on compact subsets of $(0,T) \times \mathbb{R}^n \times \overline{Y}$ to the unique solution u of* (11)*; if h does not depend on y, the convergence is uniform on compact subsets of $(0,T] \times \mathbb{R}^n \times \overline{Y}$.*

Proof. We define the weak limits in the viscosity sense, or relaxed semilimits

$$\underline{u}(t,x) := \liminf_{*\varepsilon \to 0} \inf_y u_\varepsilon(t,x,y) := \liminf_{\varepsilon \to 0,\, t' \to t,\, x' \to x} \inf_y u_\varepsilon(t',x',y)$$

and $\overline{u} = \limsup^* \sup_y u_\varepsilon$.

We redefine $\overline{u}$ at $t = T$ by setting

$$\widetilde{u}(T,x) := \limsup_{t' \to T_-,\, x' \to x} \overline{u}(t',x') \quad \text{and} \quad \widetilde{u}(t,x) := \overline{u}(t,x) \quad \text{for } 0 < t < T.$$

We will show that $\underline{u}$ is a supersolution of (11) and $\widetilde{u}$ is a subsolution of (11). By comparison this gives $\underline{u} = \overline{u} = u$ in $(0,T) \times \mathbb{R}^n$ and implies the convergence of $\{u_\varepsilon\}$ to u uniformly on compact subsets of $(0,T) \times \mathbb{R}^n \times \mathbb{R}^m$.

To prove that $\widetilde{u}$ is a subsolution of the limit H–J equation we consider a strict maximum point $(\overline{t},\overline{x})$ of $\overline{u} - \varphi$ with $0 < \overline{t} < T$ and φ smooth. We want to show that

$$-\partial_t \varphi(\overline{t},\overline{x}) + \overline{H}(\overline{x}, D\varphi(\overline{t},\overline{x})) \le 0 \tag{20}$$

and suppose by contradiction that

$$-\partial_t \varphi(\overline{t},\overline{x}) + \lambda > 0 \quad \text{for } \lambda = \overline{H}(\overline{x},\overline{p}), \quad \overline{p} = D\varphi(\overline{t},\overline{x}).$$

Let χ be the solution of the cell problem at $(\overline{x},\overline{p})$, and define the perturbed test function as

$$\varphi_\varepsilon(t,x,y)=\varphi(t,x)+\varepsilon\chi(y).$$

We claim that for some $r>0$, φ_ε is a viscosity supersolution of

$$-\partial_t\varphi_\varepsilon+H\left(x,y,D_x\varphi_\varepsilon,\frac{D_y\varphi_\varepsilon}{\varepsilon}\right)\geq 0\quad\text{in } I_r\times B(\overline{x},r)\times\overline{Y}, \tag{21}$$

where $I_r=(\overline{t}-r,\overline{t}+r)$. To prove the claim we take a smooth ψ such that $\varphi_\varepsilon-\psi$ attains its minimum over $I_r\times B(\overline{x},r)\times\overline{Y}$ at $(\widetilde{t},\widetilde{x},\widetilde{y})$, and $(\varphi_\varepsilon-\psi)(\widetilde{t},\widetilde{x},\widetilde{y})=0$. Then the function $y\mapsto\chi(y)-\varepsilon^{-1}\psi(\widetilde{t},\widetilde{x},y)$ has a minimum at $\widetilde{y}$, so the definition of χ gives

$$H\left(\overline{x},\widetilde{y},\overline{p},\frac{D_y\psi}{\varepsilon}(\widetilde{t},\widetilde{x},\widetilde{y})\right)\geq\lambda.$$

Since χ is Lipschitz continuous it is easy to check that $|\varepsilon^{-1}D_y\psi(\widetilde{t},\widetilde{x},\widetilde{y}))|\leq Lip(\chi)$. Now we set $\gamma:=(-\partial_t\varphi(\overline{t},\overline{x})+\lambda)/2$ and use the continuity of H in (x,p), uniformly for $y\in\overline{Y}$ and $|\varepsilon^{-1}D_y\psi|\leq Lip(\chi)$, to find δ such that

$$H\left(x,\widetilde{y},p,\frac{D_y\psi}{\varepsilon}(\widetilde{t},\widetilde{x},\widetilde{y})\right)\geq\lambda-\gamma$$

for $|x-\overline{x}|<\delta$ and $|p-\overline{p}|<\delta$. Now choose $0<r\leq\delta$ such that

$$\left|D_x\varphi(\overline{t},\overline{x})-D_x\varphi(\widetilde{t},\widetilde{x})\right|<\delta\quad\text{and}\quad\left|\partial_t\varphi(\overline{t},\overline{x})-\partial_t\varphi(\widetilde{t},\widetilde{x})\right|<\gamma.$$

Note that the choice of r is independent of ψ. Since $D_x\psi(\widetilde{t},\widetilde{x},\widetilde{y})=D_x\varphi(\widetilde{t},\widetilde{x})$ and $\partial_t\psi(\widetilde{t},\widetilde{x},\widetilde{y})=\partial_t\varphi(\widetilde{t},\widetilde{x})$, we get

$$\left[-\partial_t\psi+H\left(\cdot,\cdot,D_x\psi,\frac{D_y\psi}{\varepsilon}\right)\right](\widetilde{t},\widetilde{x},\widetilde{y})\geq-\partial_t\varphi(\overline{t},\overline{x})-\gamma+\lambda-\gamma=0,$$

which completes the proof of the claim.

In view of (21), we can use a comparison principle for the mixed boundary value problem with prescribed data on $\partial(I_r\times B(\overline{x},r))$ and state-constrained condition at ∂Y (see, e.g., Theorem IX.1 in [16]) to obtain

$$\sup_{I_r\times B(\overline{x},r)\times\overline{Y}}(u_\varepsilon-\varphi_\varepsilon)\leq\sup_{\partial(I_r\times B(\overline{x},r))\times\overline{Y}}(u_\varepsilon-\varphi_\varepsilon).$$

It is not hard to deduce from this and the definitions of $\overline{u}$ and φ_ε that

$$(\overline{u}-\varphi)(\overline{t},\overline{x})\leq\sup_{\partial(I_r\times B(\overline{x},r))}(\overline{u}-\varphi),$$

and this is in contradiction to the fact that $(\overline{t},\overline{x})$ is a strict maximum point of $\overline{u}-\varphi$. This completes the proof of (20).

Next we show that $\underline{u}$ is a supersolution of the limit H–J equation. Now $(\overline{t},\overline{x})$ is a strict minimum point of $\underline{u}-\varphi$ and we assume by contradiction that $-\partial_t\varphi(\overline{t},\overline{x})+\lambda<0$, where $\lambda=\overline{H}(\overline{x},D\varphi(\overline{t},\overline{x}))$ as before. We also define χ and φ_ε as before, and now claim that φ_ε is a viscosity subsolution of

$$-\partial_t\varphi_\varepsilon+H\left(x,y,D_x\varphi_\varepsilon,\frac{D_y\varphi_\varepsilon}{\varepsilon}\right)\leq 0\quad\text{in } I_r\times B(\overline{x},r)\times Y.$$

The proof is essentially the same as the proof of (21). Now we exploit the fact that u_ε is a (constrained) supersolution of the same PDE in $I_r \times B(\overline{x}, r) \times \overline{Y}$ and the comparison principle for the mixed Dirichlet-constrained boundary value problem (see, e.g., Theorem IX.1 in [16]) to get

$$\inf_{I_r \times B(\overline{x},r) \times \overline{Y}} (u_\varepsilon - \varphi_\varepsilon) \geq \inf_{\partial(I_r \times B(\overline{x},r)) \times \overline{Y}} (u_\varepsilon - \varphi_\varepsilon).$$

This implies

$$(\underline{u} - \varphi)(\overline{t}, \overline{x}) \geq \inf_{\partial(I_r \times B(\overline{x},r))} (\underline{u} - \varphi),$$

a contradiction with the choice of $(\overline{t}, \overline{x})$. Therefore

$$-\partial_t \varphi(\overline{t}, \overline{x}) + \overline{H}(\overline{x}, D\varphi(\overline{t}, \overline{x})) \geq 0,$$

and so $\underline{u}$ is a supersolution of the limit H–J equation.

Finally we check the terminal condition. The hypotheses on f and l imply easily the estimate

$$u_\varepsilon(t, x, y) \geq -(T-t)||l||_\infty + \inf\{\underline{h}(x') : |x' - x| \leq ||f||_\infty (T-t)\}$$

for all $\varepsilon > 0$. Since $\underline{h}$ is continuous, in the limit we obtain $\underline{u}(T, x) \geq \underline{h}(x)$.

For $\widetilde{u}$, we note that

$$u_\varepsilon(t, x, y) \leq (T-t)||l||_\infty + \inf_{\alpha \in \mathcal{A}_{(x,y)}} h(x, y_T^t) + \omega_h(|x - x_T^t|),$$

where (x_T^t, y_T^t) denote the position of the system at time T if the position at time t is (x, y), and ω_h is the modulus of continuity of h. Since $|x - x_T^t| \leq (T-t)||f||_\infty$, the first and third term on the right-hand side of the preceding estimate tend to 0 as $t \to T-$. To reach the conclusion we are going to prove that

$$\limsup_{\varepsilon \to 0,\, t' \to t,\, x' \to x} \sup_y \inf_{\alpha \in \mathcal{A}_{(x',y)}} h(x', y_T^{t'}) \leq \underline{h}(x)$$

for all $t < T$.

Without loss of generality we can assume that

$$B(0, r) \subset \{g(x, y, \alpha) \mid \alpha \in A\}.$$

In fact, if we take relaxed controls or use Carathéodory's theorem (as in Lemma 2.7 of [6]) to convexify $g(x, y, A)$, the value function u_ε does not change because the Hamiltonian H is the same. Then it is easy to see by means of a standard selection lemma (e.g., as in Lemma 2.8 of [6]) that any polygonal $\gamma \subset \overline{Y}$ is the trajectory of a solution y_t of the system, with speed r/ε. Therefore, by Lemma 3(ii), for some constant M the system can reach any point $\overline{y} \in \overline{Y}$ from any $y \in \overline{Y}$ within the time $M\varepsilon/r$. Then, for any initial position $y \in \overline{Y}$ of the system,

$$\inf_{\alpha \in \mathcal{A}_{(x',y)}} h(x', y_T^{t'}) = \underline{h}(x') \quad \text{if } t' \geq t + \frac{M\varepsilon}{r}.$$

For any $t < T$ we can restrict $\varepsilon < r(T-t)/M$ and get

$$\limsup_{\varepsilon \to 0,\, t' \to t,\, x' \to x} \sup_y \inf_{\alpha \in \mathcal{A}_{(x',y)}} h\left(x', y_T^{t'}\right) = \limsup_{x' \to x} \underline{h}(x') = \underline{h}(x),$$

where in the last equality we used the continuity of $\underline{h}$. Therefore $\widetilde{u}(T,x) \le \underline{h}(x)$ and the proof of the first statement is complete.

In the case of $h = h(x)$ we have, for $0 \le (T-t)||f||_\infty \le 1$, $|x-x'| \le 1$, and all y,

$$|u_\varepsilon(t,x',y) - h(x)| \le (T-t)||l||_\infty + \omega(|x-x'| + (T-t)||f||_\infty),$$

where ω is the modulus of continuity of h in $\overline{B}(x,2)$. Therefore $\underline{u}(T,x) = \overline{u}(T,x) = \widetilde{u}(T,x) = h(x)$, and the convergence of u_ε is uniform on compact subsets up to time $t = T$. □

2. Stochastic control with periodic fast variables.

2.1. The ε-control problem. For $\varepsilon > 0$ fixed, we now consider the following finite horizon stochastic control problem in $(0,T] \times \mathbb{R}^n \times \mathbb{R}^m$. Let $(\Omega, \mathcal{F}, P)$ be a probability space, endowed with a right-continuous filtration $(\mathcal{F}_t)_{0 \le t \le T}$ and an r-dimensional adapted Brownian motion W_t. Given a progressively measurable α with values in a compact set A, the stochastic differential equation

$$\begin{aligned} dx_s &= f(x_s, y_s, \alpha_s)\, ds + \sigma(x_s, y_s, \alpha_s)\, dW_s, \\ dy_s &= \varepsilon^{-1} g(x_s, y_s, \alpha_s)\, ds + \varepsilon^{-1/2} \tau(x_s, y_s, \alpha_s)\, dW_s \end{aligned}$$

for $s \ge t$, starting from $x_t = x \in \mathbb{R}^n$, $y_t = y \in \mathbb{R}^m$, has a pathwise unique adapted strong solution when the functions f, g, σ, τ are Lipschitz continuous in (x,y) uniformly in α. The variable x is called the slow variable and y the fast variable. We refer to Fleming and Soner [21] for a presentation of the theory of stochastic control and its relationship to the theory of viscosity solutions. We shall always assume that all of the functions are $\mathbb{Z}^m$-periodic in the fast variable y.

The associated value function with running cost l and terminal cost h is given by

$$u_\varepsilon(t,x,y) = \inf_\alpha E\left\{ \int_t^T l(x_s, y_s, \alpha_s)\, ds + h(x_T) \right\}.$$

Under the assumptions we recall below, it is continuous and bounded on $(0,T] \times \mathbb{R}^n \times \mathbb{R}^m$ uniformly in ε. It is also periodic in the fast variable y.

We shall make throughout the following set of assumptions that are classic in the theory of stochastic control.

- The control set A is a compact metric space.
- The functions f, g, σ, τ, and l are bounded continuous functions in $\mathbb{R}^n \times \mathbb{R}^m \times A$ with values, respectively, in $\mathbb{R}^n$, $\mathbb{R}^m$, $\mathbb{M}^{n,r}$ (the set of the $n \times r$ real matrices), $\mathbb{M}^{m,r}$, and $\mathbb{R}$. They are $\mathbb{Z}^m$-periodic in the fast variable y.
- The drift vectors f and g and the dispersion matrices σ and τ are Lipschitz continuous in (x,y), uniformly in α.
- The running cost l is uniformly continuous in (x,y), uniformly in α.
- The terminal cost $h : \mathbb{R}^n \to \mathbb{R}$ is bounded uniformly continuous.

2.2. The H–J–B equation. Consider the diffusion matrices

$$a = \frac{\sigma\sigma^T}{2}, \qquad b = \frac{\tau\tau^T}{2}, \qquad c = \frac{\tau\sigma^T}{2},$$

and associate the Hamiltonian

$$\begin{aligned} H(x,y,p,q,X,Y,Z) = \max_{\alpha \in A} \{ &-\mathrm{tr}(a(x,y,\alpha)X) - \mathrm{tr}(b(x,y,\alpha)Y) - \mathrm{tr}(c(x,y,\alpha)Z) \\ &-\mathrm{tr}(Zc(x,y,\alpha)) - (p, f(x,y,\alpha)) - (q, g(x,y,\alpha)) - l(x,y,\alpha)\}. \end{aligned}$$

It is defined on $\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{S}^n\times\mathbb{S}^m\times\mathbb{M}^{n,m}$, where $\mathbb{S}^n$ designates the set of the symmetric $n\times n$ matrices. Given a function $u(t,x,y)$ defined on $(0,T]\times\mathbb{R}^n\times\mathbb{R}^m$, we consider the partial gradients $\partial_t u$, $D_x u$, and $D_y u$. We also define the partial Hessian matrices $D^2_{xx}u$, $D^2_{yy}u$, and $D^2_{xy}u=\left(\partial^2_{x_i,y_j}u\right)_{1\le i\le n,1\le j\le m}$, so that the full Hessian matrix of u with respect to (x,y) is

$$D^2u=\begin{pmatrix} D^2_{xx}u & D^2_{xy}u\\ (D^2_{xy}u)^T & D^2_{yy}u\end{pmatrix}.$$

By the dynamic programming principle, the value function u_ε is a viscosity solution of the second order degenerate parabolic H–J–B equation

$$\begin{cases} -\partial_t u_\varepsilon + H(x,y,D_xu_\varepsilon,\frac{D_yu_\varepsilon}{\varepsilon},D^2_{xx}u_\varepsilon,\frac{D^2_{yy}u_\varepsilon}{\varepsilon},\frac{D^2_{xy}u_\varepsilon}{\sqrt{\varepsilon}})=0 \quad \text{in } (0,T)\times\mathbb{R}^n\times\mathbb{R}^m,\\ u_\varepsilon(T,\cdot)=h \quad \text{on } \mathbb{R}^n\times\mathbb{R}^m.\end{cases} \tag{22}$$

The following theorem records these facts. We refer to [21] for a proof and a detailed discussion.

THEOREM 9. *For every $\varepsilon>0$, the value function u_ε is the unique bounded continuous viscosity solution of* (22) *in* $(0,T]\times\mathbb{R}^n\times\mathbb{R}^m$.

For further use, we recall that the uniqueness statement in the theorem takes the form of the following comparison principle. If u is a bounded u.s.c. viscosity subsolution of (22) and v is a bounded l.s.c. viscosity supersolution, then we have $u\le v$ in $(0,T]\times\mathbb{R}^n\times\mathbb{R}^m$. We refer to [17, 21] for the precise definitions of a subsolution and a supersolution and for the proof of the comparison principle.

2.3. The ergodic control problem in the fast variable and the effective Hamiltonian. We shall make one of the following three assumptions to guarantee some averaging properties of the fast dynamics.

(I) The diffusions in the fast variable are uniformly nondegenerate, i.e., there is a constant $\nu>0$ such that

$$b(x,y,\alpha)\ge\nu I_m \quad \text{for all } (x,y,\alpha),$$

where I_m denotes the m-dimensional identity matrix. Moreover, the running cost $l(x,\cdot,\alpha)$ is Hölder continuous for some exponent $0<\beta\le 1$, uniformly on (x,α).

(II) The diffusions in the fast variable are independent of x ($b\equiv b(y,\alpha)$) and there is a deterministic controllable subsystem in the fast variable, i.e., there is some $r>0$ and some $A'\subset A$ so that

$$b(y,\alpha)=0 \quad \text{for all } \alpha\in A' \quad\text{and}\quad B(0,r)\subset\overline{\text{conv}}\{g(x,y,\alpha)\mid\alpha\in A'\}$$

for all (x,y).

(III) The drifts and diffusions in the fast variable do not depend on x, y ($g\equiv g(\alpha)$ and $b\equiv b(\alpha)$) and satisfy the nonresonance condition

$$\text{for every } k\in\mathbb{Z}^m \text{ there is } \alpha\in A \text{ such that } \quad (g(\alpha),k)\ne 0 \text{ or } b(\alpha)k\ne 0.$$

In terms of the Hamiltonian, case (I) corresponds to the uniform ellipticity of H in Y,

$$H(x,y,p,q,X,Y+W,Z)\le H(x,y,p,q,X,Y,Z)-\nu\text{tr}W, \qquad W\ge 0,$$

while (II) corresponds to the coercivity with respect to q:

$$H(x,y,p,q,X,Y,Z) \geq r|q| - C(1+|p|+|X|).$$

The preceding assumptions are of two different natures. Some of them demand that some quantities in the fast dynamics are independent of the slow variable. They are needed to ensure enough regularity of the averaged quantities with respect to the slow variables. It is an open question whether they can be dispensed with. The second kind of assumption is more fundamental. They guarantee the solvability of the ergodic control problem in the fast variable. These assumptions correspond to some of the cases studied by Arisawa and Lions [2]. In our context, their results read as follows.

THEOREM 10. *Assume that either* (I) *or* (II) *or* (III) *holds. Let* (x,p,X) *be fixed. For every* $\delta > 0$, *let* w_δ *denote the unique viscosity solution of the stationary problem in the fast variable*

$$\delta w_\delta + H(x,y,p,D_y w_\delta, X, D^2_{yy} w_\delta, 0) = 0 \quad \text{in } \mathbb{R}^m, \qquad w_\delta \text{ periodic.} \tag{23}$$

Then, as $\delta \to 0+$, *the family* $\{\delta w_\delta\}$ *converges to a constant* $-\overline{H}(x,p,X)$, *uniformly with respect to* y.

When one looks at the solution w_δ of the H–J–B equation (23) as the value function of a discounted control problem in the fast variable, the theorem asserts that

$$\begin{aligned}\overline{H}(x,p,X) &= \lim_{\delta\to 0+} \sup_{\alpha} \Big\{ \delta E \int_0^\infty e^{-\delta s}\big(-\operatorname{tr}(a(x,y_s,\alpha_s)X) - (p, f(x,y_s,\alpha_s)) - l(x,y_s,\alpha_s)\big)\, ds \\ &\qquad\qquad \mid dy_s = g(x,y_s,\alpha_s)\, ds + \tau(x,y_s,\alpha_s)\, dW_s,\ y_0 = y \Big\},\end{aligned} \tag{24}$$

the convergence being uniform in y.

In cases (I) and (II), one can characterize the effective Hamiltonian in the more convenient form of section 1. It is the unique constant $\overline{H}$ for which the cell problem

$$H(x,y,p,D_y\chi, X, D^2_{yy}\chi, 0) = \overline{H} \quad \text{in } \mathbb{R}^m, \qquad \chi \text{ periodic},$$

has a continuous solution χ. However, such a characterization is not available in case (III), for it may happen that the cell problem has no solution. We refer to [2] for an explicit example.

The above assumptions are three of the five cases studied by Arisawa and Lions [2]. Among the remaining two cases, the one-dimensional one in the fast variable ($m = 1$) can be handled in a similar way as (II); we omit it for simplicity. On the other hand, for the viscosity solution techniques to apply, the uniform convergence of $\{\delta w_\delta\}$ is essential. This is why we do not consider the fifth case that assumes (roughly) that at least one diffusion is uniformly nondegenerate (and not all, as in (I)), because the convergence of $\{\delta w_\delta\}$ may not be uniform (but in L^p for every $1 \leq p < \infty$). An example in which this happens is given in [2].

2.4. Examples for the effective Hamiltonian.

Example 1: *The coercive and separated case.* The first example, as in section 1, is the case of separated controls. We assume that the fast variable is controlled independently of the slow variable and that there is a controllable deterministic subsystem

(case (II)). The controls are of the form $\alpha = (\beta, \gamma)$ and $f = f(x, y, \beta)$, $\sigma = \sigma(x, y, \beta)$, $g = g(x, y, \gamma)$, and $\tau = \tau(x, y, \gamma)$. We also assume that $l = l(x, y, \beta)$. Under these assumptions, the representation formula reads as

$$\begin{aligned}
&\overline{H}(x, p, X)\\
&\quad = \lim_{\delta \to 0} \sup_{(\beta,\gamma)} \Big\{ \delta E \int_0^\infty e^{-\delta s} \big(- \operatorname{tr}(a(x, y_s, \beta_s)X) - (p, f(x, y_s, \beta_s)) - l(x, y_s, \beta_s) \big)\, ds\\
&\qquad\qquad\qquad\qquad\qquad\qquad \mid dy_s = g(x, y_s, \gamma_s)\, ds + \tau(x, y_s, \gamma_s)\, dW_s \Big\},\\
&\quad = \lim_{\delta \to 0} \sup_{\gamma} \left\{ \delta E \int_0^\infty e^{-\delta s} H_1(x, y_s, p, X)\, ds \mid dy_s = g(x, y_s, \gamma_s)\, ds + \tau(x, y_s, \gamma_s)\, dW_s \right\}
\end{aligned}$$

for

$$H_1(x, y, p, X) = \sup_{\beta} \{ -\operatorname{tr}(a(x, y, \beta)X) - (p, f(x, y, \beta)) - l(x, y, \beta) \}.$$

Arguing as in section 1, we deduce from the controllability assumption that

$$\overline{H}(x, p, X) = \sup_{y} H_1(x, y, p, X).$$

Thus the effective Hamiltonian corresponds to the original control problem where the fast variable plays the role of an additional control.

Example 2: *Uncontrolled and nondegenerate diffusion of the fast variables.* The second example assumes that the fast variable is an uncontrolled uniformly nondegenerate diffusion. Since we are in case (I), we know that the effective Hamiltonian is characterized by the solvability of the linear cell problem

$$H_1(x, y, p, X) - \operatorname{tr}(b(x, y) D^2_{yy} \chi) - (D_y \chi, g(x, y)) = \overline{H}(x, p, X) \quad \text{in } \mathbb{R}^m, \qquad \chi \text{ periodic}$$

(where H_1 is given above, with α instead of β). Assuming that the functions b and g are smooth in y, there is a unique solution μ_x (the invariant measure) of the adjoint equation

$$-\sum_{i,j} \frac{\partial^2}{\partial y_i \partial y_j} (b_{ij}(x, y)\mu_x) + \sum_i \frac{\partial}{\partial y_i} (g_i(x, y)\mu_x) = 0 \quad \text{in } \mathbb{R}^m, \qquad \mu_x \text{ periodic},$$

with mean $\int_{(0,1)^m} \mu_x(y)\, dy = 1$. This follows from the Fredholm alternative (see, for instance, [14] or [13]). A necessary and sufficient condition for the cell problem to have a solution is that

$$\overline{H}(x, p, X) = \int_{(0,1)^m} H_1(x, y, p, X)\mu_x(y)\, dy.$$

If, in addition, b and g are independent of y (and more generally when $g_i = \sum_j \frac{\partial b_{ij}}{\partial y_j}$), we have $\mu_x \equiv 1$. The effective Hamiltonian is therefore the average

$$\overline{H}(x, p, X) = \int_{(0,1)^m} H_1(x, y, p, X)\, dy.$$

This example is a variant of the results of Jensen and Lions [29] and Evans [19].

2.5. Regularity properties of the effective Hamiltonian.

PROPOSITION 11. *The effective Hamiltonian $\overline{H}$ is degenerate elliptic in X and convex in (p, X). Moreover, we have the bounds*

$$\inf_y H(x, y, p, 0, X, 0, 0) \le \overline{H}(x, p, X) \le \sup_y H(x, y, p, 0, X, 0, 0). \tag{25}$$

Proof. The bounds for $\overline{H}$ are a consequence of the observation that the constant functions

$$-\delta^{-1} \sup_y H(x, y, p, 0, X, 0, 0), \qquad -\delta^{-1} \inf_y H(x, y, p, 0, X, 0, 0)$$

are, respectively, a subsolution and a supersolution of (23). Therefore, by the comparison principle, we get

$$-\sup_y H(x, y, p, 0, X, 0, 0) \le \delta w_\delta \le -\inf_y H(x, y, p, 0, X, 0, 0).$$

Sending $\delta \to 0$ yields (25).

Degenerate ellipticity and convexity can be derived by analytical means as in [1]. They are also simple consequences of the representation formula (24). Indeed, for every δ fixed and every control α_s, the function

$$\delta E \int_0^\infty e^{-\delta s} \big(-\operatorname{tr}(a(x, y_s, \alpha_s)X) - (p, f(x, y_s, \alpha_s)) - l(x, y_s, \alpha_s) \big)\, ds$$

is linear in (p, X) and nonincreasing in X. Taking the supremum over the controls yields a function that is convex in (p, X) and nonincreasing in X, as is the limit as $\delta \to 0$. □

The continuity of the effective Hamiltonian is a consequence of the following result.

PROPOSITION 12. *There are a constant $C > 0$ and a modulus ω such that*

$$|\overline{H}(x, p', X') - \overline{H}(x, p, X)| \le C(|p' - p| + |X' - X|) \tag{26}$$

for all (x, p, p', X, X') and

$$|\overline{H}(x', p, X) - \overline{H}(x, p, X)| \le C|x' - x|(1 + |p| + |X|) + \omega(|x' - x|) \tag{27}$$

for all (x, x', p, X).

Proof. The first inequality follows at once from the representation formula (24) by taking the constant $C = \max(\|f\|_{L^\infty}, \|a\|_{L^\infty})$.

The second inequality is more delicate. When the drift and diffusion in the fast variable are independent of x (case (III)), the inequality follows from the representation formula for the constant $C = \max\big(Lip(f), Lip(a)\big)$ and for the modulus $\omega = \omega_l$. We give a second proof of this elementary result, which we shall modify in the other two cases. Since the drift and diffusion for the fast variable are independent of x, the Hamiltonian H satisfies

$$H(x', y, p, q, X, Y, 0) \le H(x, y, p, q, X, Y, 0) + C|x' - x|(1 + |p| + |X|) + \omega(|x' - x|). \tag{28}$$

Therefore, the function $w_\delta(\cdot, x, p, X)$ is a subsolution of

$$\delta w_\delta + H(x', y, p, D_y w_\delta, X, D^2_{yy} w_\delta, 0) \le C|x' - x|(1 + |p| + |X|) + \omega(|x' - x|).$$

By the comparison principle, we obtain the uniform bound

$$\delta w_\delta(\cdot,x,p,X) \le \delta w_\delta(\cdot,x',p,X) + C|x'-x|(1+|p|+|X|) + \omega(|x'-x|).$$

Sending $\delta \to 0$ yields

$$\overline{H}(x,p,X) \ge \overline{H}(x',p,X) - C|x'-x|(1+|p|+|X|) - \omega(|x'-x|).$$

We get (27) after exchanging x and x'.

We now assume (II). As g may now depend on x (but not b), we have to replace (28) by

$$H(x',y,p,q,X,Y,0) \le H(x,y,p,q,X,Y,0) + C|x'-x|\big(1+|p|+|q|+|X|\big) + \omega(|x'-x|). \tag{29}$$

The controllability assumption gives the coercivity of H in q, uniformly in Y, as

$$H(x,y,p,q,X,Y,0) \ge r|q| - C(1+|p|+|X|).$$

Since

$$\|\delta w_\delta\|_{L^\infty} \le \sup_y |H(x,y,p,0,X,0,0)| \le C(1+|p|+|X|),$$

we deduce that the solution of (23) is Lipschitz continuous with the bound

$$\|D_y w_\delta(\cdot,x,p,X)\|_{L^\infty} \le r^{-1}\big(\|\delta w_\delta\|_{L^\infty} + C(1+|p|+|X|)\big) \le C(1+|p|+|X|). \tag{30}$$

We deduce from (29) that $w_\delta(\cdot,x,p,X)$ is a subsolution of

$$\begin{aligned}\delta w_\delta + H(x',y,p,D_y w_\delta, X, D^2_{yy} w_\delta, 0) &\\ \le C|x'-x|(1+|p|+|X|+\|D_y w_\delta\|_{L^\infty}) + \omega(|x'-x|) &\\ \le C|x'-x|(1+|p|+|X|) + \omega(|x'-x|).&\end{aligned}$$

The inequality for $\overline{H}$ is deduced as before from the comparison principle.

We finally consider case (I). As g and b now depend on x, the inequality for H reads as

$$\begin{aligned}H(x',y,p,q,X,Y,0) \le H(x,y,p,q,X,Y,0)&\\ + C|x'-x|\big(1+|p|+|q|+|X|+|Y|\big) + \omega(|x'-x|).&\end{aligned} \tag{31}$$

We claim that the solution w_δ of (23) is in $C^{2,\overline{\beta}}$ for some exponent $0 < \overline{\beta} \le \beta$ with

$$\|w_\delta(\cdot,x,p,X) - w_\delta(0,x,p,X)\|_{C^{2,\overline{\beta}}(\mathbb{R}^m)} \le C(1+|p|+|X|). \tag{32}$$

Admitting this temporarily, we deduce that $w_\delta(\cdot,x,p,X)$ is a subsolution of

$$\begin{aligned}\delta w_\delta + H(x',y,p,D_y w_\delta, X, D^2_{yy} w_\delta, 0) &\\ \le C|x'-x|(1+|p|+|X|+\|D_y w_\delta\|_{L^\infty} + \|D^2_{yy} w_\delta\|_{L^\infty}) + \omega(|x'-x|) &\\ \le C|x'-x|(1+|p|+|X|) + \omega(|x'-x|).&\end{aligned}$$

Inequality (27) for $\overline{H}$ follows as before by comparison.

The proof of (32) relies on the regularity theory for uniformly elliptic H–J–B equations (see Gilbarg and Trudinger [24] and Safonov [36]). Our argument is patterned

after the one of Arisawa and Lions [2]. We give a sketch of it to exhibit the linear growth in (p, X) of the bound. The first step is to establish the uniform bound

$$\|w_\delta(\cdot, x, p, X) - w_\delta(0, x, p, X)\|_{L^\infty(\mathbb{R}^m)} \le C(1 + |p| + |X|) \tag{33}$$

for all (x, p, X) and $0 < \delta < \overline{\delta}$, for some constant C and some $\overline{\delta} > 0$. Suppose that (33) is false. Then there is a sequence $(\delta_\kappa, x_k, p_k, X_k)$ with $\delta_k \to 0$ for which the solution $w_k = w_{\delta_k}(\cdot, x_k, p_k, X_k)$ of (23) satisfies

$$\|w_k - w_k(0)\|_{L^\infty} \ge k(1 + |p_k| + |X_k|).$$

We set $\eta_k = \|w_k - w_k(0)\|_{L^\infty}^{-1}$ and $\widetilde{w}_k = \eta_k(w_k - w_k(0))$. Then, $\widetilde{w}_k(0) = 0$, $\|\widetilde{w}_k\|_{L^\infty} = 1$, and $\widetilde{w}_k$ is a solution of

$$\delta_k \widetilde{w}_k + \eta_k \delta_k w_k(0) + \sup_{\alpha \in A}\{ - \operatorname{tr}(b(x_k, y, \alpha) D^2 \widetilde{w}_k) \\ -(D\widetilde{w}_k, g(x_k, y, \alpha)) - \eta_k L(y, \alpha, x_k, p_k, X_k)\} = 0$$

for

$$L(y, \alpha, x, p, X) = \operatorname{tr}(a(x, y, \alpha)X) + (p, f(x, y, \alpha)) + l(x, y, \alpha).$$

Since

$$\|\delta w_\delta\|_{L^\infty} \le C(1 + |p| + |X|)$$

and

$$\|L(\cdot, \alpha, x, p, X)\|_{C^{0,\beta}} \le C(1 + |p| + |X|),$$

we have

$$|\eta_k \delta_k w_k(0)| + \|\eta_k L(\cdot, \alpha, x_k, p_k, X_k)\|_{C^{0,\beta}} \le C\eta_k(1 + |p_k| + |X_k|) \le \frac{C}{k}.$$

The regularity theory for uniformly elliptic H–J–B equations therefore yields the uniform boundedness of $\widetilde{w}_k$ in $C^{2,\overline{\beta}}$ for some $\overline{\beta}$, depending only on m, the ellipticity constant ν, and β. Moreover, for α_0 fixed, the families $\{b(x_k, \cdot, \alpha_0)\}$ and $\{g(x_k, \cdot, \alpha_0)\}$ are equi-bounded and equi-continuous. Therefore, along a subsequence, the functions $\widetilde{w}_k$ and their derivatives of order ≤ 2, respectively $b(x_k, \cdot, \alpha_0)$ and $g(x_k, \cdot, \alpha_0)$, converge uniformly to some function $\widetilde{w}$ in $C^{2,\overline{\beta}}$ and its derivatives of order ≤ 2, respectively to b and g. The function b is clearly $\ge \nu I_p$, while $\widetilde{w}$ is a periodic function in $C^{2,\overline{\beta}}$ such that $\widetilde{w}(0) = 0$ and $\|\widetilde{w}\|_{L^\infty} = 1$. Since $\delta_k \widetilde{w}_k$, $\eta_k \delta_k w_k(0)$, and $\eta_k L(\cdot, \alpha, x_k, p_k, X_k)$ converge to 0 uniformly, we deduce from the stability results for viscosity solutions that $\widetilde{w}$ is a classic subsolution of the uniformly elliptic linear equation

$$-\operatorname{tr}(b(y) D^2 \widetilde{w}) - (D\widetilde{w}, g(y)) \le 0.$$

Since $\widetilde{w}$ is periodic, it achieves its maximum at some point. By the strong maximum principle, it must be constant. This is impossible, for we must have $\widetilde{w}(0) = 0$ and $\|\widetilde{w}\|_{L^\infty} = 1$.

Recalling that the running cost $L(\cdot, \alpha, x, p, X)$ is Hölder continuous with $C^{0,\beta}$ norm growing linearly in p and X, we deduce from the bound (33) and from the

regularity theory for uniformly elliptic H–J–B equations that there is some $0 < \overline{\beta} \le \beta$, $\overline{\beta}$ depending only on m, ν, and β, such that

$$\begin{aligned}\|w_\delta - w_\delta(0)\|_{C^{2,\overline{\beta}}} &\le C\left(\|w_\delta - w_\delta(0)\|_{L^\infty} + |\delta w_\delta(0)| + \sup_\alpha \|L(\cdot,\alpha,x,p,X)\|_{C^{0,\beta}}\right)\\ &\le C(1+|p|+|X|).\end{aligned}$$

This is (32). □

In order to solve the limit equation with Hamiltonian $\overline{H}$, one needs to strengthen slightly the regularity of $\overline{H}$ in Proposition 12. One of the following properties is sufficient to invoke results from the theory of viscosity solutions.

— The effective Hamiltonian is uniformly elliptic and satisfies (27).
— The effective Hamiltonian is of first order and satisfies (27).
— The effective Hamiltonian is degenerate elliptic and satisfies the following structure condition (see [17]): there is a modulus ω such that, for every $\kappa > 0$ and every $x, x' \in \mathbb{R}^n$, $X, X' \in \mathbb{S}^n$ so that

$$-3\kappa \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} \le \begin{pmatrix} X & 0 \\ 0 & -X' \end{pmatrix} \le 3\kappa \begin{pmatrix} I & -I \\ -I & I \end{pmatrix},$$

we have

$$\overline{H}\left(x', \kappa(x-x'), X'\right) \le \overline{H}(x, \kappa(x-x'), X) + \omega\left(\kappa|x'-x|^2 + |x'-x|\right). \tag{34}$$

Condition (34) is the most general but it is an open question whether it is true in general. We can only prove it when the drift and diffusion in the fast variable do not depend on x. This leads us to make one of the following assumptions about the dynamics.

(IV) The diffusions in the slow variable are uniformly nondegenerate, i.e., there is a constant $\mu > 0$ such that

$$a(x,y,\alpha) \ge \mu I_n \qquad \text{for all} \quad (x,y,\alpha).$$

(V) The problem in the slow variable is deterministic ($a \equiv 0$).
(VI) The drifts and diffusions in the fast variable are independent of x ($g \equiv g(y,\alpha)$ and $b \equiv b(y,\alpha)$).

From the representation formula (24), it is obvious that $\overline{H}$ is uniformly elliptic with constant μ in case (IV) and that it is of first order in case (V). Condition (34) also follows in case (VI) from the representation formula, since the fast dynamics are independent of x (and because we have classically $-\mathrm{tr}(a(x',y,\alpha)X') \le -\mathrm{tr}(a(x,y,\alpha)X) + C\kappa|x'-x|^2$ when the matrices X and X' satisfy the inequality in (34)).

We can now invoke the theory of viscosity solutions to obtain the solvability of the limit equation. We refer to [17] as well as to [28] for the results and proofs.

PROPOSITION 13. *Assume either* (I) *or* (II) *or* (III)*, and either* (IV) *or* (V) *or* (VI)*. Then there is a unique bounded continuous viscosity solution of the limit equation*

$$-\partial_t u + \overline{H}(x, Du, D^2u) = 0 \quad \text{in } (0,T)\times\mathbb{R}^n, \qquad u(T,\cdot) = h \quad \text{on } \mathbb{R}^n. \tag{35}$$

Moreover, if u is a bounded u.s.c. subsolution and v is a bounded l.s.c. supersolution, then $u \le v$ on $(0,T]\times\mathbb{R}^n$.

2.6. Convergence.

THEOREM 14. *Assume either* (I) *or* (II) *or* (III) *and either* (IV) *or* (V) *or* (VI). *Then, as* $\varepsilon \to 0+$, *the collection* $\{u_\varepsilon\}$ *converges uniformly on the compact subsets of* $(0,T]\times\mathbb{R}^n\times\mathbb{R}^m$ *to the unique viscosity solution* u *of* (35).

Proof. The functions u_ε are bounded in $(0,T]\times\mathbb{R}^n\times\mathbb{R}^m$ uniformly in ε. We can therefore define the half-relaxed limits on $(0,T]\times\mathbb{R}^n$:

$$\underline{u}(t,x)=\liminf_{\varepsilon\to 0,\ t'\to t,\ x'\to x}\ \inf_y u_\varepsilon(t',x',y),\qquad \overline{u}(t,x)=\limsup_{\varepsilon\to 0,\ t'\to t,\ x'\to x}\ \sup_y u_\varepsilon(t',x',y).$$

As in the first section, we shall prove that $\underline{u}$ is a supersolution of (35) and that $\overline{u}$ is a subsolution of (35). By the comparison principle, we shall get $\underline{u}=\overline{u}=u$ in $(0,T]\times\mathbb{R}^n$. This gives classically the uniform convergence on the compact subsets of $(0,T]\times\mathbb{R}^n\times\mathbb{R}^m$ of $\{u_\varepsilon\}$ to u.

We only check that $\overline{u}$ is a subsolution of (35), the proof that $\underline{u}$ is a supersolution being analogous. Let $w(t,x)$ be the continuous viscosity solution of

$$-\partial_t w+\inf_y H(x,y,D_x w,0,D^2_{xx}w,0,0)=0\quad\text{in }(0,T)\times\mathbb{R}^n,\qquad w(T,\cdot)=h\quad\text{on }\mathbb{R}^n.$$

It is clearly a viscosity supersolution of (22). By the comparison principle, we have $u_\varepsilon(t,x,y)\le w(t,x)$ for all $\varepsilon>0$, $0<t\le T$, x, y. Taking the semilimit, we deduce that $\overline{u}(T,\cdot)\le h$ on $\mathbb{R}^n$. This proves that $\overline{u}$ is a subsolution at the terminal boundary.

We next prove that $\overline{u}$ is a subsolution in $(0,T)\times\mathbb{R}^n$. Let $(\overline{t},\overline{x})\in(0,T)\times\mathbb{R}^n$ be a strict maximum point of $\overline{u}(t,x)-\varphi(t,x)$ with $\overline{u}(\overline{t},\overline{x})=\varphi(\overline{t},\overline{x})$. We argue by contradiction, assuming that

$$-\partial_t\varphi(\overline{t},\overline{x})+\overline{H}(\overline{x},D\varphi(\overline{t},\overline{x}),D^2\varphi(\overline{t},\overline{x}))>0.$$

Put $\overline{H}=\overline{H}(\overline{x},D\varphi(\overline{t},\overline{x}),D^2\varphi(\overline{t},\overline{x}))$. Let v_ε be the periodic solution of

$$\varepsilon v_\varepsilon+H(\overline{x},y,D\varphi(\overline{t},\overline{x}),D_y v_\varepsilon,D^2\varphi(\overline{t},\overline{x}),D^2_{yy}v_\varepsilon,0)=\overline{H}\quad\text{in }\mathbb{R}^m.$$

By Theorem 10 and the definition of the effective Hamiltonian, we know that $\varepsilon\big(v_\varepsilon-\varepsilon^{-1}\overline{H}\big)$ converges uniformly to $-\overline{H}$. Therefore, $\varepsilon v_\varepsilon$ converges uniformly to 0. For $\varepsilon>0$, we consider the perturbed test function

$$\psi_\varepsilon(t,x,y)=\varphi(t,x)+\varepsilon v_\varepsilon(y).$$

We will show that there is a small $r\in(0,\overline{t}\wedge(T-\overline{t}))$ so that ψ_ε is a supersolution of (22) in $Q_r=(\overline{t}-r,\overline{t}+r)\times B(\overline{x},r)\times\mathbb{R}^m$ for ε small. We suppose this has been proved and reach a contradiction. Since $\{\psi_\varepsilon\}$ converges uniformly to φ on $\overline{Q_r}$, we have

$$\limsup_{\varepsilon\to 0,\ t'\to t,\ x'\to x}\ \sup_y(u_\varepsilon-\psi_\varepsilon)=\overline{u}(t,x)-\varphi(t,x).$$

But $(\overline{t},\overline{x})$ is a strict maximum point of $\overline{u}-\varphi$, so the above relaxed upper limit is <0 on ∂Q_r. By compactness (recall that u_ε and ψ_ε are periodic in y), one can find $\eta>0$ so that $u_\varepsilon-\psi_\varepsilon\le-\eta$ on ∂Q_r for ε small, i.e., $\psi_\varepsilon\ge u_\varepsilon+\eta$ on ∂Q_r. Since ψ_ε is a supersolution of (22) in Q_r, we deduce from the comparison principle that $\psi_\varepsilon\ge u_\varepsilon+\eta$ in Q_r for ε small. Taking the upper semilimit, we get $\varphi\ge\overline{u}+\eta$ in $(\overline{t}-r,\overline{t}+r)\times B(\overline{x},r)$. This is impossible for $\varphi(\overline{t},\overline{x})=\overline{u}(\overline{t},\overline{x})$.

We have to show that ψ_ε is a supersolution of (22) in Q_r for r small, for all ε small. For every $(t,x,y) \in Q_r$, we have

$$
\begin{aligned}
(36)\qquad & -\partial_t \psi_\varepsilon + H\left(x, y, D_x\psi_\varepsilon, \frac{D_y\psi_\varepsilon}{\varepsilon}, D^2_{xx}\psi_\varepsilon, \frac{D^2_{yy}\psi_\varepsilon}{\varepsilon}, \frac{D^2_{xy}\psi_\varepsilon}{\sqrt{\varepsilon}}\right) \\
& \qquad = -\partial_t\varphi(t,x) + H(x, y, D\varphi(t,x), D_y v_\varepsilon(y), D^2\varphi(t,x), D^2_{yy} v_\varepsilon(y), 0).
\end{aligned}
$$

When g and b are independent of x (case (III)), the Hamiltonian satisfies

$$
\begin{aligned}
H(x,y,p,q,X,Y,0) \ge\ & H(\overline{x}, y, \overline{p}, q, \overline{X}, Y, 0) \\
& -C|x-\overline{x}|\big(1+|\overline{p}|+|\overline{X}|\big) - \omega(|x-\overline{x}|) - C|p-\overline{p}| - C|X-\overline{X}|
\end{aligned}
$$

for $\overline{p} = D\varphi(\overline{t},\overline{x})$ and $\overline{X} = D^2\varphi(\overline{t},\overline{x})$. Therefore, the quantity in (36) is bounded from below by

$$
\begin{aligned}
& -\partial_t\varphi(\overline{t},\overline{x}) + H(\overline{x}, y, D\varphi(\overline{t},\overline{x}), D_y v_\varepsilon(y), D^2\varphi(\overline{t},\overline{x}), D^2_{yy} v_\varepsilon(y), 0) - o(1) \\
(37)\qquad & \qquad = -\partial_t\varphi(\overline{t},\overline{x}) - \varepsilon v_\varepsilon + \overline{H} - o(1),
\end{aligned}
$$

where $o(1)$ goes to 0 as $(t,x) \to (\overline{t},\overline{x})$ uniformly in ε. Since $\varepsilon v_\varepsilon$ converges uniformly to 0 and since $-\partial_t\varphi(\overline{t},\overline{x}) + \overline{H} > 0$, we can find $r > 0$ so that the quantity is ≥ 0 in Q_r for ε small. We conclude that

$$
-\partial_t\psi_\varepsilon + H\left(x, y, D_x\psi_\varepsilon, \frac{D_y\psi_\varepsilon}{\varepsilon}, D^2_{xx}\psi_\varepsilon, \frac{D^2_{yy}\psi_\varepsilon}{\varepsilon}, \frac{D^2_{xy}\psi_\varepsilon}{\sqrt{\varepsilon}}\right) \ge 0 \quad \text{in } Q_r.
$$

The inequality was derived a bit formally. Using the smoothness of φ, it is an easy exercise to check that the inequality holds in the viscosity sense (see section 1).

The modifications for the cases (I) and (II) are analogous to those performed in Proposition 12. We only sketch them here. When b is independent of x (case (II)), the Hamiltonian now satisfies (29), where the additional q term appears. In (37) there is therefore the extra term $|x-\overline{x}|\,|D_y v_\varepsilon|$. By the coercivity of H, we know that $|D_y v_\varepsilon|$ is bounded uniformly in y and ε by $C(1+|\overline{p}|+|\overline{X}|)$ (see (30)). So the extra term converges uniformly on ε and y to 0 as $x \to \overline{x}$. The above argument therefore applies and guarantees the existence of a small $r > 0$ so that ψ_ε is a supersolution in Q_r for ε small.

When both g and b may depend on x (case (I)), we must use the inequality (31) for the Hamiltonian. But one now controls $D_y v_\varepsilon$ and $D^2_{yy} v_\varepsilon$ uniformly on ε (see (32)). Thus, the extra term $|x-\overline{x}|(|D_y v_\varepsilon| + |D^2_{yy} v_\varepsilon|)$ converges uniformly to 0 as $x \to \overline{x}$, and the argument still works. □

3. Homogenization and stochastic control.

3.1. Homogenization. In the case of a periodic fast variable, a special singular perturbation problem is homogenization. We briefly illustrate this and refer to [19, 20, 8, 25, 26, 1, 38] for recent developments in the theory of homogenization of H–J equations, which was introduced by Lions, Papanicolaou, and Varadhan [35]. For an optimal control problem, homogenization corresponds to dynamics of the form

$$
dx_s = f\left(x_s, \frac{x_s}{\varepsilon}, \alpha_s\right) ds + \sigma\left(x_s, \frac{x_s}{\varepsilon}, \alpha_s\right) dW_s
$$

and the value function

$$v_\varepsilon(t,x) = \inf_\alpha E\left\{\int_t^T l\left(x_s, \frac{x_s}{\varepsilon}, \alpha_s\right)\, ds + h(x_T) \mid x_t = x\right\}.$$

All the functions are of course assumed to be periodic in the second variable. Adding the new variable $y_s = x_s/\varepsilon$, the dynamical system becomes

$$\begin{aligned} dx_s &= f(x_s, y_s, \alpha_s)\, ds + \sigma(x_s, y_s, \alpha_s)\, dW_s, \\ dy_s &= \varepsilon^{-1} f(x_s, y_s, \alpha_s)\, ds + \varepsilon^{-1}\sigma(x_s, y_s, \alpha_s)\, dW_s \end{aligned} \tag{38}$$

with starting point $x_t = x$ and $y_t = x/\varepsilon$.

When the problem is deterministic ($\sigma \equiv 0$) or when there is no drift ($f \equiv 0$), the value function v_ε can be expressed in terms of the value function u_ε of the singular perturbation problem of the preceding section (with $g \equiv f$ and $\tau \equiv \sigma$) as follows:

$$v_\varepsilon(t,x) = u_\varepsilon\left(t, x, \frac{x}{\varepsilon}\right) \qquad \text{and} \qquad v_\varepsilon(t,x) = u_{\varepsilon^2}\left(t, x, \frac{x}{\varepsilon}\right), \qquad \text{respectively.}$$

The convergence of u_ε to the solution of the limit equation (which will be uniform in y by periodicity) of course implies the convergence of v_ε to the same limit. In general, however, the scaling in (38) differs from that of the previous section. We explain briefly how the results can be adapted.

3.2. The associated singular perturbation problem. For $\varepsilon > 0$ fixed, we therefore consider a finite horizon stochastic control problem in $(0,T] \times \mathbb{R}^n \times \mathbb{R}^m$ similar to the one of the preceding section but with the dynamics

$$\begin{aligned} dx_s &= f(x_s, y_s, \alpha_s)\, ds + \sigma(x_s, y_s, \alpha_s)\, dW_s, \\ dy_s &= \varepsilon^{-1} g(x_s, y_s, \alpha_s)\, ds + \varepsilon^{-1}\tau(x_s, y_s, \alpha_s)\, dW_s. \end{aligned}$$

The value function

$$u_\varepsilon(t,x,y) = \inf_\alpha E\left\{\int_t^T l(x_s, y_s, \alpha_s)\, ds + h(x_T)\right\}$$

is now the unique bounded continuous viscosity solution of the H–J–B equation

$$\begin{cases} -\partial_t u_\varepsilon + H\left(x, y, D_x u_\varepsilon, \dfrac{D_y u_\varepsilon}{\varepsilon}, D^2_{xx} u_\varepsilon, \dfrac{D^2_{yy} u_\varepsilon}{\varepsilon^2}, \dfrac{D^2_{xy} u_\varepsilon}{\varepsilon}\right) = 0 \quad \text{in } (0,T) \times \mathbb{R}^n \times \mathbb{R}^m, \\ u_\varepsilon(T, \cdot) = h \quad \text{on } \mathbb{R}^n \times \mathbb{R}^m \end{cases} \tag{39}$$

for the Hamiltonian of the preceding section.

In an attempt to apply the method of the preceding section, we were lead to assume that $\sigma \equiv 0$ in case (II) and $g \equiv 0$ in case (III). As explained above, the scaling is unchanged under one of these assumptions. The new result therefore concerns case (I) in which drifts and diffusions appear. We recall the assumption for convenience.

(I) The diffusions in the fast variable are uniformly nondegenerate and the running cost $l(x, \cdot, \alpha)$ is Hölder continuous uniformly on (x, α).

THEOREM 15. *Assume that* (I) *holds and let* (x, p, X) *be fixed. For every* $\delta > 0$, *let* w_δ *denote the unique viscosity solution of the stationary problem in the fast variable*

$$\delta w_\delta + H\left(x, y, p, 0, X, D^2_{yy} w_\delta, 0\right) = 0 \quad \textit{in } \mathbb{R}^m, \qquad w_\delta \textit{ periodic.}$$

Then, as $\delta \to 0+$, the family $\{\delta w_\delta\}$ converges to a constant $-\overline{H}(x,p,X)$ uniformly with respect to y.

The theorem is a special case of Theorem 10 of the preceding section, because the actual Hamiltonian in the theorem corresponds to the original one with $g \equiv 0$. The new effective Hamiltonian therefore has the regularity stated in Propositions 11 and 12. As a consequence, the limit equation (35) has a unique bounded continuous viscosity solution under either (IV) or (V) or (VI), provided we drop the reference to g in this last condition:

(VI) the diffusions in the fast variable are independent of x ($b \equiv b(y,\alpha)$).

THEOREM 16. *Assume* (I) *and either* (IV) *or* (V) *or* (VI). *Then, as $\varepsilon \to 0+$, the collection $\{u_\varepsilon\}$ converges uniformly on the compact subsets of $(0,T]\times\mathbb{R}^n\times\mathbb{R}^m$ to the unique viscosity solution u of* (35).

Proof. We keep the notations of the proof of Theorem 14 and mention only the changes. To prove that $\overline{u}$ is a subsolution in $(0,T)\times\mathbb{R}^n$, we consider a strict maximum point $(\bar t,\bar x)\in(0,T)\times\mathbb{R}^n$ of $\overline{u}(t,x)-\varphi(t,x)$ with $\overline{u}(\bar t,\bar x)=\varphi(\bar t,\bar x)$ and assume that $-\partial_t\varphi(\bar t,\bar x)+\overline{H}>0$ for $\overline{H}=\overline{H}(\bar x,D\varphi(\bar t,\bar x),D^2\varphi(\bar t,\bar x))$. If v_ε is the periodic solution of

$$\varepsilon^2 v_\varepsilon + H\left(\bar x, y, D\varphi(\bar t,\bar x), 0, D^2\varphi(\bar t,\bar x), D^2_{yy}v_\varepsilon, 0\right) = \overline{H} \quad \text{in } \mathbb{R}^m,$$

the family $\{\varepsilon^2 v_\varepsilon\}$ converges uniformly to 0. A contradiction is achieved by showing that the perturbed test function

$$\psi_\varepsilon(t,x,y) = \varphi(t,x) + \varepsilon^2 v_\varepsilon(y)$$

is a supersolution of (22) in $Q_r=(\bar t-r,\bar t+r)\times B(\bar x,r)\times\mathbb{R}^m$ for some $r>0$ and for ε small. For every $(t,x,y)\in Q_r$, we compute

$$\begin{aligned}
&-\partial_t\psi_\varepsilon + H\left(x,y,D_x\psi_\varepsilon,\frac{D_y\psi_\varepsilon}{\varepsilon},D^2_{xx}\psi_\varepsilon,\frac{D^2_{yy}\psi_\varepsilon}{\varepsilon^2},\frac{D^2_{xy}\psi_\varepsilon}{\varepsilon}\right)\\
&\quad= -\partial_t\varphi(t,x) + H\left(x,y,D\varphi(t,x),\varepsilon D_y v_\varepsilon(y),D^2\varphi(t,x),D^2_{yy}v_\varepsilon(y),0\right)\\
&\quad\ge -\partial_t\varphi(t,x) + H\left(x,y,D\varphi(t,x),0,D^2\varphi(t,x),D^2_{yy}v_\varepsilon(y),0\right) - C\varepsilon|D_y v_\varepsilon|,
\end{aligned}$$

where $C=\|g\|_{L^\infty}$. The term $\varepsilon|D_y v_\varepsilon|$ converges uniformly to 0 as $\varepsilon\to 0$ because we have the bound $\|D_y v_\varepsilon\|_{L^\infty}\le C$ under (I) (see (32)). The remaining two terms correspond to the case $g\equiv 0$; they can be handled as in the proof of Theorem 14. We conclude that the expression is ≥ 0 in Q_r for some $r>0$ and for ε small, so that ψ_ε is a supersolution. This completes the proof. ∎

Acknowledgments. We are grateful to Z. Artstein for sending us the first and second drafts of the preprint [5] that stimulated this research and to V. Gaitsgory for useful talks on the singular perturbation literature and on [23].

REFERENCES

[1] O. ALVAREZ, *Homogenization of Hamilton-Jacobi equations in perforated sets*, J. Differential Equations, 159 (1999), pp. 543–577.

[2] M. ARISAWA AND P.-L. LIONS, *On ergodic stochastic control*, Comm. Partial Differential Equations, 23 (1998), pp. 2187–2217.

[3] Z. ARTSTEIN, *Invariant measures of differential inclusions applied to singular perturbations*, J. Differential Equations, 152 (1999), pp. 289–307.

[4] Z. Artstein and V. Gaitsgory, *Tracking fast trajectories along a slow dynamics: A singular perturbations approach*, SIAM J. Control Optim., 35 (1997), pp. 1487–1507.
[5] Z. Artstein and V. Gaitsgory, *The value function of singularly perturbed control systems*, Appl. Math. Optim., 41 (2000), pp. 425–445.
[6] F. Bagagiolo and M. Bardi, *Singular perturbation of a finite horizon problem with state-space constraints*, SIAM J. Control Optim., 36 (1998), pp. 2040–2060.
[7] F. Bagagiolo, M. Bardi, and I. Capuzzo-Dolcetta, *A viscosity solutions approach to some asymptotic problems in optimal control*, in Partial Differential Equations Methods in Control and Shape Analysis, G. Da Prato and J.-P. Zolesio, eds., Marcel Dekker, New York, 1997, pp. 27–37.
[8] M. Bardi, *Homogenization of quasilinear elliptic equations with possibly superquadratic growth*, in Nonlinear Variational Problems and Partial Differential Equations (Isola d'Elba, 1990), Pitman Res. Notes Math. Ser. 320, Longman, Harlow, U.K., 1995, pp. 44–56.
[9] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser Boston, Cambridge, MA, 1997.
[10] M. Bardi and P. Soravia, *A comparison result for Hamilton-Jacobi equations and applications to differential games lacking controllability*, Funkcial. Ekvac., 37 (1994), pp. 19–43.
[11] G. Barles, *Solutions de Viscosité des Equations de Hamilton-Jacobi*, Math. Appl. 17, Springer-Verlag, Paris, 1994.
[12] A. Bensoussan, *Perturbation Methods in Optimal Control*, Wiley/Gauthiers–Villars, Chichester, U.K., 1988.
[13] A. Bensoussan, L. Boccardo, and F. Murat, *Homogenization of elliptic equations with principal part not in divergence form and Hamiltonian with quadratic growth*, Comm. Pure Appl. Math., 39 (1986), pp. 769–805.
[14] A. Bensoussan, J.-L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, North–Holland, Amsterdam, The Netherlands, 1978.
[15] S. Bortoletto, *The Bellman equation for constrained deterministic optimal control problems*, Differential Integral Equations, 6 (1993), pp. 905–924.
[16] I. Capuzzo-Dolcetta and P.-L. Lions, *Hamilton-Jacobi equations with state constraints*, Trans. Amer. Math. Soc., 318 (1990), pp. 643–683.
[17] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.
[18] A. Dontchev and T. Zolezzi, *Well-Posed Optimization Problems*, Lecture Notes in Math., Springer-Verlag, Berlin, 1993.
[19] L. Evans, *The perturbed test function method for viscosity solutions of nonlinear PDE*, Proc. Roy. Soc. Edinburgh Sect. A, 111 (1989), pp. 359–375.
[20] L. Evans, *Periodic homogenisation of certain fully nonlinear partial differential equations*, Proc. Roy. Soc. Edinburgh Sect. A, 120 (1992), pp. 245–265.
[21] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, Berlin, 1993.
[22] V. Gaitsgory, *Suboptimization of singularly perturbed control systems*, SIAM J. Control Optim., 30 (1992), pp. 1228–1249.
[23] V. Gaitsgory and A. Leizarowitz, *Limit occupational measures set for a control system and averaging of singularly perturbed control systems*, J. Math. Anal. Appl., 233 (1999), pp. 461–475.
[24] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, New York, 1983.
[25] K. Horie and H. Ishii, *Homogenization of Hamilton-Jacobi equations on domains with small scale periodic structure*, Indiana Univ. Math. J., 47 (1998), pp. 1011–1058.
[26] H. Ishii, *Homogenization of the Cauchy problem for Hamilton-Jacobi equations*, in Stochastic Analysis, Control, Optimization and Applications, Systems Control Found. Appl., Birkhäuser Boston, Cambridge, MA, 1999, pp. 305–324.
[27] H. Ishii and S. Koike, *A new formulation of state constraints problems for first-order PDEs*, SIAM J. Control Optim., 34 (1996), pp. 554–571.
[28] H. Ishii and P.-L. Lions, *Viscosity solutions of fully nonlinear second-order elliptic partial differential equations*, J. Differential Equations, 83 (1990), pp. 26–78.
[29] R. Jensen and P.-L. Lions, *Some asymptotic problems in fully nonlinear elliptic equations and stochastic control*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 11 (1984), pp. 129–176.
[30] Y. Kabanov and S. Pergamenshchikov, *On convergence of attainability sets for controlled two-scale stochastic linear systems*, SIAM J. Control Optim., 35 (1997), pp. 134–159.
[31] Y. Kabanov and W. Runggaldier, *On control of two-scale stochastic systems with linear dynamics in the fast variables*, Math. Control Signals Systems, 9 (1996), pp. 107–122.

[32] P. Kokotović, H. Khalil, and J. O'Reilly, *Singular Perturbation Methods in Control: Analysis and Design*, Academic Press, London, 1986.
[33] H. Kushner, *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*, Birkhäuser Boston, Cambridge, MA, 1990.
[34] P.-L. Lions, *Generalized Solutions of Hamilton-Jacobi Equations*, Pitman, Boston, 1982.
[35] P.-L. Lions, G. Papanicolaou, and S. R. S. Varadhan, *Homogeneization of Hamilton-Jacobi Equations*, manuscript, 1986.
[36] M. V. Safonov, *On the classical solution of nonlinear elliptic equations of second-order*, Math. USSR-Izv., 33 (1989), pp. 597–612; English translation of Izv. Akad. Nauk SSSR Ser. Mat., 52 (1988), pp. 1272–1287.
[37] H. M. Soner, *Optimal control with state-space constraint* I, SIAM J. Control Optim., 24 (1986), pp. 552–561.
[38] P. E. Souganidis, *Stochastic homogenization of Hamilton-Jacobi equations and some applications*, Asympt. Anal., 20 (1999), pp. 1–11.
[39] N. Subbotina, *Asymptotic properties of minimax solutions of Isaacs-Bellman equations in differential games with fast and slow motions*, J. Appl. Math. Mech., 60 (1996), pp. 883–890.
[40] V. Veliov, *A generalization of the Tikhonov theorem for singularly perturbed differential inclusions*, J. Dynam. Control Systems, 3 (1997), pp. 291–319.
[41] A. Vigodner, *Limits of singularly perturbed control problems with statistical dynamics of fast motions*, SIAM J. Control Optim., 35 (1997), pp. 1–28.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 4, pp. 1189–1226

MINIMAX LQG CONTROL OF STOCHASTIC PARTIALLY OBSERVED UNCERTAIN SYSTEMS*

VALERY A. UGRINOVSKII† AND IAN R. PETERSEN†

Abstract. We consider an infinite-horizon linear-quadratic minimax optimal control problem for stochastic uncertain systems with output measurement. A new description of stochastic uncertainty is introduced using a relative entropy constraint. For the stochastic uncertain system under consideration, a connection between the worst-case control design problem and a specially parametrized risk-sensitive stochastic control problem is established. Using this connection, a minimax optimal LQG controller is constructed which is based on a pair of algebraic matrix Riccati equations arising in risk-sensitive control. It is shown that this minimax optimal controller absolutely stabilizes the stochastic uncertain system.

Key words. robust control, LQG control, stochastic control, stochastic risk-sensitive control, stochastic dynamic games

AMS subject classifications. 93E20, 93E05, 93C41

PII. S0363012998349352

1. Introduction.

One of the important ideas in modern robust control theory emerges from the fact that many robust control problems can be formulated as optimization problems. The advantage of this approach is that it allows one to readily convert a problem of robust controller design into a mathematically tractable game-type minimax optimization problem. For linear systems with full state measurement, this methodology leads to a robust version of the linear quadratic regulator (LQR) approach to state feedback controller design [15, 18]. However, the development of a robust version of the LQG technique appears to be a challenging problem. The problem becomes especially difficult in situations in which one wishes to take into account the fact that in real physical systems, noise disturbances entering into the controlled plant differ from Gaussian white noise. A suitable way of introducing noise disturbances in this case may be to treat the disturbances as uncertain stochastic processes. A formalization of this idea leads to the concept of an *uncertain stochastic system* introduced in recent papers [11, 12, 19].

Note that in the case of a finite time horizon, the uncertain systems framework introduced in [12, 19] allows one to extend the standard LQG design methodology into a partial information minimax optimal control methodology for stochastic uncertain systems. The problem considered in [12, 19] involves constructing a controller which minimizes worst-case performance in the face of system uncertainty which satisfies a certain stochastic uncertainty constraint. This constraint restricts the *relative entropy* between an uncertain probability measure related to the distribution of the uncertainty input and the reference probability measure. This relative entropy constraint can be thought of as a stochastic counterpart of the deterministic integral quadratic constraint uncertainty description; see [15, 23]. One advantage of the relative entropy uncertainty description is that it allows for stochastic uncertainty inputs

*Received by the editors December 17, 1998; accepted for publication (in revised form) April 16, 2001; published electronically November 28, 2001. This work was supported by the Australian Research Council.

http://www.siam.org/journals/sicon/40-4/34935.html

†School of Electrical Engineering, Australian Defence Force Academy, Canberra ACT 2600, Australia (valu@ee.adfa.edu.au, irp@ee.adfa.edu.au).

to depend dynamically on the uncertainty outputs.

In this paper, we address an infinite-horizon version of the robust LQG problems considered in [12, 19]. As we proceed from a finite time interval to an infinite time interval, the fact that the systems under consideration are those with additive noise becomes important. The solutions of such systems do not necessarily belong to $L_2[0,\infty)$. Hence, the approaches used to describe admissible uncertainties in the deterministic case (e.g., see [15]) and the multiplicative noise case [18] are not applicable here. Note that the class of admissible uncertainties defined using the approach of [15, 18] is consistent with the notion of absolute stabilizability defined in terms of the $L_2[0,\infty)$-summability of uncertainty inputs and corresponding solutions to the closed-loop system. However, in the present paper the uncertainty inputs and solutions need not be $L_2[0,\infty)$-summable. Instead, we will consider the time-averaged properties of the system solutions. This requires us to correspondingly modify the definitions of admissible uncertainty and absolute stabilizability in order to properly account for the nature of the systems under consideration. In particular, our new definition of the class of admissible uncertainties is one of the contributions of this paper. In the case of an uncertain system with additive noise considered on the infinite time interval, we use an approximating sequence of martingales to describe the class of admissible uncertainties. In particular, we give an example which shows that H^∞ norm-bounded uncertainty can be incorporated into the proposed framework by constructing a corresponding sequence of martingales.

The main result of the paper is a robust LQG control synthesis procedure based on a pair of algebraic Riccati equations arising in risk-sensitive optimal control; see [9]. We show that solutions to a certain specially parametrized risk-sensitive control problem provide us with a controller which guarantees an optimal upper bound on the time-averaged performance of the closed-loop system in the presence of admissible uncertainties.

2. Definitions. Let $(\Omega,\mathcal{F},P)$ be a complete probability space on which a p-dimensional standard Wiener process $W(\cdot)$ and a Gaussian random variable $x_0\colon \Omega\to\mathbf{R}^n$ with mean $\check{x}_0$ and nonsingular covariance matrix Y_0 are defined, $p=r+l$. The first r entries of the vector process $W(\cdot)$ correspond to the system noise, while the last l entries correspond to the measurement noise. The space Ω can be thought of as the noise space $\mathbf{R}^n\times\mathbf{R}^l\times C([0,\infty),\mathbf{R}^p)$ [1]. The probability measure P can then be defined as the product of a given probability measure on $\mathbf{R}^n\times\mathbf{R}^l$ and the standard Wiener measure on $C([0,\infty),\mathbf{R}^p)$. The space Ω is endowed with a filtration $\{\mathcal{F}_t, t\ge 0\}$ which has been completed by including all sets of probability zero. The filtration $\{\mathcal{F}_t, t\ge 0\}$ can be thought of as the filtration generated by the mappings $\{\Pi_t, t\ge 0\}$, where $\Pi_0(x,\eta,W(\cdot))=(x,\eta)$ and $\Pi_t(x,\eta,W(\cdot))=W(t)$ for $t>0$ [1]. The random variable x_0 and the Wiener process $W(\cdot)$ are stochastically independent in $(\Omega,\mathcal{F},P)$.

2.1. The nominal system. On the probability space $(\Omega,\mathcal{F},P)$ defined above, we consider the system and measurement dynamics driven by the noise input $W(\cdot)$ and a control input $u(\cdot)$. These dynamics are described by the following stochastic differential equation:

$$
\begin{aligned}
dx(t) &= (Ax(t)+B_1u(t))dt + B_2dW(t), \qquad x(0)=x_0,\\
z(t) &= C_1x(t)+D_1u(t),\\
dy(t) &= C_2x(t)dt + D_2dW(t), \qquad y(0)=0.
\end{aligned}
\tag{1}
$$

In the above equations, $x(t) \in \mathbf{R}^n$ is the state, $u(t) \in \mathbf{R}^m$ is the control input, $z(t) \in \mathbf{R}^q$ is the uncertainty output, and $y(t) \in \mathbf{R}^l$ is the measured output. System (1) is referred to as the nominal system. All coefficients in (1) are assumed to be constant matrices of corresponding dimensions. Also, we assume that $D_2 D_2' > 0$.

In the minimax optimal control problem to be considered in this paper, our attention will be restricted to linear output-feedback controllers of the form

$$\begin{aligned} d\hat{x} &= A_c \hat{x} + B_c dy, \\ u &= K\hat{x}, \end{aligned} \tag{2}$$

where $\hat{x} \in \mathbf{R}^{\hat{n}}$ is the state of the controller and $A_c \in \mathbf{R}^{\hat{n}\times\hat{n}}$, $K \in \mathbf{R}^{m\times\hat{n}}$, and $B_c \in \mathbf{R}^{\hat{n}\times q}$. Let $\mathcal{U}$ denote this class of linear controllers. Note that the controller (2) is adapted to the filtration $\{\mathcal{Y}_t, t \geq 0\}$ generated by the observation process y. The closed-loop nominal system corresponding to controller (2) is described by a linear Ito differential equation of the form

$$\begin{aligned} d\bar{x} &= \bar{A}\bar{x}dt + \bar{B}dW(t), \\ z &= \bar{C}\bar{x}, \\ u &= \begin{bmatrix} 0 & K \end{bmatrix} \bar{x} \end{aligned} \tag{3}$$

and is considered on the probability space $(\Omega, \mathcal{F}, P)$. In (3), $\bar{x} = [x' \ \hat{x}']' \in \mathbf{R}^{n+\hat{n}}$ is the state of the closed-loop system. Also, the following notation is used:

$$\bar{A} = \begin{bmatrix} A & B_1 K \\ B_c C_2 & A_c \end{bmatrix}, \qquad \bar{B} = \begin{bmatrix} B_2 \\ B_c D_2 \end{bmatrix}, \qquad \bar{C} = \begin{bmatrix} C_1 & D_1 K \end{bmatrix}. \tag{4}$$

2.2. The stochastic uncertain system. In this paper, we introduce an uncertainty description for stochastic uncertain systems with additive noise which can be regarded as an extension of the uncertainty description considered in [12, 19] to the case of an infinite time horizon. As in [12, 19], the stochastic uncertain systems to be considered are described by the nominal system (1) considered over the probability space $(\Omega, \mathcal{F}, P)$, and also by a set of perturbations of the reference probability measure P. These perturbations are defined as follows. Consider the set $\mathcal{M}$ of continuous positive martingales $(\zeta(t), \mathcal{F}_t, t \geq 0)$ such that for each $T \geq 0$, $\mathbf{E}\zeta(T) = 1$; here, $\mathbf{E}$ denotes the expectation with respect to the probability measure P. Note that the set $\mathcal{M}$ is convex.

Every martingale $\zeta(\cdot) \in \mathcal{M}$ gives rise to a probability measure Q^T on the measurable space $(\Omega, \mathcal{F}_T)$ defined by the equation

$$Q^T(d\omega) = \zeta(T) P^T(d\omega). \tag{5}$$

Here, P^T denotes the restriction of the reference probability measure P to $(\Omega, \mathcal{F}_T)$. From this definition, for every $T > 0$, the probability measure Q^T is absolutely continuous with respect to the probability measure P^T, $Q^T \ll P^T$. The uncertain system is described by the stochastic differential equation (1) considered over the probability space $(\Omega, \mathcal{F}_T, Q^T)$ for every $T > 0$. The expectation in this probability space is denoted $\mathbf{E}^{Q^T}$.

We now present an infinite-horizon uncertainty description for stochastic uncertain systems with additive noise. This uncertainty description may be regarded as an extension of the uncertainty description considered in [19] to the infinite-horizon case. Also, this uncertainty description can be thought of as an extension of the

deterministic integral quadratic constraint uncertainty description [15, 16, 23] and the stochastic integral quadratic constraint uncertainty description [18] to the case of stochastic uncertain systems with additive noise. Recall that the integral quadratic constraints arising in [15, 16, 23, 18] exploit a sequence of times $\{t_i\}_{i=1}^{\infty}$ to "localize" the uncertainty inputs and uncertainty outputs to time intervals $[0, t_i]$. The consideration of the system dynamics on these finite time intervals then allows one to deal with bounded energy processes. However, in this paper the systems under consideration are those with additive noise. For this class of stochastic systems, it is natural to consider bounded power processes rather than bounded energy processes. This motivates us to propose the relative entropy uncertainty description given below in Definition 1 to accommodate bounded power processes.

In contrast to the case of deterministic integral quadratic constraints, the uncertainty description considered in this paper exploits a sequence of continuous positive martingales $\{\zeta_i(t), \mathcal{F}_t, t \geq 0\}_{i=1}^{\infty} \subset \mathcal{M}$ which converges to a limiting martingale $\zeta(\cdot)$ in the following sense: For any $T > 0$, the sequence $\{\zeta_i(T)\}_{i=1}^{\infty}$ converges weakly to $\zeta(T)$ in $L_1(\Omega, \mathcal{F}_T, P^T)$. Using the martingales $\zeta_i(t)$, we define a sequence of probability measures $\{Q_i^T\}_{i=1}^{\infty}$ as follows:

$$Q_i^T(d\omega) = \zeta_i(T) P^T(d\omega). \tag{6}$$

From the definition of the martingales $\zeta_i(t)$, it follows that for each $T > 0$ the sequence $\{Q_i^T\}_{i=1}^{\infty}$ converges to the probability measure Q^T corresponding to a limiting martingale $\zeta(\cdot)$ in the following sense: For any bounded $\mathcal{F}_T$-measurable random variable η,

$$\lim_{i \to \infty} \int_{\Omega} \eta Q_i^T(d\omega) = \int_{\Omega} \eta Q^T(d\omega). \tag{7}$$

We denote this fact by $Q_i^T \Rightarrow Q^T$ as $i \to \infty$.

Remark 1. The property $Q_i^T \Rightarrow Q^T$ implies that the sequence of probability measures Q_i^T converges weakly to the probability measure Q^T. Indeed, consider the Polish space of probability measures on the measurable space $(\Omega, \mathcal{F}_T)$ endowed with the topology of weak convergence of probability measures. Note that Ω is a metric space. Hence, such a topology can be defined; e.g., see [2]. For the sequence $\{Q_i^T\}$ to converge weakly to Q^T, it is required that (7) hold for all bounded continuous random variables η. Obviously, this requirement is satisfied if $Q_i^T \Rightarrow Q^T$.

As in the finite-horizon case [12, 19], we describe the class of admissible uncertainties in terms of the relative entropy functional $h(\cdot\|\cdot)$; for the definition and properties of the functional $h(\cdot\|\cdot)$, see Appendix A and also [2].

DEFINITION 1. *Let d be a given positive constant. A martingale $\zeta(\cdot) \in \mathcal{M}$ is said to define an admissible uncertainty if there exists a sequence of continuous positive martingales $\{\zeta_i(t), \mathcal{F}_t, t \geq 0\}_{i=1}^{\infty} \subset \mathcal{M}$ which satisfies the following conditions:*

(i) *For each i, $h(Q_i^T \| P^T) < \infty$ for all $T > 0$;*

(ii) *For all $T > 0$, $Q_i^T \Rightarrow Q^T$ as $i \to \infty$;*

(iii) *The following* stochastic uncertainty constraint *is satisfied: For any sufficiently large $T > 0$, there exists a constant $\delta(T)$ such that $\lim_{T \to \infty} \delta(T) = 0$ and*

$$\inf_{T' > T} \frac{1}{T'} \left[\frac{1}{2} \mathbf{E}^{Q_i^{T'}} \int_0^{T'} \|z(t)\|^2 dt - h(Q_i^{T'} \| P^{T'}) \right] \geq -\frac{d}{2} + \delta(T) \tag{8}$$

for all $i = 1, 2, \ldots$. In (8), the uncertainty output $z(\cdot)$ is defined by (1) considered on the probability space $(\Omega, \mathcal{F}_T, Q_i^T)$.

In the above conditions, Q_i^T is the probability measure defined by (6) *corresponding to the martingale $\zeta_i(t)$ and time $T > 0$. We let Ξ denote the set of martingales $\zeta(\cdot) \in \mathcal{M}$ corresponding to admissible uncertainties. Elements of Ξ are also called admissible martingales.*

Observe that the reference probability measure P corresponds to the admissible martingale $\zeta(t) \equiv 1$. Hence, the set Ξ is not empty. Indeed, choose $\zeta_i(t) = 1$ for all i and t. Then $Q_i^T = P^T$ for all i. It follows from the identity $h(P^T \| P^T) = 0$ that

$$\inf_{T'>T} \frac{1}{T'} \left[\frac{1}{2} \mathbf{E}^{Q_i^{T'}} \int_0^{T'} \|z(t)\|^2 dt - h(Q_i^{T'} \| P^{T'}) \right] = \inf_{T'>T} \frac{1}{2T'} \mathbf{E} \int_0^{T'} \|z(t)\|^2 dt.$$

Note that the expectations are well defined. Also, the infimum on the right-hand side of the above equation is nonnegative for any $T > 0$. Therefore, for any constant $d > 0$, one can find a sufficiently small $\delta = \delta(T)$ such that $\lim_{T\to\infty} \delta(T) = 0$ and the constraint (8) is satisfied strictly in this case.

Remark 2. Note that condition (8) implies that

$$\liminf_{T\to\infty} \frac{1}{T} \left[\frac{1}{2} \mathbf{E}^{Q_i^T} \int_0^T \|z(t)\|^2 dt - h(Q_i^T \| P^T) \right] \geq -\frac{d}{2}$$

for all $i = 1, 2, \ldots$.

In what follows, we will use the following notation. Let $\mathcal{P}_T$ be the set of probability measures Q^T on $(\Omega, \mathcal{F}_T)$ such that $h(Q^T \| P^T) < \infty$. Also, the notation $\mathcal{M}_\infty$ will denote the set of martingales $\zeta(\cdot) \in \mathcal{M}$ such that $h(Q^T \| P^T) < \infty$ for all $T > 0$. It is readily verified that the set $\mathcal{M}_\infty$ is convex. Note that the martingales $\zeta_i(\cdot)$ from Definition 1 belong to $\mathcal{M}_\infty$.

2.3. A discussion of the class of stochastic uncertain systems under consideration. In this section, we give more insight into the class of stochastic uncertain systems under consideration. In the integral quadratic constraint approach to robust control theory, the uncertainty is described in terms of a given set of uncertainty input signals. In contrast, Definition 1 presents a martingale uncertainty description or, equivalently, a probability measure uncertainty description. The motivation behind Definition 1 is as follows. The proposed uncertain system model allows us to obtain a tractable solution to the corresponding problem of minimax optimal LQG controller design. Also, the stochastic uncertainty description presented in Definition 1 encompasses many important classes of uncertainty arising in robust control theory. In particular, it includes H^∞ norm-bounded linear time-invariant (LTI) uncertainties and cone-bounded nonlinear uncertainties. This makes the approach developed in this paper applicable to a broad range of control system design problems. We show below that H^∞ norm-bounded uncertainties satisfy the requirements of Definition 1.

The definition of admissible uncertainties given above involves a collection of martingales $\{\zeta_i(\cdot)\}_{i=1}^\infty$ which has a given uncertainty martingale $\zeta(\cdot)$ as its limit point. In the deterministic case and the multiplicative noise case, similar approximations have been defined by restricting uncertainty inputs to finite time intervals and then extending the restricted processes by zero beyond these intervals; e.g., see [15, 16, 18]. In the case of a stochastic uncertain system with additive noise considered on an infinite time interval, we apply a similar idea. However, in contrast to the deterministic and multiplicative noise cases, we use a sequence of martingales and corresponding probability measures in Definition 1. This procedure may be thought of as involving

a spatial restriction rather than the temporal restriction used previously. Indeed, a natural way to define the required sequence of martingales and corresponding probability measures is to consider martingales corresponding to the uncertainty inputs as "truncated" at certain Markov times $\mathfrak{t}_i$. For example, this can be achieved by choosing an expanding sequence of compact sets K_i in the uncertainty input space and letting $\mathfrak{t}_i$ be the Markov time when the uncertainty input reaches the boundary of the set K_i. In this case, we focus on spatial domains rather than time intervals on which the uncertainty inputs and uncertainty outputs are then constrained. An illustration of this idea will be given in section 2.3.2.

2.3.1. A connection between uncertainty input signals and martingale uncertainty. A connection between the uncertainty input signal uncertainty model and the perturbation martingale uncertainty model is based on Novikov's theorem [6]. Using the result of Novikov's theorem, a given uncertainty input $\xi(\cdot)$ satisfying the conditions of this theorem on every finite interval $[0,T]$ can be associated with a martingale $\zeta(\cdot) \in \mathcal{M}$. This result is summarized in the following lemma.

LEMMA 1. *Suppose a random process $(\xi(t), \mathcal{F}_t)$, $0 \le t \le T$, satisfies the conditions:*

$$P\left(\int_0^T \|\xi(s)\|^2 ds < \infty\right) = 1,$$

$$\mathbf{E} \exp\left(\frac{1}{2}\int_0^T \|\xi(s)\|^2 ds\right) < \infty. \tag{9}$$

Then the equation

$$\zeta(t) = 1 + \int_0^t \zeta(s)\xi(s)' dW(s) \tag{10}$$

defines a continuous positive martingale $\zeta(t)$. Furthermore, the stochastic process

$$\tilde{W}(t) = W(t) - \int_0^t \xi(t) dt \tag{11}$$

is a Wiener process with respect to the system $\{\mathcal{F}_t, 0 \le t \le T\}$ and the probability measure Q^T defined by (5)*, where $\zeta(\cdot)$ is defined by* (10)*.*

Proof. Conditions (9) are the conditions of Novikov's theorem (e.g., see Theorem 6.1 on page 216 of [6]). It follows from this theorem that the random process $(\zeta(t), \mathcal{F}_t)$, $0 \le t \le T$, defined by (10) is a continuous martingale and, in particular, $\mathbf{E}\zeta(T) = 1$. Furthermore, this martingale is given by

$$\zeta(t) = \exp\left(\int_0^t \xi'(s) dW(s) - \frac{1}{2}\|\xi(s)\|^2 ds\right). \tag{12}$$

The statement of the lemma now follows from Girsanov's theorem; e.g., see Theorem 6.3 on page 232 of [6]. □

We now consider an uncertain system with H^∞ norm-bounded LTI uncertainty and driven by a Gaussian white noise process $v(t)$ as shown in Figure 1. We will show that such an uncertain system can be described in terms of the stochastic uncertain system framework defined above. Note that if $\Delta(s) \equiv 0$ and $\xi(\cdot) = 0$, then $w(t) = v(t)$.

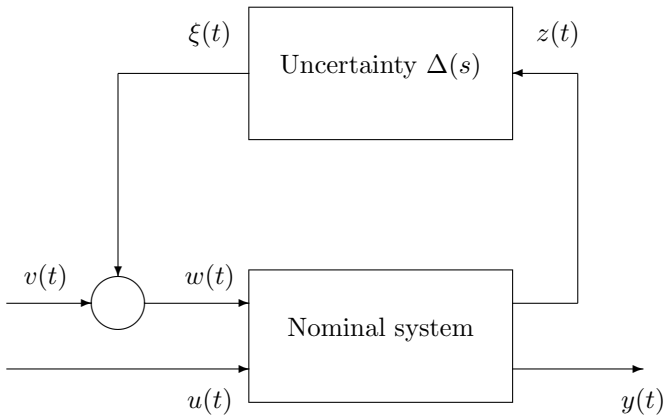


FIG. 1. *An uncertain system.*

That is, the nominal system is driven by a Gaussian white noise. However, in the presence of uncertainty, the input $w(\cdot)$ ceases to be a Gaussian white noise.

For each $T > 0$, a rigorous mathematical description of the system shown in Figure 1 can be given by the equations

$$\begin{aligned} dx &= (Ax + B_1 u + B_2 \xi)dt + B_2 d\tilde{W}(t), \\ z &= C_1 x + D_1 u, \\ dy &= (C_2 x + D_2 \xi)dt + D_2 d\tilde{W}(t) \end{aligned} \tag{13}$$

considered on the probability space $(\Omega, \mathcal{F}_T, Q^T)$, where Q^T is the probability measure constructed in Lemma 1. Also, the uncertainty input is related to the uncertainty output by the relation $\xi = \Delta(s)z$. Now, the substitution of (11) into (13) leads to a set of equations of the form (1) considered on the probability space $(\Omega, \mathcal{F}_T, Q^T)$. Thus, the uncertain system shown in Figure 1 can be considered in the stochastic uncertain system framework defined above. In what follows, we will show that an H^∞ norm bound on the LTI uncertainty $\Delta(s)$ implies the satisfaction of the relative entropy constraint described above. Note that the case $\xi(\cdot) = 0$ corresponds to $\zeta(t) \equiv 1$ and $Q^T = P^T$.

2.3.2. H^∞ norm-bounded uncertainty and the relative entropy constraint. In this section we will show that if the LTI uncertainty $\Delta(s)$ shown in Figure 1 satisfies an H^∞ norm bound, then the corresponding stochastic uncertain system satisfies the relative entropy constraint defined above. This completes the proof of our assertion that the standard H^∞ norm-bounded uncertainty description can be incorporated into the framework of Definition 1. In a similar fashion, one can also show that a cone-bounded nonlinear uncertainty defines an admissible uncertainty according to Definition 1. This proof has been removed for the sake of brevity.

In what follows, we will use the following well-known property of linear stochastic systems. On the probability space $(\Omega, \mathcal{F}, \tilde{P})$, consider the following linear system driven by the Wiener process $\tilde{W}(\cdot)$ and a disturbance input $\xi(t)$, $t \in [0, T]$:

$$d\bar{x} = (\bar{A}\bar{x} + \bar{B}\xi(t))dt + \bar{B}d\tilde{W}(t). \tag{14}$$

PROPOSITION 1. *If for some constant* $\rho > 0$

$$\tilde{\mathbf{E}} \int_0^T \|\xi(t)\|^2 dt \leq \rho, \tag{15}$$

then the corresponding solution to (14) *is mean square bounded on the interval* $[0, T]$. *Here* $\tilde{\mathbf{E}}$ *denotes the expectation with respect to the probability measure* $\tilde{P}$.

Proof. The proof of the proposition follows straightforwardly using standard Lyapunov arguments. □

Consider an uncertain system of the form (1) on the probability space $(\Omega, \mathcal{F}, P)$, driven by a controller (2). Associated with the system (1) and controller (2), consider the disturbance input $\xi(\cdot)$ defined by the convolution operator

$$\xi(t) = \int_0^t g(t-\theta) z(\theta) d\theta \tag{16}$$

corresponding to a given causal uncertainty transfer function $\Delta(s)$ which belongs to the Hardy space H^∞. In (16), $z(\cdot)$ is the uncertainty output of the closed-loop system corresponding to the system (1) and the given controller (2).

LEMMA 2. *Let an uncertainty transfer function* $\Delta(s) \in H^\infty$ *be given which satisfies the norm bound condition*

$$\|\Delta(s)\|_\infty \leq 1. \tag{17}$$

Also, suppose that the random process $(\zeta(t), \mathcal{F}_t)$ *defined by* (10) *is a martingale; here* $\xi(\cdot)$ *is the disturbance input generated by the operator* (16). *Then this martingale satisfies the conditions of Definition* 1.

Remark 3. The requirements of Lemma 2 are satisfied if $\Delta(s)$ is a stable rational transfer function satisfying condition (17). Indeed, in this case one can show that the augmented dynamics $[x'(\cdot), \hat{x}'(\cdot), \eta'(\cdot), z'(\cdot), \xi'(\cdot)]'$ are described by a linear system driven by a Wiener process, with Gaussian initial condition; here η denotes the state of the uncertainty. Hence for any $T > 0$ there exists a constant δ_T such that

$$\sup_{t \leq T} \mathbf{E} \exp(\delta_T \|\xi(t)\|^2) < \infty;$$

see the remark on page 138 of [6]. This implies that $\zeta(t)$ is a martingale; see Example 3 on page 220 of [6]. Hence, any uncertainty described by a stable rational transfer function satisfying condition (17) will belong to the class Ξ of uncertainties admissible for system (1) controlled by a linear output-feedback controller of the form (2).

Proof of Lemma 2. Since the random process $(\zeta(t), \mathcal{F}_t)$, $0 \leq t \leq T$, defined by (10) is a martingale and $\mathbf{E}\zeta(T) = 1$, it follows from Girsanov's theorem that the random process $\tilde{W}(\cdot)$ defined by (11) is a Wiener process with respect to the filtration $\{\mathcal{F}_t, 0 \leq t \leq T\}$ and the probability measure Q^T defined as in (5); see [6]. Note that on the probability space $(\Omega, \mathcal{F}_T, Q^T)$, system (1) becomes a system of the form (13).

To verify that the martingale $\zeta(t)$ corresponding to the H^∞ norm-bounded uncertainty under consideration defines an admissible uncertainty, we need to prove the existence of a sequence of martingales $\{\zeta_i(t)\}_{i=1}^\infty$ satisfying the conditions of Definition 1. To construct such a sequence, consider the following family of Markov stopping times $\{\mathfrak{t}_\rho, \rho > 0\}$ [6]. For any $\rho > 0$, define

$$\mathfrak{t}_\rho := \begin{cases} \inf\{t \geq 0 : \int_0^t \|\xi(s)\|^2 ds > \rho\} & \text{if } \int_0^\infty \|\xi(s)\|^2 ds > \rho, \\ \infty & \text{if } \int_0^\infty \|\xi(s)\|^2 ds \leq \rho. \end{cases}$$

The family $\{\mathfrak{t}_\rho\}$ is monotonically increasing and $\mathfrak{t}_\rho \to \infty$ P-a.s.

We now are in a position to construct an approximating sequence of martingales $\{\zeta_i(t)\}_{i=1}^{\infty}$ using the above sequence of Markov stopping times. First, note that the stochastic integral $\mu(t) := \int_0^t \xi(s)' dW(s)$ defines a local continuous martingale; see Definition 6 on page 69 of [6]. Also, for any stopping time $\mathfrak{t}_\rho$ defined above,

$$\mu(t \wedge \mathfrak{t}_\rho) = \int_0^{t\wedge\mathfrak{t}_\rho} \xi(s)' dW(s) = \int_0^t \xi_\rho(s)' dW(s) = \int_0^t \xi(s)' dW(s \wedge \mathfrak{t}_\rho),$$

where the process $\xi_\rho(\cdot)$ is defined as follows:

$$\xi_\rho(t) = \xi(t)\chi_{\{\mathfrak{t}_\rho \geq t\}}. \tag{18}$$

Here, χ_Λ denotes the indicator function of a set $\Lambda \subseteq \Omega$. In the above definitions, the notation $\mathfrak{t} \wedge t := \min\{t, \mathfrak{t}\}$ is used.

Associated with the positive continuous martingale $\zeta(t)$ and the family of stopping times $\{\mathfrak{t}_\rho, \rho > 0\}$ defined above, consider the stopped process

$$\zeta_\rho(t) = \zeta(t \wedge \mathfrak{t}_\rho).$$

From this definition, $\zeta_\rho(t)$ is a continuous martingale; e.g., see Lemma 3.3 on page 69 of [6]. Furthermore, using the representation (10) of the martingale $\zeta(t)$, it follows that $\zeta_\rho(t)$ is an Ito process with the stochastic differential

$$d\zeta_\rho(t) = \zeta_\rho(t)\xi_\rho'(t)dW(t) = \zeta_\rho(t)d\mu(t \wedge \mathfrak{t}_\rho); \quad \zeta_\rho(0) = 1. \tag{19}$$

From (19), the martingale $\zeta_\rho(t)$ admits the following representation:

$$\zeta_\rho(t) = \exp\left(\int_0^t \xi_\rho' dW(s) - \frac{1}{2}\int_0^t \|\xi_\rho(s)\|^2 ds\right). \tag{20}$$

Also, $\mathbf{E}\zeta_\rho(t) = \mathbf{E}\zeta_\rho(0) = 1$. Hence, $\zeta_\rho(\cdot) \in \mathcal{M}$.

Using the martingale $\zeta_\rho(t)$ defined above, we define probability measures Q_ρ^T on $(\Omega, \mathcal{F}_T)$ as follows:

$$Q_\rho^T(d\omega) = \zeta(T \wedge \mathfrak{t}_\rho)P^T(d\omega).$$

From (20), the relative entropy between the probability measures Q_ρ^T and P^T is given by

$$h(Q_\rho^T \| P^T) = \frac{1}{2}\mathbf{E}^{Q_\rho^T}\int_0^T \|\xi_\rho(s)\|^2 ds = \frac{1}{2}\mathbf{E}^{Q_\rho^T}\int_0^{\mathfrak{t}_\rho \wedge T} \|\xi(s)\|^2 ds. \tag{21}$$

From this equation and from (18), it follows that $h(Q_\rho^T \| P^T) \leq (1/2)\rho < \infty$ for all $T > 0$. Thus, condition (i) of Definition 1 is satisfied.

Also, using part 1 of Theorem 3.7 on page 62 of [6], we observe that for every $T > 0$ the family $\{\zeta(\mathfrak{t}_\rho \wedge T), \rho > 0\}$ is uniformly integrable. Also, since $\mathfrak{t}_\rho \to \infty$ with probability one as $\rho \to \infty$, then $\zeta_\rho(T) \to \zeta(T)$ with probability one. This fact together with the property of uniform integrability of the family $\{\zeta_\rho(T), \rho > 0\}$ implies that

$$\lim_{\rho\to\infty} \mathbf{E}(|\zeta(T \wedge \mathfrak{t}_\rho) - \zeta(T)|\big|\mathcal{G}) = 0 \qquad P\text{-a.s.} \tag{22}$$

for any σ-algebra $\mathcal{G} \subset \mathcal{F}_T$; see the Corollary on page 16 of [6]. We now observe that for any $\mathcal{F}_T$-measurable bounded random variable η with values in $\mathbf{R}$

$$\mathbf{E}|\eta\zeta(T \wedge \mathfrak{t}_\rho) - \eta\zeta(T)| \leq \sup_\omega |\eta| \cdot \mathbf{E}|\zeta(T \wedge \mathfrak{t}_\rho) - \zeta(T)|.$$

Therefore, it follows from the definition of the probability measures Q_ρ^T and Q^T and from (22) that $Q_\rho^T \Rightarrow Q^T$ as $\rho \to \infty$ for all $T > 0$. Thus, we have verified that the family of martingales $\zeta_\rho(t)$ satisfies condition (ii) of Definition 1.

We now consider system (1) on the probability space $(\Omega, \mathcal{F}_T, Q_\rho^T)$. Equivalently, we consider system (13) driven by the uncertainty input $\xi_\rho(t)$ on the probability space $(\Omega, \mathcal{F}_T, Q_\rho^T)$. Note that since $\int_0^T \|\xi_\rho(t)\|^2 \leq \rho$ P-a.s., Proposition 1 implies that the corresponding output $z(\cdot)$ of system (1) satisfies the conditions

$$\mathbf{E}^{Q_\rho^T} \int_0^T \|z(s)\|^2 ds < \infty, \quad \int_0^T \|z(s)\|^2 ds < \infty \quad Q_\rho^T\text{-a.s.} \tag{23}$$

for any $T > 0$. We now use the fact that condition (17) implies that for any pair $(\tilde{z}(\cdot), \tilde{\xi}(\cdot))$, $\tilde{z}(\cdot) \in L_2[0,T]$, $T > 0$, related by (16)

$$\int_0^T \|\tilde{\xi}(t)\|^2 dt \leq \int_0^T \|\tilde{z}(t)\|^2 dt;$$

e.g., see [25]. Hence from this observation and from (23), it follows that the pair $(z(\cdot), \xi(\cdot))$, where $z(\cdot)$ and $\xi(\cdot)$ are defined by system (1) and the operator (16), satisfies the condition

$$\int_0^T \|\xi(t)\|^2 dt \leq \int_0^T \|z(t)\|^2 dt \quad Q_\rho^T\text{-a.s.} \tag{24}$$

Then, the definition of the uncertainty input $\xi_\rho(\cdot)$ and condition (24) imply that for each $T > 0$

$$\frac{1}{T} \int_0^T \left[\|z(s)\|^2 - \|\xi_\rho(s)\|^2\right] ds \geq 0 \quad Q_\rho^T\text{-a.s.} \tag{25}$$

From the above condition, it follows that for each $\rho > 0$

$$\inf_{T' > T} \frac{1}{T'} \int_0^{T'} \mathbf{E}^{Q_\rho^{T'}} \left[\|z(s)\|^2 - \|\xi_\rho(s)\|^2\right] ds \geq 0.$$

Note that the expectation on the left-hand side of the above inequality exists by virtue of (23). Obviously in this case, one can find a constant $d > 0$ and a variable $\delta(T)$ which is independent of ρ and such that $\lim_{T\to\infty} \delta(T) = 0$ and

$$\inf_{T' > T} \frac{1}{2T'} \mathbf{E}^{Q_\rho^{T'}} \int_0^{T'} \left[\|z(s)\|^2 - \|\xi_\rho(s)\|^2\right] ds \geq -\frac{d}{2} + \delta(T).$$

This, along with the representation of the relative entropy between the probability measure Q_ρ^T and the reference probability measure P^T given in (21), leads us to the conclusion that for the H^∞ norm-bounded uncertainty under consideration, the corresponding martingale $\zeta_\rho(t)$, $\rho > 0$, satisfies the constraint (8). This completes the proof of the lemma. □

Remark 4. In the special case where the uncertainty is modeled by the operator (16) with L_2-induced norm less then one, and where the uncertainty output $z(\cdot)$ of the closed-loop system is known to be Q^T mean square–integrable on any interval $[0,T]$, the above proof shows that such an uncertainty can be characterized directly in terms of the martingale $\zeta(t)$ and the associated probability measures Q^T. That is, one can choose $\zeta_i(t)=\zeta(t)$ and $Q_i^T=Q^T$ in Definition 1. This will be true, for example, if the chosen controller is a stabilizing controller; see Definition 2. However, in the general case, the connection between the uncertainty output $z(\cdot)$ and the uncertainty input $\xi(\cdot)$ can be of a more complex nature than that described by (16). In this case, the Q^T mean square–integrability of the uncertainty output $z(\cdot)$ is not known a priori. Hence, one cannot guarantee that $h(Q^T\|P^T)<\infty$ for all $T>0$. Also, the expectation

$$\frac{1}{T}\left[\mathbf{E}^{Q^T}\int_0^T \|z(t)\|^2 dt - h(Q^T\|P^T)\right]$$

may not exist for all $T>0$ unless it has already been proved that the controller (2) is a stabilizing controller. In this case, the approximations of the martingale $\zeta(t)$ allow us to avoid the difficulties arising when defining an admissible uncertainty for the uncertain system (1) controlled by a generic linear output-feedback controller.

3. Absolute stability and absolute stabilizability. An important issue in any optimal control problem on an infinite time interval concerns the stabilizing properties of the optimal controller. For example, a critical issue addressed in [15, 16, 18] was to prove the absolutely stabilizing property of the optimal control schemes presented in those papers. In this paper, the systems under consideration are subject to additive noise. Hence, we need a definition of absolute stabilizability which properly accounts for this feature of the systems under consideration.

Definition 2. *A controller of the form* (2) *is said to be an* absolutely stabilizing output-feedback controller *for the stochastic uncertain system* (1), (8) *if the process* $x(\cdot)$ *defined by the closed-loop system corresponding to this controller satisfies the following condition. There exist constants* $c_1>0$, $c_2>0$ *such that for any admissible uncertainty martingale* $\zeta(\cdot)\in\Xi$

$$\limsup_{T\to\infty}\frac{1}{T}\left[\mathbf{E}^{Q^T}\int_0^T \left(\|x(t)\|^2+\|u(t)\|^2\right)dt + h(Q^T\|P^T)\right]\le c_1+c_2 d. \tag{26}$$

The property of absolute stability is defined as a special case of Definition 2 corresponding to $u(\cdot)\equiv 0$. In this case, system (1) becomes a system of the form

$$\begin{aligned} dx(t) &= Ax(t)dt + B_2 dW(t),\\ z(t) &= C_1 x(t). \end{aligned} \tag{27}$$

Definition 3. *The stochastic uncertain system corresponding to the state equations* (27) *with uncertainty satisfying the relative entropy constraint* (8) *is said to be* absolutely stable *if there exist constants* $c_1>0$, $c_2>0$ *such that for any admissible uncertainty martingale* $\zeta(\cdot)\in\Xi$

$$\limsup_{T\to\infty}\frac{1}{T}\left[\mathbf{E}^{Q^T}\int_0^T \|x(t)\|^2 dt + h(Q^T\|P^T)\right]\le c_1+c_2 d. \tag{28}$$

In what follows, the following property of mean square stable systems will be used; see [21]. For the sake of completeness, the proof of the following lemma is given in Appendix B.

LEMMA 3. *Suppose the stochastic nominal system* (27) *is mean square stable; i.e.,*

$$\limsup_{T\to\infty} \frac{1}{T}\mathbf{E}\int_0^T \|x(t)\|^2 dt < \infty. \tag{29}$$

Also, suppose the pair (A, B_2) *is stabilizable. Then, the matrix* A *must be stable.*

4. Infinite-horizon minimax optimal control problem. Associated with the stochastic uncertain system (1), (8), consider a cost functional $J(\cdot)$ of the form

$$J(u(\cdot),\zeta(\cdot)) = \limsup_{T\to\infty} \frac{1}{2T}\mathbf{E}^{Q^T}\int_0^T F(x(t),u(t))dt, \tag{30}$$

defined on solutions $x(\cdot)$ to (1). In (30),

$$F(x,u) := x'Rx + u'Gu,$$

and R and G are positive-definite symmetric matrices, $R \in \mathbf{R}^{n\times n}$, $G \in \mathbf{R}^{m\times m}$. Also, in (30), Q^T is the probability measure corresponding to the martingale $\zeta(\cdot)$; see (5).

In this paper, we are concerned with a minimax optimal control problem associated with system (1), cost functional (30), and uncertainty set Ξ. In this problem, we seek to find a controller $u^*(\cdot)$ of the form (2) which minimizes the worst-case value of the cost functional J in the face of uncertainty $\zeta(\cdot) \in \Xi$ satisfying the constraint (8):

$$\sup_{\zeta(\cdot)\in\Xi} J(u^*(\cdot),\zeta(\cdot)) = \inf_{u(\cdot)\in\mathcal{U}} \sup_{\zeta(\cdot)\in\Xi} J(u(\cdot),\zeta(\cdot)). \tag{31}$$

The derivation of a solution to the above minimax optimal control problem relies on a duality relationship between free energy and relative entropy established in [1]; see Lemma 8 of Appendix A. Associated with system (1), consider the parameter-dependent risk-sensitive cost functional

$$\Im_{\tau,T}(u(\cdot)) := \frac{2\tau}{T}\log\mathbf{E}\left\{\exp\left(\frac{1}{2\tau}\int_0^T F_\tau(x(t),u(t))dt\right)\right\}, \tag{32}$$

where $\tau > 0$ is a given constant and

$$F_\tau(x,u) := x'R_\tau x + 2x'\Upsilon_\tau u + u'G_\tau u, \tag{33}$$

$$R_\tau := R + \tau C_1'C_1,$$

$$G_\tau := G + \tau D_1'D_1,$$

$$\Upsilon_\tau := \tau C_1'D_1. \tag{34}$$

We will apply the duality result of Lemma 8 of Appendix A; also, see [1]. When applied to system (1) and the risk-sensitive cost functional (32) (see Corollary 3.1 and Remark 3.2 of [1]), this result states that for each admissible controller $u(\cdot)$

$$\sup_{Q^T\in\mathcal{P}_T} J_{\tau,T}(u(\cdot),Q^T) = \frac{1}{2}\Im_{\tau,T}(u(\cdot)), \tag{35}$$

where

$$J_{\tau,T}(u(\cdot),Q^T) := \frac{1}{T}\left[\frac{1}{2}\mathbf{E}^{Q^T}\int_0^T F_\tau(x(t),u(t))dt - \tau h(Q^T\|P^T)\right]. \tag{36}$$

The use of the duality result (35) is a key step that enables us to replace the minimax optimal control problem by a risk-sensitive optimal control problem. Hence, we will be interested in constructing an output-feedback controller of the form (2) solving the following stochastic risk-sensitive optimal control problem:[1]

$$\inf_{u(\cdot)\in\mathcal{U}} \lim_{T\to\infty} \Im_{\tau,T}(u(\cdot)). \tag{37}$$

5. A connection between risk-sensitive optimal control and minimax optimal control. In this section, we present results establishing a connection between the risk-sensitive optimal control problem (37) and the minimax optimal control problem (31).

For a given constant $\tau > 0$, let V_τ denote an optimal value of the risk-sensitive control problem (37); i.e.,

$$\begin{aligned} V_\tau &:= \inf_{u(\cdot)\in\mathcal{U}} \lim_{T\to\infty} \Im_{\tau,T}(u(\cdot)) \\ &= \inf_{u(\cdot)} \lim_{T\to\infty} \frac{2\tau}{T}\log\mathbf{E}\left\{\exp\left[\frac{1}{2\tau}\int_0^T F_\tau(x(s),u(s))ds\right]\right\}. \end{aligned}$$

THEOREM 1. *Suppose that for a given $\tau > 0$ the risk-sensitive control problem* (37) *admits an optimal controller $u_\tau(\cdot) \in \mathcal{U}$ of the form* (2) *which guarantees a finite optimal value: $V_\tau < \infty$. Then this controller is an absolutely stabilizing controller for the stochastic uncertain system* (1) *satisfying the relative entropy constraint* (8). *Furthermore,*

$$\sup_{\zeta(\cdot)\in\Xi} J(u_\tau(\cdot),\zeta(\cdot)) \le \frac{1}{2}(V_\tau + \tau d). \tag{38}$$

Proof. It follows from the condition of the theorem that

$$V_\tau = \lim_{T\to\infty} \frac{2\tau}{T}\log\mathbf{E}\left\{\exp\left[\frac{1}{2\tau}\int_0^T F_\tau(x(s),u_\tau(s))ds\right]\right\} < \infty,$$

where $u_\tau(\cdot)$ is the risk-sensitive optimal controller of the form (2) corresponding to the given τ. We wish to prove that this risk-sensitive optimal controller satisfies condition (26) of Definition 2.

[1] A risk-sensitive control problem of the form (37) was considered in [9]. That paper defines the class of admissible infinite-horizon risk-sensitive controllers as those controllers which satisfy a certain causality condition. This causality condition is formulated in terms of corresponding martingales and ensures that the probability measure transformations required in [9] are well defined. As observed in [9], linear controllers satisfy this causality condition. Furthermore, it is shown in [9] that a solution to the risk-sensitive optimal control problem (37), in the broader class of nonlinear output-feedback controllers satisfying such a causality condition, is attained by a linear controller of the form (2). This implies that the class of admissible controllers in the risk-sensitive control problem (37) can be restricted to include only linear output-feedback controllers.

Using the duality result (35), we obtain

$$\lim_{T\to\infty}\sup_{Q^T\in\mathcal{P}_T}\frac{1}{T}\left[\frac{1}{2}\mathbf{E}^{Q^T}\int_0^T F_\tau(x(s),u_\tau(s))ds-\tau h(Q^T\|P^T)\right]=\frac{V_\tau}{2}. \tag{39}$$

Equation (39) implies that, for any sufficiently large $T>0$, one can choose a sufficiently small positive constant $\hat{\delta}=\hat{\delta}(T)>0$ such that $\lim_{T\to\infty}\hat{\delta}(T)=0$ and

$$\sup_{Q^{T'}\in\mathcal{P}_{T'}}\frac{1}{T'}\left[\frac{1}{2}\mathbf{E}^{Q^{T'}}\int_0^{T'} F_\tau(x(s),u_\tau(s))ds-\tau h(Q^{T'}\|P^{T'})\right]\le\frac{V_\tau}{2}+\hat{\delta}(T) \tag{40}$$

for all $T'>T$. Thus, for the chosen constants $T>0$ and $\hat{\delta}(T)>0$ and for all $T'>T$,

$$\frac{1}{T'}\left[\frac{1}{2}\mathbf{E}^{Q^{T'}}\int_0^{T'} F_\tau(x(s),u_\tau(s))ds-\tau h(Q^{T'}\|P^{T'})\right]\le\frac{V_\tau}{2}+\hat{\delta}(T)$$

for any $Q^{T'}\in\mathcal{P}_{T'}$. Furthermore, if $Q^{T'}\in\mathcal{P}_{T'}$ for all $T'>T$, then

$$\sup_{T'>T}\frac{1}{T'}\left[\frac{1}{2}\mathbf{E}^{Q^{T'}}\int_0^{T'} F_\tau(x(s),u_\tau(s))ds-\tau h(Q^{T'}\|P^{T'})\right]\le\frac{V_\tau}{2}+\hat{\delta}(T). \tag{41}$$

Let $\zeta(\cdot)\in\Xi$ be a given admissible uncertainty martingale and let $\zeta_i(\cdot)$ be a corresponding sequence of martingales as in Definition 1. Recall that the corresponding probability measures Q_i^T belong to the set $\mathcal{P}_T$ for all $T>0$. Hence each probability measure Q_i^T satisfies condition (41); i.e.,

$$\sup_{T'>T}\frac{1}{T'}\left[\frac{1}{2}\mathbf{E}^{Q_i^{T'}}\int_0^{T'} F_\tau(x(s),u_\tau(s))ds-\tau h(Q_i^{T'}\|P^{T'})\right]\le\frac{V_\tau}{2}+\hat{\delta}(T). \tag{42}$$

Note that in condition (42), $\hat{\delta}(T)$ and T are the constants which are independent of i.

Since $F(x,u)\ge 0$ and $\tau>0$, condition (42) implies

$$\sup_{T'>T}\frac{1}{T'}\left[\frac{1}{2}\mathbf{E}^{Q_i^{T'}}\int_0^{T'}\|z(s)\|^2ds-h(Q_i^{T'}\|P^{T'})\right]<\infty.$$

From this, it follows from (42) that for each integer i

$$\begin{aligned}
&\sup_{T'>T}\frac{1}{2T'}\mathbf{E}^{Q_i^{T'}}\int_0^{T'}F(x(s),u_\tau(s))ds\\
&\quad+\tau\inf_{T'>T}\frac{1}{T'}\left[\frac{1}{2}\mathbf{E}^{Q_i^{T'}}\int_0^{T'}\|z(s)\|^2ds-h(Q_i^{T'}\|P^{T'})\right]\\
&\le\sup_{T'>T}\frac{1}{T'}\left[\frac{1}{2}\mathbf{E}^{Q_i^{T'}}\int_0^{T'}\left(F(x(s),u_\tau(s))+\tau\|z(s)\|^2\right)ds-\tau h(Q_i^{T'}\|P^{T'})\right]\\
&\le\frac{1}{2}V_\tau+\hat{\delta}(T).
\end{aligned}$$

This implies

$$\begin{aligned}
&\sup_{T'>T} \frac{1}{2T'} \mathbf{E}^{Q_i^{T'}} \int_0^{T'} F(x(s), u_\tau(s)) ds \\
&\leq \frac{V_\tau}{2} + \hat{\delta}(T) - \tau \inf_{T'>T} \frac{1}{T'} \left[\frac{1}{2} \mathbf{E}^{Q_i^{T'}} \int_0^{T'} \|z(s)\|^2 ds - h(Q_i^{T'} \| P^{T'}) \right] \\
&\leq \frac{1}{2}(V_\tau + \tau d) + \hat{\delta}(T) - \tau \delta(T). \qquad (43)
\end{aligned}$$

The derivation of the last line of inequality (43) uses the fact that the probability measure Q_i^T satisfies condition (8). Also, note that in condition (43), the constants $\hat{\delta}(T)$, $\delta(T)$, and T are independent of i and $T' > T$.

We now let $i \to \infty$ in inequality (43). This leads to the following proposition.

PROPOSITION 2. *For any admissible uncertainty* $\zeta(\cdot) \in \Xi$,

$$\sup_{T'>T} \frac{1}{2T'} \mathbf{E}^{Q^{T'}} \int_0^{T'} F(x(s), u_\tau(s)) ds \leq \frac{1}{2}(V_\tau + \tau d) + \hat{\delta}(T) - \tau \delta(T). \qquad (44)$$

To establish this proposition, consider the space $L_1(\Omega, \mathcal{F}_{T'}, P^{T'})$ endowed with the topology of weak convergence of random variables, where $T' > T$. We denote this space by L_1^w. Define the functional

$$\phi(\nu) := \frac{1}{T'} \mathbf{E} \left[\nu \int_0^{T'} F(x(s), u_\tau(s)) ds \right], \qquad (45)$$

mapping L_1^w into the space of extended reals $\mathfrak{R} = \mathbf{R} \cup \{-\infty, \infty\}$. Also, consider a sequence of functionals mapping $L_1^w \to \mathfrak{R}$ defined by

$$\phi_N(\nu) := \frac{1}{T'} \mathbf{E} \left[\nu \int_0^{T'} F_N(x(s), u_\tau(s)) ds \right], \quad N = 1, 2, \ldots, \qquad (46)$$

where each function $F_N(\cdot)$ is defined as follows:

$$F_N(x, u) := \begin{cases} F(x, u) & \text{if } F(x, u) \leq N, \\ N & \text{if } F(x, u) > N. \end{cases}$$

Note that from the above definition, the sequence $\phi_N(\nu)$ is monotonically increasing in N for each ν. Also, we note that for any $N > 0$

$$P\left(\frac{1}{T'} \int_0^{T'} F_N(x(s), u_\tau(s)) ds \leq N \right) = 1.$$

Hence, if $\nu_i \to \nu$ weakly, then $\phi_N(\nu_i) \to \phi_N(\nu)$. That is, each functional $\phi_N(\cdot)$ is continuous on the space L_1^w. Therefore, the functional

$$\phi(\nu) = \lim_{N \to \infty} \phi_N(\nu)$$

is lower semicontinuous; e.g., see Theorem 10 on page 330 of [13]. Now let $\nu = \zeta(T')$ be the Radon–Nikodým derivative of the probability measure $Q^{T'}$, and let $\nu_i = \zeta_i(T')$

be the Radon–Nikodým derivative of the probability measure $Q_i^{T'}$. Then, the fact that $\zeta_i(T') \to \zeta(T')$ weakly implies

$$\frac{1}{2T'}\mathbf{E}^{Q^{T'}}\int_0^{T'} F(x(s),u_\tau(s))ds \le \liminf_{i\to\infty}\frac{1}{2T'}\mathbf{E}^{Q_i^{T'}}\int_0^{T'} F(x(s),u_\tau(s))ds$$
$$\le \frac{1}{2}(V_\tau+\tau d)+\hat{\delta}(T)-\tau\delta(T). \tag{47}$$

Since the constants on the right-hand side of (47) are independent of $T' > T$, condition (44) of the proposition now follows. This completes the proof of the proposition.

Note that from the above proposition, (38) follows. Indeed, for any $\zeta(\cdot) \in \Xi$, Proposition 2 and the fact that $\hat{\delta}(T), \delta(T) \to 0$ as $T \to \infty$ together imply

$$J(u_\tau(\cdot),\zeta(\cdot)) = \lim_{T\to\infty}\sup_{T'>T}\frac{1}{2T'}\mathbf{E}^{Q^{T'}}\int_0^{T'} F(x(s),u_\tau(s))ds$$
$$\le \frac{1}{2}(V_\tau+\tau d). \tag{48}$$

From condition (48), equation (38) of the theorem follows.

We now establish the absolute stabilizing property of the risk-sensitive optimal controller $u_\tau(\cdot)$. Indeed, since the matrices R and G are positive-definite, inequality (44) implies

$$\limsup_{T\to\infty}\frac{1}{T}\mathbf{E}^{Q^T}\int_0^T \left(\|x(s)\|^2+\|u_\tau(s)\|^2\right)ds \le \alpha\left(V_\tau+\tau d\right), \tag{49}$$

where α is a positive constant which depends only on R and G.

To complete the proof, it remains to prove that there exist constants $c_1, c_2 > 0$ such that

$$\limsup_{T\to\infty}\frac{1}{T}h(Q^T\|P^T) < c_1 + c_2 d. \tag{50}$$

To this end, we note that for any sufficiently large T and for all $T' > T$, the constraint (8) implies

$$\frac{1}{T'}h(Q_i^{T'}\|P^{T'}) \le \frac{1}{2T'}\mathbf{E}^{Q_i^{T'}}\int_0^{T'}\|z(s)\|^2 ds + \frac{d}{2} - \delta(T) \tag{51}$$

for all $i = 1, 2, \ldots$. We now observe that condition (43) implies that for all $T' > T$

$$\frac{1}{2T'}\mathbf{E}^{Q_i^{T'}}\int_0^{T'}\|z(s)\|^2 ds \le \bar{c}\left(\frac{1}{2}(V_\tau+\tau d)+\hat{\delta}(T)-\tau\delta(T)\right), \tag{52}$$

where $\bar{c}$ is a positive constant determined only by the matrices R, G, C_1, and D_1. From conditions (51), (52), Remark 1, and the fact that the relative entropy functional is lower semicontinuous, it follows that

$$\frac{1}{T'}h(Q^{T'}\|P^{T'}) \le \liminf_{i\to\infty}\frac{1}{T'}h(Q_i^{T'}\|P^{T'})$$
$$\le \bar{c}\left(\frac{1}{2}(V_\tau+\tau d)+\hat{\delta}(T)-\tau\delta(T)\right)+\frac{d}{2}-\delta(T)$$

for any $T' > T$. This inequality and the fact that $\delta(T) \to 0$ and $\hat{\delta}(T) \to 0$ as $T \to \infty$ together imply that

$$\limsup_{T\to\infty} \frac{1}{T} h(Q^T \| P^T) \leq \frac{1}{2}(\bar{c}V_\tau + (1+\bar{c}\tau)d).$$

Combining this condition and inequality (49), we obtain condition (26), where the constants c_1, c_2 are defined by V_τ, τ, α, $\bar{c}$, and hence independent of $\zeta(\cdot) \in \Xi$. □

Remark 5. It is straightforward to extend the result of Theorem 1 to the case in which the uncertainty output is structured; i.e.,

$$\begin{aligned} z_1(t) &= C_{1,1}x(t) + D_{1,1}u(t), \\ &\vdots \\ z_k(t) &= C_{1,k}x(t) + D_{1,k}u(t). \end{aligned}$$

In this case, we need k relative entropy uncertainty constraints of the form (8) to define the admissible uncertainty. The corresponding risk-sensitive control problem involves k scaling parameters $\tau_1 \geq 0, \ldots, \tau_k \geq 0$, $\sum_{j=1}^{k} \tau_j > 0$.

To formulate conditions under which a converse to Theorem 1 holds, we consider the closed-loop system corresponding to system (1) and a linear time-invariant output-feedback controller of the form (2). Recall that the closed-loop nominal system corresponding to controller (2) is described by the linear Ito differential equation (3) on the probability space $(\Omega, \mathcal{F}, P)$. In what follows, we will consider the class of linear controllers of the form (2) satisfying the following assumptions: the matrix $\bar{A}$ is stable, the pair $(\bar{A}, \bar{B})$ is controllable, and the pair $(\bar{A}, \bar{R})$ is observable, where

$$\bar{R} = \begin{bmatrix} R & 0 \\ 0 & K'GK \end{bmatrix}. \tag{53}$$

Also, let $\mathfrak{D}_0$ be the set of all linear functions $\phi(\bar{x}) = \Phi\bar{x}$ such that the matrix $\bar{A} + \bar{B}\Phi$ is stable. Note that the pair $(\bar{A} + \bar{B}\Phi, \bar{B})$ is controllable since the pair $(\bar{A}, \bar{B})$ is controllable. Under these assumptions, the Markov process generated by the linear system

$$d\bar{x}_\phi(t) = (\bar{A} + \bar{B}\Phi)\bar{x}_\phi(t)dt + \bar{B}dW(t) \tag{54}$$

has a unique invariant probability measure ν^ϕ on $\mathbf{R}^{n+\hat{n}}$; e.g., see [24]. It is shown in [24] that the probability measure ν^ϕ is a Gaussian probability measure.

LEMMA 4. *For every function $\phi(\bar{x}) = \Phi\bar{x}$, $\phi(\cdot) \in \mathfrak{D}_0$, there exists a martingale $\zeta(\cdot) \in \mathcal{M}_\infty$ such that for any $T > 0$ the process*

$$\tilde{W}(t) = W(t) - \int_0^t \Phi\bar{x}(s)ds \tag{55}$$

is a Wiener process with respect to $\{\mathcal{F}_t, t \in [0,T]\}$ and the probability measure Q^T corresponding to the martingale $\zeta(\cdot)$. In (55), $\bar{x}(\cdot)$ is the solution to the nominal closed-loop system (3) with initial probability distribution ν^ϕ.

Furthermore,

$$d\bar{x} = (\bar{A} + \bar{B}\Phi)\bar{x}dt + \bar{B}d\tilde{W}(t), \tag{56}$$

considered on the probability space $(\Omega, \mathcal{F}_T, Q^T)$*, admits a stationary solution* $\bar{x}_\zeta(\cdot)$ *such that*

$$Q^T(\bar{x}_\zeta(t) \in d\bar{x}) = \nu^\phi(d\bar{x}). \tag{57}$$

Proof. Let ν^ϕ be the Gaussian invariant probability measure corresponding to a given $\phi(\cdot) \in \mathfrak{D}_0$. Consider a stochastic process $\bar{x}(t)$ defined by (3) and having initial probability distribution ν^ϕ; i.e., $P(\bar{x}(0) \in d\bar{x}) = \nu^\phi(d\bar{x})$. Since the probability measure ν^ϕ is Gaussian, there exists a constant $\delta_0 > 0$ such that

$$\mathbf{E} \exp(\delta_0 \|\bar{x}(0)\|^2) = \int \exp(\delta_0 \|\bar{x}\|^2) \nu^\phi(d\bar{x}) < \infty.$$

Hence, using the multivariate version of Theorem 4.7 on page 137 of [6] along with Example 3 on page 220 of [6], this leads to the satisfaction of the conditions of Lemma 1, which shows that the random process $\tilde{W}(\cdot)$ defined by (55) is a Wiener process with respect to $\{\mathcal{F}_t, t \in [0,T]\}$ and the probability measure Q^T defined in Lemma 1.

We now consider system (56) on the probability space $(\Omega, \mathcal{F}_T, Q^T)$ with initial distribution ν^ϕ. Also, consider system (54) on the probability space $(\Omega, \mathcal{F}_T, P^T)$ with initial distribution ν^ϕ. It follows from Proposition 3.10 on page 304 of [4] that the stochastic process $\bar{x}_\zeta(\cdot)$ defined by (56) and the corresponding stochastic process $\bar{x}_\phi(\cdot)$ defined by (54) have the same probability distribution under their respective probability measures. Also, as in [2, 1],

$$h(Q^T \| P^T) = \frac{1}{2} \mathbf{E}^{Q^T} \int_0^T \|\Phi \bar{x}(t)\|^2 dt = \frac{1}{2} \int \|\Phi \bar{x}\|^2 \nu^\phi(d\bar{x}) < \infty$$

for each $T < \infty$, since $\bar{x}(t)$ is the solution to system (3) with Gaussian initial distribution ν^ϕ. Thus, $Q^T \in \mathcal{P}_T$ for all $T > 0$. Hence, $\zeta(\cdot) \in \mathcal{M}_\infty$. From this, the lemma follows. □

We now present a converse to Theorem 1.

THEOREM 2. *Suppose that there exists a controller* $u^*(\cdot) \in \mathcal{U}$ *such that the following conditions are satisfied:*

(i) $\sup_{\zeta(\cdot) \in \Xi} J(u^*(\cdot), \zeta(\cdot)) < c < \infty$.

(ii) *The controller* $u^*(\cdot)$ *is an absolutely stabilizing controller such that the corresponding closed-loop matrix* $\bar{A}$ *is stable, the pair* $(\bar{A}, \bar{B})$ *is controllable, and the pair* $(\bar{A}, \bar{R})$ *is observable.*

Then there exists a constant $\tau > 0$ *such that the corresponding risk-sensitive optimal control problem* (37) *has a solution which guarantees a finite optimal value. Furthermore,*

$$\frac{1}{2}(V_\tau + \tau d) < c. \tag{58}$$

The proof of this theorem follows along the same lines as the proof of the necessity part of the main result of [21]. For the sake of completeness, the modification of this proof adapted to the condition of Theorem 2 is presented below.

We first establish the following lemma.

LEMMA 5. *Consider the uncertain closed-loop system* (3), (8) *in which the pair* $(\bar{A}, \bar{B})$ *is controllable. Also, consider a nonnegative-definite matrix* $\bar{R}$ *such that the pair* $(\bar{A}, \bar{R})$ *is observable. If the system* (3), (8) *is absolutely stable, then there exists a positive constant* $\tau_0 > 0$ *such that the Riccati equation*

$$\bar{A}'\Pi + \Pi\bar{A} + \bar{R} + \tau_0\bar{C}'\bar{C} + \frac{1}{\tau_0}\Pi\bar{B}\bar{B}'\Pi = 0 \tag{59}$$

admits a positive-definite stabilizing solution.

Proof. Since the uncertain system (3), (8) is absolutely stable, there exists a positive constant $\tilde{c}$ such that for all $\zeta(\cdot) \in \Xi$

$$\liminf_{T\to\infty} \frac{1}{2T}\mathbf{E}^{Q^T}\int_0^T \bar{x}(s)'\bar{R}\bar{x}(s)ds + \bar{\varepsilon}\liminf_{T\to\infty} \frac{1}{2T}\mathbf{E}^{Q^T}\int_0^T \|\bar{x}(t)\|^2 dt \le \tilde{c}. \tag{60}$$

Here $\bar{\varepsilon} > 0$ is a sufficiently small positive constant.

Consider the functionals

$$\begin{aligned}\mathcal{G}_0(\zeta(\cdot)) &:= \tilde{c} - \liminf_{T\to\infty} \frac{1}{2T}\mathbf{E}^{Q^T}\int_0^T \bar{x}(s)'\bar{R}\bar{x}(s)ds \\ &\quad - \bar{\varepsilon}\liminf_{T\to\infty} \frac{1}{2T}\mathbf{E}^{Q^T}\int_0^T \|\bar{x}(t)\|^2 dt, \\ \mathcal{G}_1(\zeta(\cdot)) &:= -\frac{d}{2} - \liminf_{T\to\infty} \frac{1}{T}\left[\frac{1}{2}\mathbf{E}^{Q^T}\int_0^T \|z(s)\|^2 ds - h(Q^T\|P^T)\right].\end{aligned} \tag{61}$$

Note that since the system (3), (8) is absolutely stable, both of these functionals are well defined on the set Ξ.

Now consider a martingale $\zeta(\cdot) \in \mathcal{M}_\infty$ such that

$$\mathcal{G}_1(\zeta(\cdot)) \le 0. \tag{62}$$

This condition implies that the martingale $\zeta(\cdot)$ satisfies the conditions of Definition 1 with $\zeta_i(\cdot) = \zeta(\cdot)$. Indeed, condition (i) of Definition 1 is satisfied since $\zeta(\cdot) \in \mathcal{M}_\infty$. Condition (ii) is trivial in this case. Also, let $\delta(T)$ be any function chosen to satisfy the conditions $\lim_{T\to\infty}\delta(T) = 0$ and

$$\inf_{T'>T} \frac{1}{T'}\left[\frac{1}{2}\mathbf{E}^{Q^{T'}}\int_0^{T'} \|z(s)\|^2 ds - h(Q^{T'}\|P^{T'})\right] \ge -\frac{d}{2} + \delta(T)$$

for all sufficiently large $T > 0$. The existence of such a function $\delta(T)$ follows from condition (62). Then condition (8) of Definition 1 is also satisfied. Thus, condition (62) implies that each martingale $\zeta(\cdot) \in \mathcal{M}_\infty$ satisfying this condition is an admissible uncertainty martingale. That is, $\zeta(\cdot) \in \Xi$. From condition (60), it follows that $\mathcal{G}_0(\zeta(\cdot)) \ge 0$.

We have now shown that the satisfaction of condition (60) implies that the following condition is satisfied:

$$\text{If } \mathcal{G}_1(\zeta(\cdot)) \le 0, \text{ then } \mathcal{G}_0(\zeta(\cdot)) \ge 0. \tag{63}$$

Furthermore, the set of martingales satisfying condition (62) has an interior point $\zeta(t) \equiv 1$; see the remark following Definition 1. Also, it follows from the properties

of the relative entropy functional that the functionals $\mathcal{G}_0(\cdot)$ and $\mathcal{G}_1(\cdot)$ are convex. We have now verified all of the conditions needed to apply the Lagrange multiplier result (e.g., see [7]). Indeed, Theorem 1 on page 217 of [7] implies that there exists a constant $\tau_0 \geq 0$ such that

$$\mathcal{G}_0(\zeta(\cdot)) + \tau_0 \mathcal{G}_1(\zeta(\cdot)) \geq 0 \tag{64}$$

for all $\zeta(\cdot) \in \mathcal{M}_\infty$. We now show that the conditions of the theorem guarantee that $\tau_0 > 0$.

PROPOSITION 3. *In inequality* (64), $\tau_0 > 0$.

Consider system (56) where $\phi(\bar{x}) := \Phi\bar{x}$ belongs to $\mathfrak{D}_0$. From Lemma 4, the corresponding martingale $\zeta_\Phi(\cdot)$ belongs to the set $\mathcal{M}_\infty$.

Now consider the quantity

$$\liminf_{T\to\infty} \frac{1}{2T} \mathbf{E}^{Q_\Phi^T} \int_0^T \bar{x}(t)' \bar{R} \bar{x}(t) dt.$$

Here, Q_Φ^T is the probability measure corresponding to the martingale $\zeta_\Phi(\cdot)$, and $\bar{x}(\cdot)$ is the solution to the corresponding system (1) considered on the probability space $(\Omega, \mathcal{F}_T, Q_\Phi^T)$. Also, consider the Lyapunov equation

$$(\bar{A} + \bar{B}\Phi)'\Pi + \Pi(\bar{A} + \bar{B}\Phi) + \bar{R} = 0. \tag{65}$$

Since the matrix $\bar{A} + \bar{B}\Phi$ is stable, then this matrix equation admits a nonnegative-definite solution Π. Using Ito's formula, it is straightforward to show that (65) leads to the inequality

$$\liminf_{T\to\infty} \frac{1}{T} \mathbf{E}^{Q_\Phi^T} \int_0^T \bar{x}(t)' \bar{R} \bar{x}(t) dt \geq \operatorname{tr} \bar{B}\bar{B}'\Pi.$$

This condition implies that

$$\sup_{\zeta(\cdot)\in\mathcal{M}_\infty} \liminf_{T\to\infty} \frac{1}{2T} \mathbf{E}^{Q^T} \int_0^T \bar{x}(t)' \bar{R} \bar{x}(t) dt \geq \sup_{\Phi:\bar{A}+\bar{B}\Phi \text{ is stable}} \frac{1}{2} \operatorname{tr} \bar{B}\bar{B}'\Pi = \infty. \tag{66}$$

Using (66), the proposition follows. Indeed, suppose that $\tau_0 = 0$. Then condition (64) implies that

$$\sup_{\zeta(\cdot)\in\mathcal{M}_\infty} \left[\begin{array}{l} \liminf_{T\to\infty} \frac{1}{2T} \mathbf{E}^{Q^T} \int_0^T \bar{x}(t)' \bar{R} \bar{x}(t) dt \\ +\bar{\varepsilon} \liminf_{T\to\infty} \frac{1}{2T} \mathbf{E}^{Q^T} \int_0^T \|\bar{x}(t)\|^2 dt \end{array} \right] \leq \tilde{c} < \infty. \tag{67}$$

Inequality (67) leads to a contradiction with condition (66). From this, it follows that $\tau_0 > 0$.

PROPOSITION 4. *The Riccati equation* (59) *with* τ_0 *defined above admits a positive-definite stabilizing solution.*

We first note that the pair $(\bar{A}, \bar{R} + \tau_0 \bar{C}'\bar{C})$ is observable, since the pair $(\bar{A}, \bar{R})$ is observable. Hence, if $\Pi \geq 0$ satisfies (59), then $\Pi > 0$. Thus, it is sufficient to prove that (59) admits a nonnegative-definite stabilizing solution. This is true if and only if the following bound on the H^∞ norm of the corresponding transfer function is satisfied:

$$\|\mathcal{H}_{\tau_0}(s)\|_\infty \leq 1, \tag{68}$$

where

$$\mathcal{H}_{\tau_0}(s) := \begin{bmatrix} \frac{1}{\sqrt{\tau_0}}\bar{R}^{1/2} \\ \bar{C} \\ \frac{\sqrt{\bar{\varepsilon}}}{\sqrt{\tau_0}} I \end{bmatrix} (sI - \bar{A})^{-1}\bar{B};$$

see Lemma 5 and Theorem 5 of [22].

In order to prove the above claim, we note that condition (64) implies that for any martingale $\zeta(\cdot) \in \mathcal{M}_\infty$

$$\begin{aligned} \liminf_{T\to\infty} \frac{1}{T}\mathbf{E}^{Q^T} \int_0^T \frac{1}{2\tau_0}\bar{x}(t)'\bar{R}\bar{x}(t)dt + \frac{\bar{\varepsilon}}{\tau_0} \liminf_{T\to\infty} \frac{1}{2T}\mathbf{E}^{Q^T} \int_0^T \|\bar{x}(t)\|^2 dt \\ + \liminf_{T\to\infty} \frac{1}{T}\left[\frac{1}{2}\mathbf{E}^{Q^T}\int_0^T \|z(s)\|^2 ds - h(Q^T\|P^T)\right] \\ \leq \frac{\tilde{c}}{\tau_0} - \frac{d}{2}. \end{aligned} \tag{69}$$

We will show that the satisfaction of condition (68) follows from (69).

Suppose condition (68) is not true. That is, suppose that

$$\|\mathcal{H}_{\tau_0}(s)\|_\infty > 1. \tag{70}$$

Consider a set $\mathfrak{P}^+$ of deterministic power signals $\xi(t)$, $t \in (-\infty, \infty)$, for which the autocorrelation matrix exists and is finite and for which the power spectral density function exists. Furthermore, $\xi(t) = 0$ if $t < 0$. It can be shown that $\|\mathcal{H}_{\tau_0}\|_{\mathfrak{P}^+} = \|\mathcal{H}_{\tau_0}\|_\infty$, where $\|\mathcal{H}_{\tau_0}\|_{\mathfrak{P}^+}$ denotes the induced norm of the convolution operator $\mathfrak{P}^+ \to \mathfrak{P}^+$ defined by the transfer function $\mathcal{H}_{\tau_0}(s)$. The proof of this fact is a minor variation of the proof of the corresponding fact given in [25].

Now consider the following state space realization of the transfer function $\mathcal{H}_{\tau_0}(s)$:

$$\begin{aligned} \frac{d\bar{x}_1}{dt} &= \bar{A}\bar{x}_1 + \bar{B}\xi(t), \\ z_1 &= \begin{bmatrix} \frac{1}{\sqrt{\tau_0}}\bar{R}^{1/2} \\ \bar{C} \\ \frac{\sqrt{\bar{\varepsilon}}}{\sqrt{\tau_0}} I \end{bmatrix} \bar{x}_1. \end{aligned} \tag{71}$$

Then, the fact that $\|\mathcal{H}_{\tau_0}\|_{\mathfrak{P}^+} = \|\mathcal{H}_{\tau_0}\|_\infty > 1$ leads to the following conclusion:

$$\sup_{\xi(\cdot)\in\mathfrak{P}^+} \lim_{T\to\infty} \frac{1}{T}\int_0^T \left(\|z_1(t)\|^2 dt - \|\xi(t)\|^2\right) dt = \infty. \tag{72}$$

In (72), $z_1(\cdot)$ is the output of system (71) corresponding to the input $\xi(\cdot) \in \mathfrak{P}^+$ and an arbitrarily chosen initial condition $\bar{x}_1(0)$.[2] That is, for any $N > 0$ there exists an uncertainty input $\xi_N(\cdot) \in \mathfrak{P}^+$ such that

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T \left(\|z_1(t)\|^2 dt - \|\xi_N(t)\|^2\right) dt > N.$$

[2]Note that the limit on the left-hand side of (72) is independent of the initial condition of system (71).

This condition implies that for a sufficiently small $\varepsilon > 0$ there exists a constant $T(\varepsilon, N) > 0$ such that

$$\frac{1}{T}\int_0^T \left(\|z_1(t)\|^2 dt - \|\xi_N(t)\|^2\right) dt > N - \varepsilon \tag{73}$$

for all $T > T(\varepsilon, N)$. We now suppose that the initial condition of system (71) is a random variable $\bar{x}_0$. This system is driven by the input $\xi_N(\cdot)$. In this case, system (71) gives rise to an $\mathcal{F}_0$-measurable stochastic process $\bar{x}_1(\cdot)$. Furthermore, for all $T > T(\varepsilon, N)$, inequality (73) holds with probability one. Now note that the signal $\xi_N(\cdot)$ is a deterministic signal; hence, it satisfies the conditions of Lemma 1. Therefore, for this process, the martingale $\zeta_N(\cdot) \in \mathcal{M}$, the probability measure Q_N^T, and the Wiener process $\tilde{W}(\cdot)$ can be constructed as described in Lemma 1. Also, since $\xi_N(\cdot) \in \mathfrak{P}^+$ and is deterministic, then for any $T > 0$

$$\mathbf{E}^{Q_N^T}\int_0^T \|\xi_N(t)\|^2 dt = \int_0^T \|\xi_N(t)\|^2 dt < \infty.$$

From this observation, it follows that $\zeta_N(\cdot) \in \mathcal{M}_\infty$, and also the random variable on the left-hand side of inequality (73) has the finite expectation with respect to the probability measure Q_N^T. Furthermore, using inequality (73), one can prove that the system

$$d\bar{x} = (\bar{A}\bar{x} + \bar{B}\xi_N(t))dt + \bar{B}d\tilde{W}(t), \quad \bar{x}(0) = \bar{x}_0, \tag{74}$$

considered on the probability space $(\Omega, \mathcal{F}_T, Q_N^T)$, satisfies the following condition:

$$\frac{1}{T}\mathbf{E}^{Q_N^T}\int_0^T \left(\bar{x}'(t)\left(\frac{1}{\tau_0}\bar{R} + \bar{C}'\bar{C}\right)\bar{x}(t) + \frac{\bar{\varepsilon}}{\tau_0}\|\bar{x}(t)\|^2 - \|\xi_N(t)\|^2\right) dt > N - \varepsilon.$$

This condition can be established using the same arguments as those used in proving the corresponding fact in [19]. Hence,

$$\lim_{T\to\infty}\frac{1}{T}\mathbf{E}^{Q_N^T}\int_0^T \left(\bar{x}'(t)\left(\frac{1}{\tau_0}\bar{R} + \bar{C}'\bar{C}\right)\bar{x}(t) + \frac{\bar{\varepsilon}}{\tau_0}\|\bar{x}(t)\|^2 - \|\xi_N(t)\|^2\right) dt \geq N. \tag{75}$$

Letting $N \to \infty$ in (75) and using the representation of the relative entropy $h(Q_N^T\|P^T)$, we obtain a contradiction with (69):

$$\begin{aligned}
&\sup_{\zeta\in\mathcal{M}_\infty}\left\{\begin{array}{l}
\liminf_{T\to\infty}\frac{1}{T}\mathbf{E}^{Q^T}\int_0^T \frac{1}{2\tau_0}\bar{x}(t)'\bar{R}\bar{x}(t)dt \\
+\frac{\bar{\varepsilon}}{\tau_0}\liminf_{T\to\infty}\frac{1}{2T}\mathbf{E}^{Q^T}\int_0^T \|\bar{x}(s)\|^2 ds \\
+\liminf_{T\to\infty}\frac{1}{T}\left[\frac{1}{2}\mathbf{E}^{Q^T}\int_0^T \|z(s)\|^2 ds - h(Q^T\|P^T)\right]
\end{array}\right\} \\
&\geq \sup_{N>0}\left\{\begin{array}{l}
\lim_{T\to\infty}\frac{1}{T}\mathbf{E}^{Q_N^T}\int_0^T \frac{1}{2\tau_0}\bar{x}(t)'\bar{R}\bar{x}(t)dt \\
+\frac{\bar{\varepsilon}}{\tau_0}\lim_{T\to\infty}\frac{1}{2T}\mathbf{E}^{Q_N^T}\int_0^T \|\bar{x}(s)\|^2 ds \\
+\lim_{T\to\infty}\frac{1}{2T}\mathbf{E}^{Q_N^T}\int_0^T (\|z(s)\|^2 - \|\xi_N(s)\|^2) ds
\end{array}\right\} \\
&= \infty.
\end{aligned}$$

Thus, condition (68) holds. As observed above, the proposition follows from this condition. Consequently, Lemma 5 follows from Proposition 4. □

Proof of Theorem 2. This proof exploits a large deviation result established in [14].

We first note that since the given controller $u^*(\cdot)$ is an absolutely stabilizing controller and the pair $(\bar{A}, \bar{R})$ is observable, the uncertain closed-loop system (3), (8) is absolutely stable. Furthermore, condition (i) of the theorem implies that there exists a sufficiently small positive constant $\bar{\varepsilon} > 0$ such that for all $\zeta(\cdot) \in \Xi$

$$\liminf_{T\to\infty} \frac{1}{2T}\mathbf{E}^{Q^T} \int_0^T \bar{x}(s)'\bar{R}\bar{x}(s)ds + \bar{\varepsilon} \liminf_{T\to\infty} \frac{1}{2T}\mathbf{E}^{Q^T} \int_0^T \|\bar{x}(t)\|^2 dt \le c - \bar{\varepsilon}. \tag{76}$$

Here $\bar{R}$ is the matrix corresponding to the controller $u^*(\cdot)$ as defined in (53). Also, $c > 0$ is the constant defined in condition (i) of the theorem. Then, it follows from Lemma 5 that there exists a positive constant $\tau_0 > 0$ such that the Riccati equation (59) has a positive-definite stabilizing solution. The existence of such a constant τ_0 is established using condition (76) in the same manner as in the proof of Lemma 5. Also, as in the proof of Lemma 5, it follows that for any martingale $\zeta(\cdot) \in \mathcal{M}_\infty$,

$$\begin{aligned} \liminf_{T\to\infty} \frac{1}{T}\mathbf{E}^{Q^T} \int_0^T \frac{1}{2}\bar{x}(t)'\bar{R}\bar{x}(t)dt + \bar{\varepsilon} \liminf_{T\to\infty} \frac{1}{2T}\mathbf{E}^{Q^T} \int_0^T \|\bar{x}(t)\|^2 dt \\ + \liminf_{T\to\infty} \frac{1}{T}\left[\frac{1}{2}\mathbf{E}^{Q^T} \int_0^T \|z(s)\|^2 ds - h(Q^T\|P^T)\right] \\ \le c - \bar{\varepsilon} - \frac{d}{2}\tau_0. \end{aligned} \tag{77}$$

Furthermore, the matrix $\bar{A}$ is stable, the pair $(\bar{A}, \bar{B})$ is controllable, and the pair $(\bar{A}, \bar{R} + \tau_0\bar{C}'\bar{C})$ is observable. The above conditions and the condition that Riccati equation (59) has a positive-definite stabilizing solution are the conditions of Example 2.2 of [14]. It follows from this example that

$$\begin{aligned} \lim_{T\to\infty} \frac{\tau_0}{T} \log \mathbf{E} \exp\left\{\frac{1}{2\tau_0}\int_0^T \bar{x}'(t)(\bar{R} + \tau_0\bar{C}'\bar{C})\bar{x}(t)dt\right\} \\ = \frac{1}{2}\int \left[\bar{x}'(\bar{R} + \tau_0\bar{C}'\bar{C})\bar{x} - \tau_0\|\phi(x)\|^2\right] \nu^\phi(d\bar{x}), \end{aligned} \tag{78}$$

where $\phi(\bar{x}) = 1/\tau_0\bar{B}'\Pi\bar{x}$ and Π is the positive-definite stabilizing solution to Riccati equation (59). On the left-hand side of (78), $\bar{x}(\cdot)$ is the solution to (3) corresponding to the given controller of form (2) and a given initial condition. It is shown in [14] that the value on both sides of (78) is independent of this initial condition.

For the function $\phi(\cdot)$ defined above, consider the martingale $\zeta(\cdot) \in \mathcal{M}_\infty$ and the corresponding stationary solution $\bar{x}(\cdot)$ to system (56) with initial distribution ν^ϕ constructed as in Lemma 4. For this martingale $\zeta(\cdot)$ and stationary solution $\bar{x}(\cdot)$, condition (78) leads to the following expression for the risk-sensitive cost:

$$
\begin{aligned}
&\lim_{T\to\infty}\frac{\tau_0}{T}\log\mathbf{E}\exp\left\{\frac{1}{2\tau_0}\int_0^T \bar{x}'(t)(\bar{R}+\tau_0\bar{C}'\bar{C})\bar{x}(t)dt\right\}\\
&=\frac{1}{2}\int\left[\bar{x}'(\bar{R}+\tau_0\bar{C}'\bar{C})\bar{x}-\tau_0\|\phi(x)\|^2\right]\nu^{\phi}(d\bar{x})\\
&=\liminf_{T\to\infty}\frac{1}{2T}\mathbf{E}^{Q^T}\int_0^T F(x(s),u^*(s))ds\\
&\quad+\tau_0\liminf_{T\to\infty}\frac{1}{T}\left[\frac{1}{2}\mathbf{E}^{Q^T}\int_0^T\|z(s)\|^2ds-h(Q^T\|P^T)\right].
\end{aligned}
\tag{79}
$$

Also, note that the right-hand side of the above equation is independent of the initial condition of system (56). This fact is readily established using Ito's formula and the fact that the matrix $\bar{A}+\frac{1}{\tau_0}\bar{B}\bar{B}'\Pi$ is stable. Therefore, on the right-hand side of inequality (79), the stationary process $\bar{x}(\cdot)$ can be replaced by the solution $\bar{x}(\cdot)$ to system (56) corresponding to the given initial condition. Then, (79) and (77) imply that

$$
\lim_{T\to\infty}\frac{\tau_0}{T}\log\mathbf{E}\exp\left\{\frac{1}{2\tau_0}\int_0^T \bar{x}'(t)(\bar{R}+\tau_0\bar{C}'\bar{C})\bar{x}(t)dt\right\}\le c-\bar{\varepsilon}-\frac{\tau_0}{2}d. \tag{80}
$$

Thus,

$$
V_{\tau_0}\le\lim_{T\to\infty}\frac{2\tau_0}{T}\log\mathbf{E}\exp\left[\frac{1}{2\tau_0}\int_0^T \bar{x}'(t)(\bar{R}+\tau_0\bar{C}'\bar{C})\bar{x}(t)dt\right]<2c-\tau_0 d.
$$

Hence the optimal value of the corresponding risk-sensitive control problem (37) is finite. □

6. Design of the infinite-horizon minimax optimal controller. In this section, we present the main result of the paper. This result shows that the solution to an infinite-horizon minimax optimal control problem of the form (31) can be obtained via optimization over solutions to a scaled risk-sensitive control problem of the form (37). Therefore, this result extends the corresponding result of [19] to the case where the underlying system is considered on an infinite time interval.

Consider the class $\mathcal{U}$ of linear controllers of the form (2). In what follows, we will focus on linear output feedback controllers of the form (2) having a controllable and observable state-space realization. The class of such controllers is denoted by $\mathcal{U}_0$.

The derivation of the main result of this paper makes use of parameter-dependent algebraic Riccati equations. Let $\tau>0$ be a constant. We consider the algebraic Riccati equations

$$
\begin{aligned}
&(A-B_2D_2'(D_2D_2')^{-1}C_2)Y_\infty+Y_\infty(A-B_2D_2'(D_2D_2')^{-1}C_2)'\\
&-Y_\infty\left(C_2'(D_2D_2')^{-1}C_2-\frac{1}{\tau}R_\tau\right)Y_\infty+B_2(I-D_2'(D_2D_2')^{-1}D_2)B_2'=0,
\end{aligned}
\tag{81}
$$

$$
\begin{aligned}
&X_\infty(A-B_1G_\tau^{-1}\Upsilon_\tau')+(A-B_1G_\tau^{-1}\Upsilon_\tau')'X_\infty\\
&+(R_\tau-\Upsilon_\tau G_\tau^{-1}\Upsilon_\tau')-X_\infty\left(B_1G_\tau^{-1}B_1'-\frac{1}{\tau}B_2B_2'\right)X_\infty=0.
\end{aligned}
\tag{82}
$$

The subsequent development relies on Theorem 3 of [9]. We now present a version of this theorem adapted to the notation used in this paper. We first note that some

of the conditions of Theorem 3 of [9] are automatically satisfied. Indeed, using the notation

$$\tilde{C}_1 := \begin{bmatrix} \frac{1}{\sqrt{\tau}} R^{1/2} \\ 0 \\ C_1 \end{bmatrix}, \quad \tilde{D}_1 := \begin{bmatrix} 0 \\ \frac{1}{\sqrt{\tau}} G^{1/2} \\ D_1 \end{bmatrix},$$

we obtain $R_\tau - \Upsilon_\tau G_\tau^{-1} \Upsilon_\tau' = \tau \tilde{C}_1'(I - \tilde{D}_1(\tilde{D}_1'\tilde{D}_1)^{-1}\tilde{D}_1')\tilde{C}_1 \geq 0$. Also, the pair $(A - B_1 G_\tau^{-1}\Upsilon_\tau', R_\tau - \Upsilon_\tau G_\tau^{-1}\Upsilon_\tau')$ is detectable since the matrix

$$\begin{bmatrix} A - sI & B_1 \\ R^{1/2} & 0 \\ 0 & G^{1/2} \\ \sqrt{\tau} C_1 & \sqrt{\tau} D_1 \end{bmatrix}$$

has full column rank for all s such that $\mathrm{Re} s \geq 0$.

LEMMA 6. *Consider the risk-sensitive optimal control problem* (37) *with underlying system* (1). *Suppose the pair*

$$(A - B_2 D_2'(D_2 D_2')^{-1} C_2, B_2(I - D_2'(D_2 D_2')^{-1} D_2)) \tag{83}$$

is stabilizable. Also, suppose that there exists a constant $\tau > 0$ *such that the following assumptions are satisfied:*

(i) *Algebraic Riccati equation* (81) *admits a minimal positive-definite solution* Y_∞.
(ii) *Algebraic Riccati equation* (82) *admits a minimal nonnegative-definite solution* X_∞.
(iii) *The matrix* $I - \frac{1}{\tau} Y_\infty X_\infty$ *has only positive eigenvalues; that is, the spectral radius of the matrix* $Y_\infty X_\infty$ *satisfies the condition*

$$\rho(Y_\infty X_\infty) < \tau; \tag{84}$$

where $\rho(\cdot)$ *denotes the spectral radius of a matrix.*

If $Y_\infty \geq Y_0$, *then there exists a controller solving risk-sensitive optimal control problem* (37) *where the infimum is taken over the set* $\mathcal{U}$. *This optimal risk-sensitive controller is a controller of the form* (2) *with*

$$\begin{aligned} K &:= -G_\tau^{-1}(B_1' X_\infty + \Upsilon_\tau'), \\ A_c &:= A + B_1 K - B_c C_2 + \frac{1}{\tau}(B_2 - B_c D_2) B_2' X_\infty, \\ B_c &:= \left(I - \frac{1}{\tau} Y_\infty X_\infty\right)^{-1} (Y_\infty C_2' + B_2 D_2')(D_2 D_2')^{-1}. \end{aligned} \tag{85}$$

The corresponding optimal value of the risk-sensitive cost is given by

$$\begin{aligned} V_\tau &:= \inf_{u \in \mathcal{U}} \lim_{T \to \infty} \Im_{\tau,T}(u(\cdot)) \\ &= \mathrm{tr} \begin{bmatrix} Y_\infty R_\tau + \\ (Y_\infty C_2' + B_2 D_2')(D_2 D_2')^{-1}(C_2 Y_\infty + D_2 B_2') X_\infty (I - \frac{1}{\tau} Y_\infty X_\infty)^{-1} \end{bmatrix}. \end{aligned} \tag{86}$$

Proof. See Theorem 3 of [9]. □

Remark 6. The condition $Y_\infty \geq Y_0$ required by Lemma 6 is a technical condition needed to apply the results of [9] to risk-sensitive control problem (37). However, it can be seen from Lemma 6 that the resulting optimal risk-sensitive controller and the optimal risk-sensitive cost are independent of the matrix Y_0. Therefore, the condition of Lemma 6 requiring $Y_\infty \geq Y_0$ can always be satisfied by a suitable choice of the matrix Y_0.

Reference [9] does not address the issue of stability for the closed-loop system corresponding to the optimal risk-sensitive controller. However, Theorem 1 shows that the controller (2), (85) leads to a robustly stable closed-loop system. This fact is consistent with results showing that risk-sensitive controllers enjoy certain robustness properties; e.g., see [3, 20]. The following results show that the conditions of Lemma 6 are not only sufficient conditions, but also necessary conditions for the existence of a solution to the risk-sensitive optimal control problem under consideration, if such a solution is sought in the class of linear stabilizing controllers; cf. [8].

LEMMA 7. *Suppose the pair* (83) *is controllable and for some* $\tau' > 0$ *there exists an absolutely stabilizing controller* $\tilde{u}(\cdot) \in \mathcal{U}_0$ *such that*

$$V^0_{\tau'} := \lim_{T\to\infty} \Im_{\tau',T}(\tilde{u}) = \inf_{u\in\mathcal{U}_0} \lim_{T\to\infty} \Im_{\tau',T}(u) < +\infty. \tag{87}$$

Then there exists a constant $\tau > 0$ *which satisfies conditions* (i)–(iii) *of Lemma* 6.

Furthermore, if for this τ *the corresponding pairs* (A_c, B_c) *and* (A_c, K) *defined by* (85) *are controllable and observable, respectively, then*

$$V^0_\tau = \operatorname{tr}\left[\begin{array}{l} Y_\infty R_\tau + \\ (Y_\infty C_2' + B_2 D_2')(D_2 D_2')^{-1}(C_2 Y_\infty + D_2 B_2')X_\infty (I - \frac{1}{\tau} Y_\infty X_\infty)^{-1} \end{array}\right]. \tag{88}$$

In the proof of Lemma 7, the following proposition is used.

PROPOSITION 5. *Suppose the pair* (83) *is controllable. Then, for any controller* $u(\cdot) \in \mathcal{U}_0$, *the pair* $(\bar{A}, \bar{B})$ *in the corresponding closed-loop system is controllable and the pair* $(\bar{A}, \bar{R})$ *is observable.*

The proof of this proposition is given in Appendix B.

Proof of Lemma 7. We prove the lemma by contradiction. Suppose that for any $\tau > 0$ at least one of conditions (i)–(iii) of Lemma 6 does not hold. That is, either (81) does not admit a positive-definite stabilizing solution, or (82) does not admit a nonnegative-definite stabilizing solution, or (84) fails to hold. Note that conditions (i)–(iii) of Lemma 6 are standard conditions arising in H^∞ control. Since for any stabilizing controller $u(\cdot) \in \mathcal{U}_0$ the corresponding matrix $\bar{A}$ is stable (see Proposition 5 and Lemma 3), then it follows from standard results on H^∞ control that if at least one of conditions (i)–(iii) of Lemma 6 fails to hold, then for any controller of the form (2)

$$\left\| \left[\begin{array}{cc} \tilde{C}_1 & \tilde{D}_1 K \end{array}\right] (j\omega I - \bar{A})^{-1} \bar{B} \right\|_\infty \geq 1; \tag{89}$$

see Theorem 3.1 of [10]. It is straightforward to verify that the conditions of Theorem 3.1 of [10] are satisfied. Furthermore, the strict bounded real lemma implies that the Riccati equation

$$\bar{A}'\bar{X} + \bar{X}\bar{A} + \frac{1}{\tau}\bar{X}\bar{B}\bar{B}'\bar{X} + \bar{R} + \tau\bar{C}'\bar{C} = 0 \tag{90}$$

does not have a stabilizing positive definite solution. In this case, Lemma 5 implies that none of the controllers $u(\cdot) \in \mathcal{U}_0$ leads to an absolutely stable closed-loop system. This leads to a contradiction with the assumption that an absolutely stabilizing

controller exists and belongs to $\mathcal{U}_0$. This completes the proof by contradiction that there exists a constant τ which satisfies conditions (i)–(iii) of Lemma 6.

It remains to prove (88). Note that Lemma 6 states that for each $\tau > 0$ satisfying the conditions of that lemma, the optimal controller solving risk-sensitive control problem (86) is the controller (2), (85). Furthermore, it is assumed that the state-space realization of this controller is controllable and observable, and hence the optimal controller from Lemma 6 belongs to the set $\mathcal{U}_0$. Therefore,

$$V_\tau^0 = \inf_{u\in\mathcal{U}_0} \lim_{T\to\infty} \Im_{\tau,T}(u(\cdot)) = \inf_{u\in\mathcal{U}} \lim_{T\to\infty} \Im_{\tau,T}(u(\cdot)) = V_\tau. \tag{91}$$

From this observation, (88) follows. □

We now define a set $\mathcal{T} \subset \mathbf{R}$ as the set of constants $\tau \in \mathbf{R}$ satisfying the conditions of Lemma 6. It follows from Lemma 6 that, for any $\tau \in \mathcal{T}$, the controller of form (2) with coefficients given by (85) represents an optimal controller in the risk-sensitive control problem (37), which guarantees the optimal value (88).

THEOREM 3. *Assume that the pair* (83) *is controllable.*

(i) *Suppose that the set* $\mathcal{T}$ *is nonempty and that* $\tau_* \in \mathcal{T}$ *attains the infimum in*

$$\inf_{\tau\in\mathcal{T}} \frac{1}{2}(V_\tau^0 + \tau d), \tag{92}$$

where V_τ^0 *is defined in* (88). *Then the corresponding controller* $u^*(\cdot) := u_{\tau_*}(\cdot)$ *of the form* (2) *defined by* (85), *with the pair* (A_c, B_c) *being controllable and the pair* (A_c, K) *being observable, is an output-feedback controller guaranteeing that*

$$\inf_{u\in\mathcal{U}_0} \sup_{\zeta\in\Xi} J(u,\zeta) \le \sup_{\zeta\in\Xi} J(u^*,\zeta) \le \inf_{\tau\in\mathcal{T}} \frac{1}{2}(V_\tau^0 + \tau d). \tag{93}$$

Furthermore, this controller is an absolutely stabilizing controller for the stochastic uncertain system (1), (8).

(ii) *Conversely, if there exists an absolutely stabilizing minimax optimal controller* $\tilde{u}(\cdot) \in \mathcal{U}_0$ *for the stochastic uncertain system* (1), (8) *such that*

$$\sup_{\zeta\in\Xi} J(\tilde{u},\zeta) < \infty,$$

then the set $\mathcal{T}$ *is nonempty. Moreover,*

$$\inf_{\tau\in\mathcal{T}} \frac{1}{2}(V_\tau^0 + \tau d) \le \sup_{\zeta\in\Xi} J(\tilde{u},\zeta). \tag{94}$$

Proof. Part (i). The conditions of this part of the theorem guarantee that $u^*(\cdot) \in \mathcal{U}_0$. Then $V_{\tau_*}^0 = V_{\tau_*}$. This fact together with Theorem 1 implies that

$$\inf_{u\in\mathcal{U}_0} \sup_{\zeta\in\Xi} J(u,\zeta) \le \sup_{\zeta\in\Xi} J(u^*,\zeta) \le \frac{1}{2}(V_{\tau_*}^0 + \tau_* d) = \inf_{\tau\in\mathcal{T}} \frac{1}{2}(V_\tau^0 + \tau d). \tag{95}$$

Also from Theorem 1, the controller $u^*(\cdot)$ solving the corresponding risk-sensitive control problem is an absolutely stabilizing controller. From this observation, part (i) of the theorem follows.

Part (ii). Note that the controller $\tilde{u}(\cdot) \in \mathcal{U}_0$ satisfies the conditions of Theorem 2; see Proposition 5. Let c be a constant such that

$$\sup_{\zeta\in\Xi} J(\tilde{u},\zeta) < c.$$

When proving Theorem 2, it was shown that there exists a constant $\tau > 0$ such that Riccati equation (90) has a stabilizing positive-definite solution and

$$\frac{1}{2}\lim_{T\to\infty}\Im_{\tau,T}(\tilde{u}) < c - \frac{\tau}{2}d < \infty; \tag{96}$$

see (80). Hence, $V_\tau^0 < \infty$. From the above conditions and using Lemma 7, we conclude that the set $\mathcal{T}$ is nonempty.

We now prove (94). Consider a sequence $\{c_i\}$, $i = 1, 2, \ldots$, such that

$$c_i \downarrow \sup_{\zeta\in\Xi} J(\tilde{u}, \zeta) \quad \text{as } i \to \infty.$$

From (96) it follows that

$$\inf_{\tau\in\mathcal{T}} \frac{1}{2}(V_\tau^0 + \tau d) < c_i.$$

Hence, letting i approach infinity leads to the satisfaction of (94). □

The first part of Theorem 3 provides a sufficient condition for the existence of an optimal solution to the minimax LQG control problem considered in this section. This condition is given in terms of certain Riccati equations. This makes the result useful in practical controller design since there is a wide range of software available for solving such Riccati equations.

In the control literature, there is a great deal of interest concerning the issue of conservatism in robust controller design. For example, a significant issue considered in [15, 16, 18] is to prove that the results on the minimax optimal control considered in those papers are not conservative, in that the corresponding Riccati equations fail to have stabilizing solutions if the minimax optimal controller does not exist. Thus, the conditions for the existence of a minimax optimal controller presented in those sections are necessary and sufficient conditions. The second part of Theorem 3 is analogous to the necessity results of [15, 16, 18, 19]. It follows from this part of Theorem 3 that the controller $u^*(\cdot)$ constructed in the first part of Theorem 3 represents a minimax optimal controller in the subclass $\mathcal{U}_{0,\text{stab}} \subset \mathcal{U}_0$ of stabilizing linear output feedback controllers. This result is summarized in the following theorem.

THEOREM 4. *Assume that the conditions of part* (i) *of Theorem* 3 *are satisfied. Then, the controller* $u^*(\cdot)$ *constructed in part* (i) *of Theorem* 3 *is the minimax optimal controller such that*

$$\inf_{u\in\mathcal{U}_{0,\text{stab}}}\sup_{\zeta\in\Xi} J(u,\zeta) = \sup_{\zeta\in\Xi} J(u^*,\zeta) = \inf_{\tau\in\mathcal{T}}\frac{1}{2}(V_\tau^0 + \tau d). \tag{97}$$

Proof. It was shown in part (i) of Theorem 3 that the controller $u^*(\cdot)$ belongs to the set $\mathcal{U}_{0,\text{stab}}$. Hence,

$$\inf_{u\in\mathcal{U}_{0,\text{stab}}}\sup_{\zeta\in\Xi} J(u,\zeta) \le \sup_{\zeta\in\Xi} J(u^*,\zeta) \le \inf_{\tau\in\mathcal{T}}\frac{1}{2}(V_\tau^0 + \tau d). \tag{98}$$

Furthermore, condition (98) implies that

$$\inf_{u\in\mathcal{U}_{0,\text{stab}}}\sup_{\zeta\in\Xi} J(u,\zeta) < \infty.$$

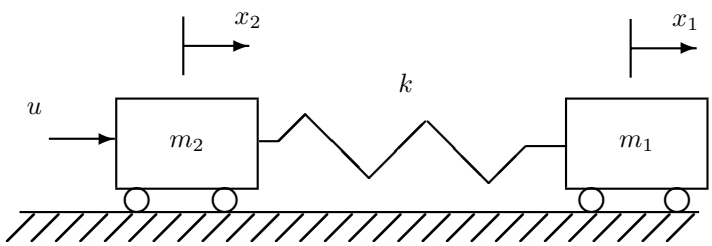


FIG. 2. *A two mass spring system.*

That is, for any sufficiently small $\varepsilon > 0$, there exists a controller $\tilde{u}(\cdot) \in \mathcal{U}_{0,\text{stab}}$ such that

$$\sup_{\zeta\in\Xi} J(\tilde{u},\zeta) \le \inf_{u\in\mathcal{U}_{0,\text{stab}}} \sup_{\zeta\in\Xi} J(u,\zeta) + \varepsilon.$$

This controller satisfies the conditions of part (ii) of Theorem 3. Therefore, it follows from Theorem 3 that

$$\inf_{\tau\in\mathcal{T}} \frac{1}{2}(V_\tau^0 + \tau d) \le \inf_{u\in\mathcal{U}_{0,\text{stab}}} \sup_{\zeta\in\Xi} J(u,\zeta) + \varepsilon.$$

The above inequality holds for any infinitesimal $\varepsilon > 0$. Therefore,

$$\inf_{\tau\in\mathcal{T}} \frac{1}{2}(V_\tau^0 + \tau d) \le \inf_{u\in\mathcal{U}_{0,\text{stab}}} \sup_{\zeta\in\Xi} J(u,\zeta).$$

This inequality together with (98) implies (97). □

7. Illustrative example. We now consider the tracking problem which was used as an illustrative example in [15, 17]. In this tracking problem, the goal is to design an output-feedback controller so that the controlled output of a two-cart system tracks a reference step input. The system to be controlled is shown in Figure 2.

As in [15, 17], the masses of the carts are assumed to be $m_1 = 1$ and $m_2 = 1$. Furthermore, the spring constant k is treated as an uncertain parameter subject to the bound $0.5 \le k \le 2.0$. From this, a corresponding uncertain system was derived in [17]. This uncertain system is described by the following state equations:

$$\begin{aligned}
\dot{x} &= \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1.25 & 1.25 & 0 & 0 \\ 1.25 & -1.25 & 0 & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} u + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -0.70 & 0 & 0 \\ 0.80 & 0 & 0 \end{bmatrix} \xi, \\
z &= \begin{bmatrix} 1 & -1 & 0 & 0 \end{bmatrix} x, \\
y &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} x + \begin{bmatrix} 0 & 0.05 & 0 \\ 0 & 0 & 0.05 \end{bmatrix} \xi, \\
y_T &= \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} x.
\end{aligned} \tag{99}$$

Here, the uncertainty is subject to an integral quadratic constraint which will be specified below. The output y_T is the output which is required to track a step input.

The control problem solved in [17] involved finding a controller which absolutely stabilized the system and also ensured that the output y_T tracks a reference step input. In [17], the system was transformed into the form:

$$\begin{aligned}
\dot{\bar{x}} &= \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -2.5 & 0 \end{bmatrix} \bar{x} + \begin{bmatrix} 0 \\ 0.5 \\ 0 \\ -0.5 \end{bmatrix} u + \begin{bmatrix} 0 & 0 & 0 \\ 0.05 & 0 & 0 \\ 0 & 0 & 0 \\ -0.75 & 0 & 0 \end{bmatrix} \xi, \\
z &= \begin{bmatrix} 0 & 0 & 2 & 0 \end{bmatrix} \bar{x}, \\
\bar{y} &= \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 \end{bmatrix} \bar{x} + \begin{bmatrix} 0 & 0.05 & 0 \\ 0 & 0 & 0.05 \end{bmatrix} \xi, \\
y_T - \tilde{y}_T &= \begin{bmatrix} 1 & 0 & 1 & 0 \end{bmatrix} \bar{x},
\end{aligned} \tag{100}$$

where

$$\bar{y} := y - \begin{bmatrix} 1 \\ 1 \end{bmatrix} \eta.$$

Here, η denotes the state of the reference input signal model:

$$\begin{aligned}
&\dot{\eta} = 0, \quad \eta(0) = 1, \\
&\tilde{y}_T = \eta.
\end{aligned} \tag{101}$$

The above transformation involved the following change of variables:

$$\begin{aligned}
\bar{x}_1 &= (x_1 + x_2)/2 - \eta, \\
\bar{x}_2 &= (\dot{x}_1 + \dot{x}_2)/2 = (x_3 + x_4)/2, \\
\bar{x}_3 &= (x_1 - x_2)/2, \\
\bar{x}_4 &= (\dot{x}_1 - \dot{x}_2)/2 = (x_3 - x_4)/2.
\end{aligned} \tag{102}$$

To construct the required controller, the following cost function was used:

$$\int_0^\infty [(y_T - \tilde{y}_T)^2 + 0.1\|\bar{x}\|^2 + u^2]dt. \tag{103}$$

Hence, the matrices R and G are as defined in [17]:

$$R = \begin{bmatrix} 1.1 & 0 & 1 & 0 \\ 0 & 0.1 & 0 & 0 \\ 1 & 0 & 1.1 & 0 \\ 0 & 0 & 0 & 0.1 \end{bmatrix} > 0, \qquad G = 1.$$

In (100), the uncertainty input $\xi(\cdot)$ has three components,

$$\xi(\cdot) = [\xi_1(\cdot),\ \xi_2(\cdot),\ \xi_3(\cdot)]'.$$

The uncertainty input $\xi_1(\cdot)$ describes the uncertainty in the spring rate. This uncertainty satisfies the constraint

$$|\xi_1(t)| \le |z(t)|.$$

The components ξ_2 and ξ_3 of the uncertainty input vector ξ are fictitious uncertainty inputs which were added to system (99) in [17] in order to fit this system into the framework of the method presented in that paper. Specifically, it was assumed in [17] that the uncertainty input $\xi(\cdot)$ satisfies the following integral quadratic constraint:

$$\int_0^{t_i} \|\xi(t)\|^2 dt \leq \int_0^{t_i} \|z(t)\|^2 dt + \bar{x}_0' S \bar{x}_0, \tag{104}$$

where $\{t_i\}$ is a sequence of times as discussed in [17]. Also, in [17], the initial condition of system (100) was chosen to be $\bar{x}_0 = [-1\ 0\ 0\ 0]'$. This choice of the initial condition corresponds to a zero initial condition on the system dynamics and an initial condition of $\eta(0) = 1$ on the reference input dynamics. Also, the mismatch matrix S was chosen to be

$$S = \begin{bmatrix} 0.1 & 0 & 0 & 0 \\ 0 & 0.1 & 0 & 0 \\ 0 & 0 & 0.1 & 0 \\ 0 & 0 & 0 & 0.1 \end{bmatrix} > 0.$$

The output-feedback robust controller designed in [17] was a suboptimal time-varying controller. We now apply the controller design procedure presented in this paper to design a time-invariant output-feedback minimax optimal controller solving the above tracking problem. We will use the state space transformation (102), which reduces the original tracking problem to a regulator problem. However, in order to apply the results of this paper to this robust control problem, we must introduce a stochastic description of the system. To satisfy this requirement, a noise input will be added to the system, and the controller will be designed for the system with additive noise. That is, we replace the nominal system corresponding to (100) with $\xi(\cdot) \equiv 0$ with a stochastic system described by the following stochastic differential equation:

$$\begin{aligned}
d\bar{x} &= \left[\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -2.5 & 0 \end{bmatrix} \bar{x} + \begin{bmatrix} 0 \\ 0.5 \\ 0 \\ -0.5 \end{bmatrix} u \right] dt + \begin{bmatrix} 0 & 0 & 0 \\ 0.05 & 0 & 0 \\ 0 & 0 & 0 \\ -0.75 & 0 & 0 \end{bmatrix} dW(t), \\
z &= \begin{bmatrix} 0 & 0 & 2 & 0 \end{bmatrix} \bar{x}, \\
d\bar{y} &= \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 \end{bmatrix} \bar{x} dt + \begin{bmatrix} 0 & 0.05 & 0 \\ 0 & 0 & 0.05 \end{bmatrix} dW, \\
y_T - \tilde{y}_T &= \begin{bmatrix} 1 & 0 & 1 & 0 \end{bmatrix} \bar{x},
\end{aligned} \tag{105}$$

where $W(t) = [W_1(t),\ W_2(t),\ W_3(t)]'$ is a 3-dimensional Wiener process on a certain measurable space $(\Omega, \mathcal{F}, P)$. Here, P is the reference probability measure. Also, the uncertain system (100) is replaced by an uncertain system of the form (105) considered on an uncertain measurable space defined using an uncertain martingale $\zeta(\cdot)$. Also, as noted in section 2, uncertain systems of this type can be described using a stochastic differential equation of the form (13). System (105) is a system of the form (1) to which the design technique presented in this paper is applicable.

Note that in this example, a robust controller is sought which stabilizes the system in the face of stochastic uncertainty. It can readily be shown using Lemma 5 that the absolute stability of the stochastic closed-loop system consisting of system

(105) and this controller implies the robust stability of the closed-loop system corresponding to the deterministic system (100) driven by the same linear output-feedback controller. Indeed, Lemma 5 shows that the corresponding Riccati equation (59) has a nonnegative-definite stabilizing solution. Then, using the strict bounded real lemma leads to the conclusion that the corresponding deterministic closed-loop system with norm-bounded uncertainty is quadratically stable [5]. Also, the corresponding deterministic closed-loop system with the uncertainty modeled using an integral quadratic constraint of the form (104) is absolutely stable [23]. It follows from this observation that a robust output-feedback controller designed for the uncertain stochastic system (105) also serves as a robust controller for the original uncertain system (100). Thus, a controller designed for stochastic uncertain system (105) will solve the original tracking problem.

We now proceed to the derivation of a robust output-feedback controller for system (105). We first replace the integral quadratic constraint (104) by the following stochastic uncertainty constraint: For any $T > 0$

$$\frac{1}{2T}\mathbf{E}^{Q^T}\int_0^T \|\xi(t)\|^2 dt \le \frac{1}{2T}\mathbf{E}^{Q^T}\int_0^T \|z(t)\|^2 dt + d, \quad d = \frac{1}{2}\bar{x}_0' S \bar{x}_0. \tag{106}$$

It was shown in section 2 that the uncertainty class defined by the constraint (106) can be embedded into an uncertainty class described by the corresponding relative entropy uncertainty constraint of the form (8).

The cost functional is chosen to have the form

$$\limsup_{T\to\infty} \frac{1}{2T}\int_0^T \mathbf{E}^{Q^T}[(y_T - \tilde{y}_T)^2 + 0.1\|\bar{x}\|^2 + u^2]dt. \tag{107}$$

We are now in a position to apply the design procedure outlined in Theorem 3. For each value of $\tau > 0$, the Riccati equations (81) and (82) are solved, and then a line search is carried out to find the value of $\tau > 0$ which attains the minimum of the function $1/2(V_\tau^0 + \tau d)$ defined in Theorem 3. A graph of $1/2(V_\tau^0 + \tau d)$ versus τ for this example is shown in Figure 3. It was found that the optimal value of the parameter τ is $\tau = 5.6931$.

With this optimal value of τ, the following positive-definite stabilizing solutions to Riccati equations (82) and (81) were obtained:

$$X_\infty = \begin{bmatrix} 4.0028 & 6.8156 & -6.3708 & 3.8312 \\ 6.8156 & 18.3891 & -20.6541 & 9.3784 \\ -6.3708 & -20.6541 & 48.5330 & -5.8268 \\ 3.8312 & 9.3784 & -5.8268 & 12.5738 \end{bmatrix},$$

$$Y_\infty = \begin{bmatrix} 0.0007 & 0.0003 & -0.0005 & -0.0014 \\ 0.0003 & 0.0008 & -0.0017 & -0.0108 \\ -0.0005 & -0.0017 & 0.0077 & 0.0236 \\ -0.0014 & -0.0108 & 0.0236 & 0.1641 \end{bmatrix}.$$

Furthermore, a corresponding time-invariant controller of the form (2), (85) was

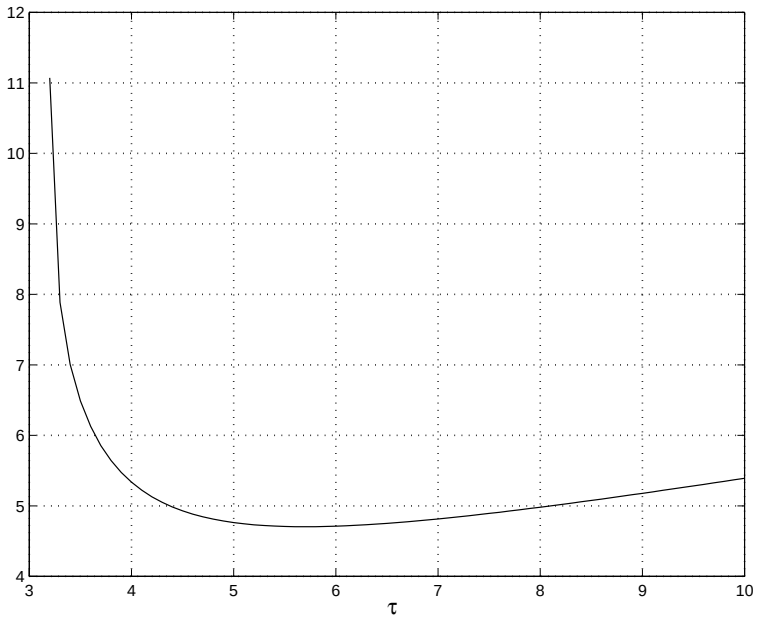


FIG. 3. *Cost bound* $\frac{1}{2}(V_\tau^0 + \tau d)$ *versus the parameter* τ.

constructed to be

$$d\hat{x} = \begin{bmatrix} -0.5868 & 1.0000 & 0.4581 & 0 \\ -1.0384 & -2.3064 & 5.7466 & 0.7202 \\ 0.4581 & 0 & -7.6627 & 1.0000 \\ 2.6530 & 3.0582 & -34.7464 & 0.3817 \end{bmatrix} \hat{x}dt + \begin{bmatrix} 0.0643 & 0.5225 \\ -0.8702 & 1.1403 \\ 3.6023 & -4.0604 \\ 13.2633 & -14.8366 \end{bmatrix} dy(t),$$

$$u = \begin{bmatrix} -1.4922 & -4.5053 & 7.4137 & 1.5977 \end{bmatrix} \hat{x}.$$

Then referring to system (99), the required tracking control system is constructed by replacing the time-varying controller of [17] with the above time-invariant controller as shown in Figure 4. To verify the robust tracking properties of this control system, Figure 5 shows the step response of the system for various values of the spring constant parameter k. It can be seen from these plots that the stochastic minimax optimization approach of this paper leads to a robust tracking system which exhibits transient behavior similar to the behavior of the tracking system designed using the deterministic approach of [17]. However, the controller designed using the approach of this paper is time-invariant.

Appendix A. Relative entropy. This appendix presents a result on the duality between free energy and relative entropy which is exploited in this paper. This result is taken from [1].

Let $(\Omega, \mathcal{F})$ be a measurable space, and let $\mathcal{P}(\Omega)$ be the set of probability measures on $(\Omega, \mathcal{F})$.

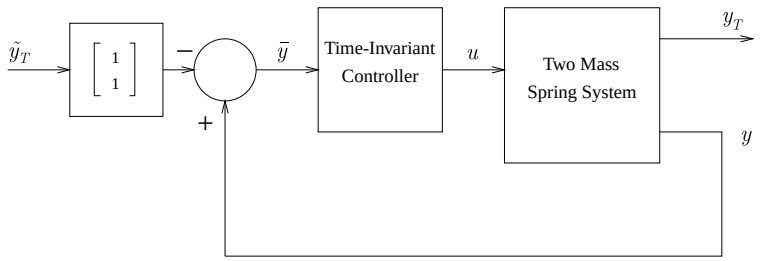


FIG. 4. *Block diagram of a tracking control system.*

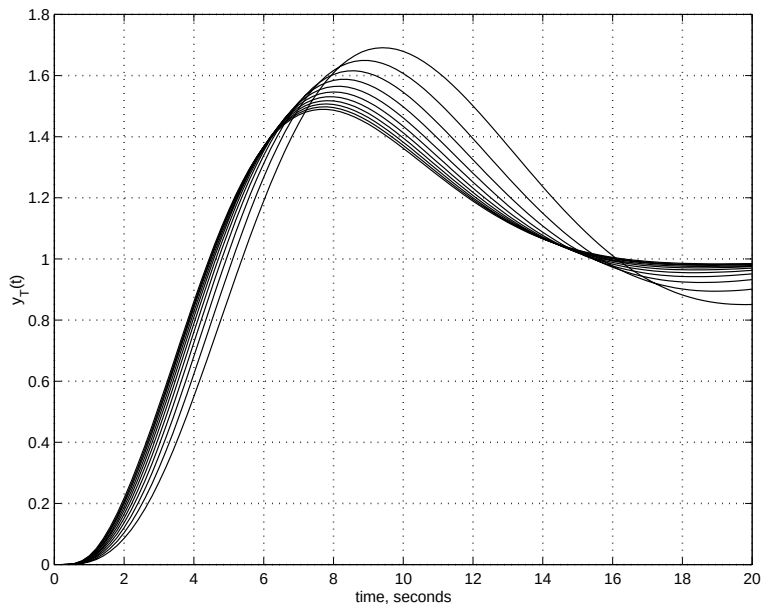


FIG. 5. *Control system step response for various spring constants.*

DEFINITION 4. *Let $P \in \mathcal{P}(\Omega)$, and $\psi : \Omega \to \mathbf{R}$ be a measurable function. The quantity*

$$\mathbb{E} := \log\left(\int e^{\psi} P(d\omega)\right)$$

is called the free energy of ψ with respect to P.

DEFINITION 5. *Given any two probability measures $Q, P \in \mathcal{P}(\Omega)$, the relative entropy of the probability measure Q with respect to the probability measure P is defined by*

$$h(Q\|P) := \begin{cases} \int \log\left(\frac{dQ}{dP}\right) Q(d\omega) & \text{if } Q \ll P \text{ and } \log\left(\frac{dQ}{dP}\right) \in L_1(\Omega, \mathcal{F}, Q), \\ +\infty & \text{otherwise.} \end{cases} \tag{A.1}$$

In the above definition, $\frac{dQ}{dP}$ is the Radon–Nikodým derivative of the probability measure Q with respect to the probability measure P. Note that the relative entropy is a convex, lower semicontinuous functional of Q; e.g., see [2]. It is shown in [1]

that the functions $\mathbb{E}(\psi)$ and $h(Q\|P)$ are in duality with respect to a Legendre-type transform as follows.

LEMMA 8.

(i) *For every* $Q \in \mathcal{P}(\Omega)$,

$$h(Q\|P) = \sup_{\substack{e^{\psi} \in L_1(\Omega,\mathcal{F},P),\\ \psi \text{ bounded below}}} \left\{ \int \psi Q(d\omega) - \mathbb{E}(\psi) \right\}; \tag{A.2}$$

(ii) *For every* ψ *bounded from below,*

$$\mathbb{E}(\psi) = \sup_{h(Q\|P)<\infty} \left\{ \int \psi Q(d\omega) - h(Q\|P) \right\}. \tag{A.3}$$

Moreover, if $\psi e^{\psi} \in L_1(\Omega, \mathcal{F}, P)$, *then the supremum in* (A.3) *is attained at* Q^* *given by*

$$\frac{dQ^*}{dP} = \frac{e^{\psi}}{\int e^{\psi} P(d\omega)}.$$

Proof. See [1]. □

Appendix B. Proofs.

Proof of Lemma 3. Since the stochastic nominal system (27) satisfies condition (29), then for any vector y

$$\limsup_{T\to\infty} \frac{1}{T} \int_0^T y' \mathbf{E}\left[x(t)x'(t)\right] y dt < \infty. \tag{B.1}$$

We will prove that the stability of the matrix A follows from condition (B.1). This proof is by contradiction.

Suppose that the matrix A is not stable and therefore it has a left eigenvalue λ such that $\mathrm{Re}\lambda \geq 0$. Consider the left eigenvector y of the matrix A corresponding to the eigenvalue λ. Hence, $y'A = y'\lambda$. Here, y' denotes the Hermitian conjugate of y. Since the pair (A, B_2) is stabilizable, it follows that $y'B_2 \neq 0$ and, consequently,

$$y' B_2 B_2' y > 0. \tag{B.2}$$

We now consider the following two cases.

Case 1. $\mathrm{Re}\lambda > 0$. In this case, we obtain the following bound on $y'\mathbf{E}\left[x(t)x'(t)\right] y$:

$$\begin{aligned} y'\mathbf{E}\left[x(t)x'(t)\right] y &= y' e^{At} \mathbf{E}\left[x(0)x'(0)\right] e^{A't} y + \int_0^t y' e^{A(t-s)} B_2 B_2' e^{A'(t-s)} y ds \\ &= e^{2\mathrm{Re}\lambda t} y' \mathbf{E}\left[x(0)x'(0)\right] y + \int_0^t e^{2\mathrm{Re}\lambda(t-s)} y' B_2 B_2' y ds \\ &\geq \frac{e^{2\mathrm{Re}\lambda t} - 1}{2\mathrm{Re}\lambda} y' B_2 B_2' y. \end{aligned}$$

Thus for any $T > 0$

$$\frac{1}{T} \int_0^T y' \mathbf{E}\left[x(t)x'(t)\right] y dt \geq y' B_2 B_2' y \cdot \frac{1}{T} \int_0^T \frac{e^{2\mathrm{Re}\lambda t} - 1}{2\mathrm{Re}\lambda} dt. \tag{B.3}$$

Case 2. $\mathrm{Re}\lambda = 0$. In this case, we obtain the following bound on $y'\mathbf{E}\left[x(t)x'(t)\right]y$:

$$\begin{aligned} y'\mathbf{E}\left[x(t)x'(t)\right]y &= y'e^{At}\mathbf{E}\left[x(0)x'(0)\right]e^{A't}y + \int_0^t y'e^{A(t-s)}B_2B_2'e^{A'(t-s)}y ds \\ &= y'\mathbf{E}\left[x(0)x'(0)\right]y + \int_0^t y'B_2B_2'y ds \\ &\geq t\cdot y'B_2B_2'y. \end{aligned}$$

Thus for any $T > 0$

$$\frac{1}{T}\int_0^T y'\mathbf{E}\left[x(t)x'(t)\right]y dt \geq y'B_2B_2'y\cdot\frac{1}{T}\frac{T^2}{2} = y'B_2B_2'y\frac{T}{2}. \tag{B.4}$$

Since $y'B_2B_2'y > 0$, the expressions on the right-hand side of inequalities (B.3) and (B.4) both approach infinity as $T\to\infty$. That is, in both cases,

$$\limsup_{T\to\infty}\frac{1}{T}\int_0^T y'\mathbf{E}\left[x(t)x'(t)\right]y dt = \infty.$$

This yields the desired contradiction with (B.1). □

Proof of Proposition 5. Note that by definition, for any controller $u(\cdot)\in\mathcal{U}_0$, the corresponding pair (A_c, B_c) is controllable and the pair (A_c, K) is observable.

To prove the controllability of the pair $(\bar{A},\bar{B})$, we first note that the matrix

$$\begin{bmatrix} A'-sI & C_2' \\ B_2' & D_2' \end{bmatrix} \tag{B.5}$$

has full column rank for all $s\in\mathbb{C}$ [25].

Next, consider the matrix pair

$$(\bar{A},\bar{B}) = \left(\begin{bmatrix} A & B_1K \\ B_cC_2 & A_c \end{bmatrix}, \begin{bmatrix} B_2 \\ B_cD_2 \end{bmatrix}\right). \tag{B.6}$$

For this matrix pair to be controllable, the equations

$$(A'-sI)x_1 + C_2'B_c'x_2 = 0, \tag{B.7a}$$

$$B_2'x_1 + D_2'B_c'x_2 = 0, \tag{B.7b}$$

$$K'B_1'x_1 + (A_c'-sI)x_2 = 0 \tag{B.7c}$$

must imply that $x_1 = 0$ and $x_2 = 0$ for every $s\in\mathbb{C}$. Equations (B.7a) and (B.7b) can be written as follows:

$$\begin{bmatrix} A'-sI & C_2' \\ B_2' & D_2' \end{bmatrix}\begin{bmatrix} x_1 \\ B_c'x_2 \end{bmatrix} = 0.$$

It was noted above that the matrix (B.5) has full column rank for all $s\in\mathbb{C}$. Hence, the above equation and (B.7c) imply that

$$x_1 = 0,\quad B_c'x_2 = 0,\quad (A_c'-sI)x_2 = 0.$$

Since the pair (A_c, B_c) is controllable, then the two last equations imply that $x_2 = 0$. Thus, the pair (B.6) is controllable.

In order to prove the observability of the pair $(\bar{A}, \bar{R})$, we need to show that the equations

$$(A - sI)x_1 + B_1 K x_2 = 0, \tag{B.8a}$$

$$B_c C_2 x_1 + (A_c - sI)x_2 = 0, \tag{B.8b}$$

$$R^{1/2} x_1 = 0, \tag{B.8c}$$

$$G^{1/2} K x_2 = 0 \tag{B.8d}$$

imply that $x_1 = 0$, $x_2 = 0$ for every $s \in \mathbb{C}$. Indeed, since the matrices R, G are positive-definite, then it follows from (B.8c) and (B.8d) that $x_1 = 0$ and $Kx_2 = 0$. Using these equations, we also obtain from (B.8b) that $(A_c - sI)x_2 = 0$. Since the pair (A_c, K) is observable, this implies that $x_1 = 0$ and $x_2 = 0$. Thus, the pair $(\bar{A}, \bar{R})$ is observable. □

REFERENCES

[1] P. Dai Pra, L. Meneghini, and W. Runggaldier, *Connections between stochastic control and dynamic games*, Math. Control Signals Systems, 9 (1996), pp. 303–326.

[2] P. Dupuis and R. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*, Wiley, New York, 1997.

[3] P. Dupuis, M. R. James, and I. R. Petersen, *Robust properties of risk–sensitive control*, in Proceedings of the IEEE Conference on Decision and Control, Tampa, FL, 1998, pp. 2365–2370.

[4] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York, 1988.

[5] P. P. Khargonekar, I. R. Petersen, and K. Zhou, *Robust stabilization of uncertain systems and H^∞ optimal control*, IEEE Trans. Automat. Control, AC-35 (1990), pp. 356–361.

[6] R. S. Liptser and A. N. Shiryayev, *Statistics of Random Processes.* I. *General Theory*, Springer-Verlag, New York, 1977.

[7] D. G. Luenberger, *Optimization by Vector Space Methods*, Wiley, New York, 1969.

[8] D. Mustafa and K. Glover, *Minimum Entropy H_∞ Control*, Springer-Verlag, Berlin, 1990.

[9] Z. Pan and T. Başar, *Model simplification and optimal control of stochastic singularly perturbed systems under exponentiated quadratic cost*, SIAM J. Control Optim., 34 (1996), pp. 1734–1766.

[10] I. R. Petersen, B. D. O. Anderson, and E. A. Jonckheere, *A first principles solution to the non-singular H^∞ control problem*, Internat. J. Robust Nonlinear Control, 1 (1991), pp. 171–185.

[11] I. R. Petersen and M. R. James, *Performance analysis and controller synthesis for nonlinear systems with stochastic uncertainty constraints*, Automatica, 32 (1996), pp. 959–972.

[12] I. R. Petersen, M. R. James, and P. Dupuis, *Minimax optimal control of stochastic uncertain systems with relative entropy constraints*, IEEE Trans. Automat. Control, 45 (2000), pp. 398–412.

[13] J. F. Randolph, *Basic Real and Abstract Analysis*, Academic Press, New York, London, 1968.

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

[15] A. V. Savkin and I. R. Petersen, *Minimax optimal control of uncertain systems with structured uncertainty*, Internat. J. Robust Nonlinear Control, 5 (1995), pp. 119–137.

[16] A. V. Savkin and I. R. Petersen, *Nonlinear versus linear control in the absolute stabilizability of uncertain linear systems with structured uncertainty*, IEEE Trans. Automat. Control, 40 (1995), pp. 122–127.

[17] A. V. Savkin and I. R. Petersen, *Output feedback guaranteed cost control of uncertain systems on an infinite time interval*, Internat. J. Robust Nonlinear Control, 7 (1997), pp. 43–58.

[18] V. A. Ugrinovskii and I. R. Petersen, *Absolute stabilization and minimax optimal control of uncertain systems with stochastic uncertainty*, SIAM J. Control Optim., 37 (1999), pp. 1089–1122.

[19] V. A. Ugrinovskii and I. R. Petersen, *Finite horizon minimax optimal control of stochastic partially observed time varying uncertain systems*, Math. Control Signals Systems, 12 (1999), pp. 1–23.

[20] V. A. Ugrinovskii and I. R. Petersen, *Robust output feedback stabilization via risk-sensitive control*, in Proceedings of the IEEE Conference on Decision and Control, Phoenix, AZ, 1999, pp. 546–551.

[21] V. A. Ugrinovskii and I. R. Petersen, *Robust stability and performance of stochastic uncertain systems on an infinite time interval*, Systems Control Lett., to appear.

[22] J. C. Willems, *Least-squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, AC-16 (1971), pp. 621–634.

[23] V. A. Yakubovich, *Dichotomy and absolute stability of nonlinear systems with periodically nonstationary linear part*, Systems Control Lett., 11 (1988), pp. 221–228.

[24] M. Zakai, *A Lyapunov criterion for the existence of stationary probability distributions for systems perturbed by noise*, SIAM J. Control Optim., 7 (1969), pp. 390–397.

[25] K. Zhou, J. Doyle, and K. Glover, *Robust and Optimal Control*, Prentice-Hall, Upper Saddle River, NJ, 1996.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 4, pp. 1227–1249

A CHARACTERIZATION OF THE LIE ALGEBRA RANK CONDITION BY TRANSVERSE PERIODIC FUNCTIONS*

PASCAL MORIN† AND CLAUDE SAMSON†

Abstract. The Lie algebra rank condition plays a central role in nonlinear systems control theory. The present paper establishes that the satisfaction of this condition by a set of smooth control vector fields is equivalent to the existence of smooth transverse periodic functions. The proof here enclosed is constructive and provides an explicit method for the synthesis of such functions.

Key words. controllability, driftless system, transversality, Lie algebra

AMS subject classifications. 93B05, 93B29, 93C10

PII. S0363012900366054

1. Introduction. Let $X_1, \dots, X_m$ denote smooth vector fields (v.f.) on a smooth n-dimensional manifold M. By definition, the Lie algebra rank condition at a point $p_0 \in M$ ($LARC(p_0)$) is the property that[1]

$$M_{p_0} = \mathrm{Span}\{X(p_0) : X \in \mathrm{Lie}(X_1, \dots, X_m)\},$$

where $\mathrm{Lie}(X_1, \dots, X_m)$ denotes the Lie algebra of v.f. generated by $X_1, \dots, X_m$. This condition plays a major role in the study of controllability properties of nonlinear control systems, as shown in the classical works of Chow [2], Lobry [10], Hermann [4], Sussmann and Jurdjevic [18], and others. For example, the well-known "Chow's theorem" states that if $LARC(p_0)$ is satisfied for the v.f. $X_1, \dots, X_m$, then the set of points reachable from p_0 by trajectories of the control system

$$\dot{p} = \sum_{i=1}^{m} u_i X_i(p) \tag{1}$$

contains a neighborhood of p_0. While the Lie algebra rank condition provides a systematic tool to test the controllability of system (1), its use at the control design level is usually not direct. For instance, even though $LARC(p_0)$ implies the existence of elements $X_{m+1}, \dots, X_{\bar{n}}$ of $\mathrm{Lie}(X_1, \dots, X_m)$ such that

$$\forall p \in \mathcal{V}, \quad M_p = \mathrm{Span}\{X_1(p), \dots, X_m(p)\} + \mathrm{Span}\{X_{m+1}(p), \dots, X_{\bar{n}}(p)\}, \tag{2}$$

where $\mathcal{V}$ denotes a neighborhood of p_0, the "generation of motion" in the direction of the v.f. $X_{m+1}, \dots, X_{\bar{n}}$ by means of the control variables u_i is not simple. Although general results have been obtained for this problem in both the open-loop [9] and closed-loop [11] contexts, their application to physical systems usually raises several difficult issues—complexity, robustness, etc.

In this paper, we present a characterization of the Lie algebra rank condition which allows us to consider the control of system (1) from a slightly different perspective. More precisely, the following result is proved.

*Received by the editors January 20, 2000; accepted for publication (in revised form) May 22, 2001; published electronically December 7, 2001.
http://www.siam.org/journals/sicon/40-4/36605.html
†INRIA, B.P. 93, 06902 Sophia-Antipolis Cedex, France (pascal.morin@inria.fr,claude.samson@inria.fr).

[1]Throughout the paper, the notation N_q is used to denote the tangent space of a manifold N at q, whereas T_qF denotes the tangent mapping of a smooth map F at q.

THEOREM 1. *Let $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$ denote the one-dimensional torus, and let $X_1, \ldots, X_m$ denote smooth v.f. on a smooth n-dimensional manifold M, such that the accessibility distribution $\Delta(p) \triangleq Span\{X(p) : X \in Lie(X_1, \ldots, X_m)\}$ is of constant dimension n_0 in a neighborhood of p_0. Then the following properties are equivalent:*

1. *$n_0 = n$; i.e., the Lie algebra rank condition at p_0, $LARC(p_0)$, is satisfied for the v.f. $X_1, \ldots, X_m$.*
2. *There exist $\bar{n} \in \mathbb{N}$ and, for any neighborhood $\mathcal{U}$ of p_0, a function $F \in \mathcal{C}^\infty(\mathbb{T}^{\bar{n}-m}; \mathcal{U})$ such that*

$$\forall \theta \in \mathbb{T}^{\bar{n}-m}, \quad M_{F(\theta)} = Span\{X_1(F(\theta)), \ldots, X_m(F(\theta))\} + T_\theta F(\mathbb{T}^{\bar{n}-m}_\theta)\,. \tag{3}$$

REMARK 1.

1. *Relation (3) is reminiscent of the transversality property for functions—see, e.g.,* [1, Section 3.5] *for a definition.*
2. *It is clear that $\bar{n}$ is at least equal to n. For some systems—in particular, for free systems introduced later—it can be chosen equal to n, so that the sum in the right-hand side of (3) becomes direct, and F is an immersion.*

Roughly speaking, by comparison with (2), equality (3) implies that at any point $F(\theta) \in M$, the directions $X_{m+1}(F(\theta)), \ldots, X_{\bar{n}}(F(\theta))$, which are not directly available for control, are spanned by the partial derivatives of the smooth function F. An important property of this characterization is that the function F can be directly used for control design purposes. In order to briefly illustrate this fact (for more details on potential applications, the reader is referred to [13]), let us consider the well-known chain system on $\mathbb{R}^3$, where $p = (p_1, p_2, p_3)^T \in \mathbb{R}^3$:

$$\dot{p} = u_1 X_1(p) + u_2 X_2\,, \quad X_1(p) = (1, 0, p_2)^T\,, X_2 = (0, 1, 0)^T \tag{4}$$

for which LARC(0) is clearly satisfied. For this system, (3) is satisfied with $\bar{n} = 3$—so that $\mathbb{T}^{\bar{n}-m} = \mathbb{T}$—and, for example, any function F_ϵ ($\epsilon > 0$) defined by

$$F_\epsilon(\theta) = \begin{pmatrix} \epsilon \sin\theta \\ \epsilon \cos\theta \\ \dfrac{\epsilon^2}{4} \sin 2\theta \end{pmatrix}.$$

Indeed, (3) is in this case equivalent to the condition

$$\forall \theta \in \mathbb{T}\,, \quad \mathrm{Det}\left(H(\theta) \triangleq \left[X_1(F_\epsilon(\theta)) \;\; X_2 \;\; -\frac{\partial F_\epsilon}{\partial \theta}(\theta)\right]\right) \neq 0, \tag{5}$$

the satisfaction of which is readily verified. Let us now introduce a new state vector φ defined by

$$\varphi(p, \theta) \triangleq \begin{pmatrix} p_1 - F_{\epsilon,1}(\theta) \\ p_2 - F_{\epsilon,2}(\theta) \\ p_3 - F_{\epsilon,3}(\theta) - p_1\,(p_2 - F_{\epsilon,2}(\theta)) \end{pmatrix}.$$

A direct calculation shows that for any function of time $\theta(.)$ the time derivative of φ along any solution to (4) satisfies

$$\dot{\varphi}(p, \theta) = C(p) H(\theta) (u_1, u_2, \dot{\theta})^T$$

with

$$C(p) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -p_1 & 1 \end{pmatrix}.$$

Since both matrices $C(p)$ and $H(\theta)$ are invertible for any $p \in \mathbb{R}^3$ and any $\theta \in \mathbb{T}$, it is straightforward, by considering $(u_1, u_2, \dot{\theta})$ as a new control vector, to globally asymptotically stabilize φ to zero. For instance, uniform exponential stabilization of $\varphi = 0$ is obtained by setting

$$(u_1, u_2, \dot{\theta})^T = -kH^{-1}(\theta)C^{-1}(p)\varphi(p, \theta), \qquad k > 0\,.$$

In terms of the state p, this yields a control law which globally stabilizes a neighborhood of the origin, the size of which can be made arbitrarily small by choosing ϵ as small as needed. Let us remark that, although it was not formalized in this way, this idea has been used implicitly in [3] for the problem of tracking a unicycle-type vehicle.

Based on this simple example, potential applications of Theorem 1 to various control problems are easily envisioned. Direct applications concern practical feedback stabilization of either systems without drift—as illustrated in the above example—or systems with a nonvanishing drift v.f. (see, e.g., [13], where potential application to nonholonomic motion planning is also briefly discussed). Other applications in the domain of nonlinear observer design or control of PDEs might also be considered.

This paper is organized as follows: Theorem 1 is proved[2] in section 2, and an example to illustrate the construction of transverse functions F is provided in section 3. Let us finally indicate that a presentation of Theorem 1 was accepted at the IEEE Conference on Decision and Control 2000 [12] in the form of a regular paper which did not contain the proof.

The following notation is used throughout the paper.

- δ_i^j denotes the Kronecker delta.
- $B_n(0, \delta)$ denotes the closed ball in $\mathbb{R}^n$ centered at zero and of radius δ.
- For $h \in \mathcal{C}^\infty(\mathbb{R}^n; \mathbb{R}^m)$ and $g \in \mathcal{C}^\infty(\mathbb{R}^n; \mathbb{R})$ with $g(x) \neq 0$ for $x \neq 0$, we write $h = o(g)$ when $|h(x)|/|g(x)| \longrightarrow 0$ as $x \longrightarrow 0$.
- $\mathbf{d}$ denotes the exterior derivative.

2. Proof of Theorem 1. By considering a system of local coordinates $x = (x_1, \dots, x_n)$ on M, which maps p_0 to $0 \in \mathbb{R}^n$, and a—globally defined—frame[3] $\{\frac{\partial}{\partial\theta_{m+1}}, \dots, \frac{\partial}{\partial\theta_{\bar{n}}}\}$ on $\mathbb{T}^{\bar{n}-m}$, Theorem 1 rewrites as follows.

COROLLARY 1. *Let $g_1, \dots, g_m$ denote smooth v.f. on $\mathbb{R}^n$ such that the accessibility distribution is of constant dimension in a neighborhood of the origin. Then the following properties are equivalent:*

1. *LARC(0): the system*

$$S: \qquad \dot{x} = \sum_{i=1}^{m} u_i g_i(x)$$

satisfies the Lie algebra rank condition at the origin.

[2]Note added in proof: A simpler proof has recently been obtained. More details are available from the authors.

[3]The dual basis—coframe—will be denoted $(\mathrm{d}\theta_{m+1}, \dots, \mathrm{d}\theta_{\bar{n}})$.

2. *TC(0): there exist* $\bar{n} \in \mathbb{N}$ *and a family of functions* $f_\epsilon \in \mathcal{C}^\infty(\mathbb{T}^{\bar{n}-m}; B_n(0,\epsilon))$ $(\epsilon > 0)$ *such that, for any* $\epsilon > 0$*, the following transversality condition holds:*

$$\forall \theta \in \mathbb{T}^{\bar{n}-m}, \quad \text{Rank} \left(g_1(f_\epsilon(\theta)) \quad \dots \quad g_m(f_\epsilon(\theta)) \quad \frac{\partial f_\epsilon}{\partial \theta_{m+1}}(\theta) \quad \dots \quad \frac{\partial f_\epsilon}{\partial \theta_{\bar{n}}}(\theta) \right) = n \,. \tag{6}$$

We now focus on the proof of this equivalent formulation of Theorem 1.

2.1. TC(0) $\Longrightarrow$ LARC(0). We assume that $LARC(0)$ is not satisfied and show that $TC(0)$ cannot be satisfied either. By assumption, the accessibility distribution is of constant dimension n_0 in a neighborhood of the origin. Therefore, if $n_0 < n$, the Frobenius theorem guarantees the existence of local coordinates $\phi(x)$ such that ϕ_n is constant along the trajectories of S, i.e., for some neighborhood $\mathcal{U}$ of the origin,

$$\forall x \in \mathcal{U},\ \forall i = 1, \dots, m, \qquad \frac{\partial \phi_n}{\partial x}(x) \neq 0, \quad \text{and} \quad \frac{\partial \phi_n}{\partial x}(x) g_i(x) = 0 \,. \tag{7}$$

Now assume that $TC(0)$ is satisfied, and choose any f_ϵ satisfying (6) and such that $B_n(0,\epsilon) \subset \mathcal{U}$. By the compactness of $\mathbb{T}^{\bar{n}-m}$, the smooth function $\theta \longmapsto \phi_n(f_\epsilon(\theta))$ from $\mathbb{T}^{\bar{n}-m}$ to $\mathbb{R}$ attains its maximum value for some $\bar{\theta}$, i.e.,

$$\forall i = m+1, \dots \bar{n}, \qquad \frac{\partial \phi_n}{\partial x}(f_\epsilon(\bar{\theta})) \frac{\partial f_\epsilon}{\partial \theta_i}(\bar{\theta}) = 0 \ . \tag{8}$$

From (8) and from (7) evaluated at $x = f_\epsilon(\bar{\theta})$, we obtain

$$\frac{\partial \phi_n}{\partial x}(f_\epsilon(\bar{\theta})) \left(g_1(f_\epsilon(\bar{\theta})) \quad \dots \quad g_m(f_\epsilon(\bar{\theta})) \quad \frac{\partial f_\epsilon}{\partial \theta_{m+1}}(\bar{\theta}) \quad \dots \quad \frac{\partial f_\epsilon}{\partial \theta_{\bar{n}}}(\bar{\theta}) \right) = 0 \ ,$$

which is in contradiction with $TC(0)$. □

2.2. LARC(0) $\Longrightarrow$ TC(0).

2.2.1. Notation and recalls. Prior to addressing the proof itself, we specify some notation and recall a few basic definitions and results that are extensively used in what follows. These recalls are about homogeneity on one hand and free Lie algebras on the other hand. For a more complete survey about these issues, we refer the reader to [5, 6] for the properties associated with homogeneity, and to [7, 17] for the role of free Lie algebras in control theory.

About homogeneity. Given $\mu > 0$ and a *weight vector* $r = (r_1, \dots, r_n)$ $(r_i > 0\ \forall i)$, a *dilation* Δ^r_μ on $\mathbb{R}^n$ is a map from $\mathbb{R}^n$ to $\mathbb{R}^n$ defined by $\forall z = (z_1, \dots, z_n) \in \mathbb{R}^n$, $\Delta^r_\mu z \triangleq (\mu^{r_1} z_1, \dots, \mu^{r_n} z_n)$. A function $f \in \mathcal{C}^0(\mathbb{R}^n; \mathbb{R})$ is *homogeneous of degree* l *with respect to the family of dilations* $(\Delta^r_\mu)_{\mu>0}$ or, more concisely, Δ^r*-homogeneous of degree* l if $\forall \mu > 0$, $f(\Delta^r_\mu z) = \mu^l f(z)$. A Δ^r*-homogeneous norm* is defined as a positive definite function on $\mathbb{R}^n$, Δ^r-homogeneous of degree one. A smooth v.f. X on $\mathbb{R}^n$ is Δ^r*-homogeneous of degree* d if, $\forall i = 1, \dots, n$, the function $x \longmapsto X_i(x)$ is Δ^r-homogeneous of degree $d + r_i$. The system

$$S_{ap}: \qquad \dot{z} = \sum_{i=1}^{m} b_i(z) u_i \tag{9}$$

is a Δ^r-*homogeneous approximation* of S if there exists a change of coordinates $\phi: x \longmapsto z$ which transforms S into

$$\dot{z} = \sum_{i=1}^{m} (b_i(z) + h_i(z))\, u_i\,, \tag{10}$$

where b_i is Δ^r-homogeneous of degree -1, and h_i denotes higher-order terms; i.e., for any j, the jth component $h_{i,j}$ of h_i satisfies $h_{i,j} = o(\rho^{r_j - 1})$, where ρ is any Δ^r-homogeneous norm.

The main motivation for introducing such approximations comes from the following result.

PROPOSITION 1 (see [5, 15]). *For any system S of smooth v.f. which satisfies LARC(0), there exists a Δ^r-homogeneous approximation S_{ap} which also satisfies LARC(0).*

Finally, we say that a set $\{b_1, \dots, b_m\}$ of v.f., or the associated system (9), is *nilpotent of order* $d+1$ if any Lie bracket of these v.f. of length larger than, or equal to, $d+1$ is identically zero. It is simple to verify that any set $\{b_1, \dots, b_m\}$ of smooth v.f. with the b_i's Δ^r-homogeneous of degree -1 is nilpotent of order $1 + \mathrm{Max}\{r_i : i = 1, \dots, n\}$.

About free Lie algebras. Let us consider a finite set of indeterminates $X_1, \dots, X_m$ and denote by $\mathrm{Lie}(X)$ the free Lie algebra over $\mathbb{R}$ generated by the X_i's. We also denote by $\mathcal{F}(X)$ the set of formal brackets in the X_i's. For any set $\mathbf{b} \stackrel{\Delta}{=} \{b_1, \dots, b_m\}$ of smooth v.f. and any $B \in \mathcal{F}(X)$, we denote by $\mathrm{Ev}_{\mathbf{b}}(B)$ the *evaluation map*, i.e., $\mathrm{Ev}_{\mathbf{b}}(X_i) = b_i$, and

$$\mathrm{Ev}_{\mathbf{b}}([B_\lambda, B_\rho]) = [\mathrm{Ev}_{\mathbf{b}}(B_\lambda), \mathrm{Ev}_{\mathbf{b}}(B_\rho)]\,.$$

The definition of a (generalized) P. Hall basis of $\mathrm{Lie}(X)$ is recalled below.

DEFINITION 1. *A P. Hall basis $\mathcal{B}$ of Lie(X) is a totally ordered subset of $\mathcal{F}(X)$ such that*

1. *each X_i belongs to $\mathcal{B}$;*
2. *if $B = [B_\lambda, B_\rho] \in \mathcal{F}$ with $B_\lambda, B_\rho \in \mathcal{F}$, then $B \in \mathcal{B}$ if and only if $B_\lambda, B_\rho \in \mathcal{B}$ with $B_\lambda < B_\rho$, and either* (i) *B_ρ is one of the X_i's or* (ii) *$B_\rho = [B_{\lambda\rho}, B_{\rho^2}]$ with $B_{\lambda\rho} \leq B_\lambda$;*
3. *if $B \in \mathcal{B}$ is a bracket of length $\ell(B) \geq 2$, i.e., $B = [B_\lambda, B_\rho]$, with $B_\lambda, B_\rho \in \mathcal{B}$, then $B_\lambda < B$.*

In order to simplify the forthcoming analysis we choose a specific P. Hall basis $\mathcal{B}$ associated with a specific total order. The P. Hall basis so obtained is in fact a Hall basis in the original (narrow) sense (see, e.g., [14, Section IV.5]).

Specific order.

$$\begin{cases} \ell(B) < \ell(B') \Longrightarrow B < B', \\ X_i < X_j \Longleftrightarrow i < j, \\ \text{For } \ell(B) = \ell(B') > 1, B < B' \Longleftrightarrow B_\lambda < B'_\lambda, \text{ or } B_\lambda = B'_\lambda \text{ and } B_\rho < B'_\rho\,. \end{cases} \tag{11}$$

We denote by

$$\mathcal{B} = \{B_1, B_2, \dots, B_q, \dots\},\ B_1 < B_2 < \cdots < B_q < \cdots\,, \tag{12}$$

the P. Hall basis associated with the total order (11), and also by $\ell(i)$ the length of any bracket B_i of this basis. From (11) and the definition of a P. Hall basis, we deduce the following properties which will be extensively used in what follows:

$$i \in \{1, \dots, m\} \Longleftrightarrow \ell(i) = 1 \Longleftrightarrow B_i = X_i, \tag{13}$$

$$i > m \Longleftrightarrow \ell(i) > 1 \Longleftrightarrow B_i = [B_{\lambda(i)}, B_{\rho(i)}], \tag{14}$$

where $\lambda(i)$ and $\rho(i)$ are uniquely defined integers. By extension of this notation, and whenever this will make sense, we will use the symbols $\lambda^2(i), \lambda\rho(i), \rho^2(i), \dots$, to index the elements of $\mathcal{B}$. For instance, if $\ell(\rho(i)) \geq 2$, we can write $B_{\rho(i)} = [B_{\lambda\rho(i)}, B_{\rho^2(i)}]$. Finally, it also follows from (11) and the definition of a P. Hall basis that

$$\ell(i) > 1 \Longrightarrow \lambda(i) < \rho(i) < i\,.$$

Letting $0 < d \in \mathbb{N}$, we denote by $\mathrm{Lie}_d(X)$ the subspace of $\mathrm{Lie}(X)$ generated by brackets of length at most equal to d. Then the subset of $\mathcal{B}$ composed of all brackets B_j such that $\ell(j) \leq d$ is a basis of $\mathrm{Lie}_d(X)$ denoted as $\mathcal{B}_d$. Let $n(d)$ denote the dimension of $\mathrm{Lie}_d(X)$ so that

$$\mathcal{B}_d = \{B_1, \dots, B_{n(d)}\} \quad \text{and } \ell(n(d)) = d\,.$$

One can associate the following *free system* with $\mathcal{B}_d$:

$$\left\{ \begin{array}{rcll} \dot{x}_i & = & u_i, & i = 1, \dots, m, \\ \dot{x}_i & = & x_{\lambda(i)} \dot{x}_{\rho(i)}, & i = m+1, \dots, n(d). \end{array} \right. \tag{15}$$

REMARK 2. *Since there is a one-to-one correspondence between the components of the state vector x associated with the free system* (15) *and the element of $\mathcal{B}_d$, it would be natural to index each component of x by the corresponding element of $\mathcal{B}_d$, as done, for example, in* [7]. *We have preferred here to write B_i for an element of $\mathcal{B}_d$ and x_i for the corresponding component of x in order to lighten the notation.*

It is straightforward to verify that (15) defines a control-affine driftless system:

$$S(m,d): \qquad \dot{x} = \sum_{i=1}^{m} u_i b_i(x), \tag{16}$$

where the components $b_{i,j}$ of the v.f. b_i are defined by

$$b_{i,j}(x) = \left\{ \begin{array}{ll} \delta_i^j & \text{if } \ell(j) = 1, \\ x_{\lambda(j)} b_{i,\rho(j)} & \text{otherwise.} \end{array} \right. \tag{17}$$

The following properties of free systems will be used in what follows. For the first two properties, we refer to [7]. The third property has been proved in [8, Section 3] in a formal algebraic framework. A proof of the fourth property is given in the appendix.

LEMMA 1. *For $i = m+1, \dots, n(d)$, let b_i denote the v.f. $Ev_{\mathbf{b}}(B_i)$, where $\mathbf{b} = \{b_1, \dots, b_m\}$. Then the following properties hold.*

1. *For any $i \in \{1, \dots, n(d)\}$, $b_i = a_i \partial/\partial x_i + \sum_{j>i} b_{i,j} \partial/\partial x_j$ for some nonzero constant a_i and some smooth functions $b_{i,j}$ so that $S(m,d)$ satisfies $LARC(x)$ for any $x \in \mathbb{R}^{n(d)}$.*

2. *The v.f. b_i are Δ-homogeneous of degree $-\ell(i)$ with Δ_μ ($\mu > 0$), the dilation defined by*

$$\Delta_\mu x = (\mu^{\ell(1)} x_1, \dots, \mu^{\ell(n(d))} x_{n(d)}) \tag{18}$$

so that $S(m,d)$ is nilpotent of order $d+1$.

3. *For any $p \in \mathcal{C}^\infty(\mathbb{R}^{n(d)};\mathbb{R})$, Δ-homogeneous of degree $d' < d$, and any $j \in \{1,\dots,m\}$, there exists $q^j \in \mathcal{C}^\infty(\mathbb{R}^{n(d)};\mathbb{R})$, Δ-homogeneous of degree $d'+1$, such that*

$$\forall i \in \{1,\dots,m\}, \qquad L_{b_i} q^j = \begin{cases} p & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{19}$$

4. *For any $i \in \{1,\dots,n(d)\}$ and any $p \in \mathcal{C}^\infty(\mathbb{R}^{n(d)};\mathbb{R})$, Δ-homogeneous of degree $d' - \ell(i)$ with $d' \le d$, there exist h_1 and $h_{2,j}$ ($1 < \ell(j) \le d'$) in $\mathcal{C}^\infty(\mathbb{R}^{n(d)};\mathbb{R})$, Δ-homogeneous of degree d' and $d' - \ell(j)$, respectively, such that*

$$p(x)\mathrm{d}x_i = \mathbf{d}h_1 + \sum_{j:1<\ell(j)\le d'} h_{2,j}(x)\left(\mathrm{d}x_j - x_{\lambda(j)}\mathrm{d}x_{\rho(j)}\right). \tag{20}$$

Remark 3.

1. *The functions p, q^j, h_1, and $h_{2,j}$ in properties 3 and 4 are polynomial in x because they are smooth and homogeneous.*
2. *Since the smooth functions q^j in property 3 are homogeneous of degree $d'+1$, it can depend only on the $n(d'+1)$ first components of x.*

After these preliminary recalls, we can now proceed with the proof of Theorem 1. It is composed of three steps which are summarized in the following three propositions.

Proposition 2. *If $TC(0)$ holds for a homogeneous approximation S_{ap} of a system S, then $TC(0)$ holds for S also.*

Proposition 3. *If, for any $d \in \mathbb{N}-\{0\}$, $TC(0)$ holds for the free system $S(m,d)$ with $\bar{n} = n(d)$, then $TC(0)$ holds for any smooth driftless system S_{hom} which satisfies $LARC(0)$ and whose control v.f. are Δ^r-homogeneous of degree -1 for some dilation Δ^r_μ.*

Proposition 4. *For any $d \in \mathbb{N}-\{0\}$, $TC(0)$ holds for the free system $S(m,d)$ with $\bar{n} = n(d)$.*

From Proposition 1, if S satisfies $LARC(0)$, it has a homogeneous approximation which also satisfies $LARC(0)$. This property, combined with the three propositions above, clearly implies that $LARC(0) \Longrightarrow TC(0)$. There remains to prove these three propositions.

2.2.2. Proof of Proposition 2. S rewrites, in some coordinates $z = \phi(x)$, as

$$\dot{z} = \sum_{i=1}^{m} u_i \left(\tilde{b}_i(z) + h_i(z)\right), \tag{21}$$

where the $\tilde{b}_i$'s, Δ^r-homogeneous of degree -1 (for some dilation Δ^r), are the v.f. of the homogeneous approximation S_{ap}, and h_i denotes higher-order terms, i.e.,

$$h_{i,j} = o(\rho^{r_j - 1}), \tag{22}$$

with ρ denoting any Δ^r-homogeneous norm. We want to show that if $TC(0)$ holds for S_{ap}, then it also holds for S. Since $TC(0)$ is independent of the system of coordinates,

it is sufficient to show that $TC(0)$ holds in the coordinates z. Let $\bar{n}$ and $(f_\epsilon)_{\epsilon>0}$ denote an integer and a family of functions which satisfy (6) with the v.f. of the approximation S_{ap}. We show below that S satisfies $TC(0)$ by considering the same integer $\bar{n}$ and the family of functions $(\bar{f}_\epsilon)_{\epsilon>0}$ defined by

$$\bar{f}_\epsilon(\theta) = \Delta^r_{\mu(\epsilon)} f_1(\theta)\,, \tag{23}$$

with $\mu(\epsilon)$ denoting a strictly positive number which is (i) smaller than some adequately chosen $\mu_0 > 0$ and (ii) such that $\sup_{\theta\in\mathbb{T}^{\bar{n}-m}} |\Delta^r_{\mu(\epsilon)} f_1(\theta)| \le \epsilon$. Note that $\mu(\epsilon)$ always exists because $f_1(\mathbb{T}^{\bar{n}-m})$ is a compact set so that $\lim_{\mu\to 0} \sup_{\theta\in\mathbb{T}^{\bar{n}-m}} |\Delta^r_\mu f_1(\theta)| = 0$.

With z denoting a vector in $\mathbb{R}^n$, one deduces from (22) that

$$\lim_{\mu\to 0} \frac{h_{i,j}(\Delta^r_\mu z)}{\rho^{r_j-1}(\Delta^r_\mu z)} = \lim_{\mu\to 0} \frac{h_{i,j}(\Delta^r_\mu z)}{\mu^{r_j-1}\rho^{r_j-1}(z)} = 0.$$

Therefore,

$$h_{i,j}(\Delta^r_\mu z) = c_{i,j}(\mu, z)\mu^{r_j-1},$$

where $|c_{i,j}(\mu, z)|$ tends to zero as μ tends to zero. Moreover, the convergence is uniform with respect to the z variable when $z \in B_n(0, 1)$. The above equation can also be written in vectorial form as

$$h_i(\Delta^r_\mu z) = \mu^{-1}\Delta^r_\mu c_i(\mu, z) \tag{24}$$

with $c_i = (c_{i,1}, \ldots, c_{i,n})^T$.

Let us now evaluate the rank of the matrix

$$A(\epsilon,\theta) \triangleq \left((\tilde{b}_1 + h_1)(\bar{f}_\epsilon(\theta)) \quad \ldots \quad (\tilde{b}_m + h_m)(\bar{f}_\epsilon(\theta)) \quad \frac{\partial \bar{f}_\epsilon}{\partial\theta_{m+1}}(\theta) \quad \ldots \quad \frac{\partial \bar{f}_\epsilon}{\partial\theta_{\bar{n}}}(\theta) \right).$$

Using (23), (24), and the fact that each $\tilde{b}_i$ is homogeneous of degree -1,

$$A(\epsilon,\theta) \;=\; \bar{A}(\epsilon,\theta)\; D(\mu(\epsilon))$$

with

$$\begin{aligned}
&\bar{A}(\epsilon,\theta)\\
&\quad\triangleq \left(\Delta^r_{\mu(\epsilon)}\tilde{b}_1(f_1(\theta)) \quad \ldots \quad \Delta^r_{\mu(\epsilon)}\tilde{b}_m(f_1(\theta)) \quad \Delta^r_{\mu(\epsilon)}\frac{\partial f_1}{\partial\theta_{m+1}}(\theta) \quad \ldots \quad \Delta^r_{\mu(\epsilon)}\frac{\partial f_1}{\partial\theta_{\bar{n}}}(\theta) \right)\\
&\quad+ \left(\Delta^r_{\mu(\epsilon)}c_1(\mu(\epsilon), f_1(\theta)) \quad \ldots \quad \Delta^r_{\mu(\epsilon)}c_m(\mu(\epsilon), f_1(\theta)) \quad 0 \quad \ldots \quad 0 \right),
\end{aligned}$$

and

$$D(\mu(\epsilon)) \triangleq \mathrm{diag}\{1/\mu(\epsilon), \ldots, 1/\mu(\epsilon), 1, \ldots, 1\}\,.$$

Since $D(\mu(\epsilon))$ is nonsingular, it readily follows that

$$\begin{aligned}
\mathrm{Rank}\, A(\epsilon,\theta) = \mathrm{Rank} \Big(&\tilde{b}_1(f_1(\theta)) + c_1(\mu(\epsilon), f_1(\theta)) \quad \ldots \quad \tilde{b}_m(f_1(\theta)) + c_m(\mu(\epsilon), f_1(\theta))\\
&\frac{\partial f_1}{\partial\theta_{m+1}}(\theta) \quad \ldots \quad \frac{\partial f_1}{\partial\theta_{\bar{n}}}(\theta) \Big).
\end{aligned} \tag{25}$$

Now, by assumption,

(26)
$$\forall \theta \in \mathbb{T}^{\bar{n}-m}, \qquad \text{Rank} \left(\tilde{b}_1(f_1(\theta)) \quad \dots \quad \tilde{b}_m(f_1(\theta)) \quad \frac{\partial f_1}{\partial \theta_{m+1}}(\theta) \quad \dots \quad \frac{\partial f_1}{\partial \theta_{\bar{n}}}(\theta) \right) = n\,.$$

In view of (25) and (26) and using the facts that $f_1(\theta) \in B_n(0,1)$ and that $|c_{i,j}(\mu, z)|$ tends uniformly (with respect to $z \in B_n(0,1)$) to zero as μ tends to zero, there exists a strictly positive number μ_0 such that

$$\mu(\epsilon) \leq \mu_0 \Longrightarrow \forall \theta \in \mathbb{T}^{\bar{n}-m}\,,\ \text{Rank}\ A(\epsilon, \theta) = n.$$

This concludes the proof of Proposition 2. □

REMARK 4. *The previous analysis implies—by setting* $\forall i,\ h_i \equiv 0$ *in* (21)*—that for a homogeneous system, if a function* $f \in \mathcal{C}^\infty(\mathbb{T}^{\bar{n}-m}; \mathbb{R}^n)$ *satisfies* (6)*, then, for any* $\mu > 0$, $\Delta_\mu f$ *also satisfies* (6)*. Therefore,* $TC(0)$ *is satisfied for this homogeneous system with the functions* $f_\epsilon \stackrel{\Delta}{=} \Delta_{\mu(\epsilon)} f$*, where* $\mu(\epsilon)$ *is any strictly positive value such that* $\sup_{\theta \in \mathbb{T}^{\bar{n}-m}} |\Delta_{\mu(\epsilon)} f(\theta)| \leq \epsilon$.

2.2.3. Proof of Proposition 3. Consider a smooth driftless system

$$S_{hom}: \qquad \dot{z} = \sum_{i=1}^{m} \tilde{b}_i(z) u_i, \tag{27}$$

whose v.f. $\tilde{b}_i$ $(i = 1, \dots, m)$ are Δ^r-homogeneous of degree -1 for some dilation Δ^r_μ and satisfy $LARC(0)$. Since S_{hom} is nilpotent of some order $d+1$, it can be associated with the free system $S(m,d)$ whose v.f. b_i are defined in (17). We show below that any family $(f_\epsilon)_{\epsilon>0}$ which satisfies $TC(0)$ for the free system $S(m,d)$ induces a family $(\tilde{f}_\epsilon)_{\epsilon>0}$ which satisfies $TC(0)$ for S_{hom}. In fact, from Remark 4 above, we need only to show the existence of a single function $\tilde{f} \in \mathcal{C}^\infty(\mathbb{T}^{n(d)-m}; \mathbb{R}^n)$, which satisfies the transversality condition (6) for S_{hom}.

Let f denote any of the functions f_ϵ which satisfy the transversality condition for $S(m,d)$. From property 1 of Lemma 1, the vectors $b_1(x), \dots, b_{n(d)}(x)$ are linearly independent at any $x \in \mathbb{R}^{n(d)}$. Therefore, there exist (unique) smooth functions $u_{i,j}$ such that

$$\forall j = m+1, \dots, n(d),\ \forall \theta \in \mathbb{T}^{n(d)-m}\,, \qquad \frac{\partial f}{\partial \theta_j}(\theta) = \sum_{i=1}^{n(d)} u_{i,j}(\theta) b_i(f(\theta)). \tag{28}$$

Also, using the fact that f satisfies the transversality condition (6) for $S(m,d)$,

$$\forall \theta \in \mathbb{T}^{n(d)-m}, \qquad \text{Det}\, U(\theta) \neq 0 \quad \text{with} \quad U(\theta) \stackrel{\Delta}{=} \big(u_{i,j}(\theta)\big)_{i,j=m+1,\dots,n(d)}. \tag{29}$$

Let us now define the function $\tilde{f}$. To this purpose, let us pick an arbitrary couple $(\theta_0, z_0) \in (\mathbb{T}^{n(d)-m} \times \mathbb{R}^n)$ and consider an element θ of $\mathbb{T}^{n(d)-m}$. Consider also a smooth path $\gamma : t \in [0,1] \longrightarrow \gamma(t) \in \mathbb{T}^{n(d)-m}$ which connects θ_0 to θ, i.e., such that $\gamma(0) = \theta_0$ and $\gamma(1) = \theta$. Let $z_\gamma(t)$ denote the solution, for $t \in [0,1]$, of

$$\dot{z} = \sum_{i=1}^{n(d)} \bar{U}_i(\gamma(t), \dot{\gamma}(t))\ \tilde{b}_i(z), \qquad z(0) = z_0\,, \tag{30}$$

where

$$\bar{U}_i(\gamma,\dot{\gamma}) = \sum_{j=m+1}^{n(d)} u_{i,j}(\gamma)\mathrm{d}\theta_j(\dot{\gamma})\,, \tag{31}$$

and, for $i = m+1,\dots,n(d)$, $\tilde{b}_i \triangleq Ev_{\tilde{\mathbf{b}}}(B_i)$. Note that $z_\gamma(t)$ is well defined for $t \in [0,1]$. Indeed, finite-time escape is not possible because the v.f. $\tilde{b}_i$ are homogeneous of negative degree (by assumption). Let us show that $z_\gamma(1)$ is independent of the path γ chosen to connect θ_0 to θ. To this purpose, consider two paths γ_i $(i=1,2)$ which map 0 to θ_0 and 1 to θ. We must show that the solution $z_{\gamma_1}(1)$ of (30) at $t=1$ with $\gamma=\gamma_1$ is the same as the solution $z_{\gamma_2}(1)$ of (30) at $t=1$ with $\gamma=\gamma_2$. To show this, we will use the properties stated in the following lemma, which are easily derived from well-known results. (See the appendix for details.)

LEMMA 2. *Consider the P. Hall basis $\mathcal{B}$ of Lie$(X_1,\dots,X_m)$ defined by* (12). *Then there exist mappings $(T,u) \longmapsto c_i(T,u)$ such that, for any set $\mathbf{g} = \{g_1,\dots,g_m\}$ of v.f. nilpotent of order $d+1$, and any $u \in \mathcal{C}^\infty([0,T];\mathbb{R}^{n(d)})$, the solution at time T of*

$$\dot{x} = \sum_{i=1}^{n(d)} u_i(t)\, g_i(x)\,, \qquad x(0) = x_0, \tag{32}$$

is

$$x(T) = \exp\left(\sum_{i=1}^{n(d)} c_i(T,u)\, g_i\right) x_0, \tag{33}$$

where $g_i \triangleq Ev_{\mathbf{g}}(B_i)$ $(i = m+1,\dots,n(d))$. Furthermore, if $g_1,\dots,g_m$ are the control v.f. of the $(n(d)$-dimensional) free system $S(m,d)$, then for any $x_0 \in \mathbb{R}^{n(d)}$ the mapping

$$(c_1,\dots,c_{n(d)}) \longmapsto \exp\left(\sum_{i=1}^{n(d)} c_i\, g_i\right) x_0 \tag{34}$$

from $\mathbb{R}^{n(d)}$ to $\mathbb{R}^{n(d)}$ is one-to-one.

Applying the first result stated in the lemma to (30) yields

$$\forall k = 1,2, \qquad z_{\gamma_k}(1) = \exp\left(\sum_{i=1}^{n(d)} c_i\left(1, \bar{U}(\gamma_k,\dot{\gamma}_k)\right)\, \tilde{b}_i\right) z_0\,. \tag{35}$$

Consider now the following equation parameterized by $k=1,2$ (compare with (30)):

$$\dot{x} = \sum_{i=1}^{n(d)} \bar{U}_i(\gamma_k(t),\dot{\gamma}_k(t))\; b_i(x), \qquad x(0) = f(\theta_0)\,. \tag{36}$$

From (28) and (31), $f(\gamma_k(.))$ is a solution to (36). Therefore, applying the first result stated in the lemma to this equation and using the fact that $f(\theta) = f(\gamma_k(1))$ for $k=1,2$ yields

$$\exp\left(\sum_{i=1}^{n(d)} c_i\left(1, \bar{U}(\gamma_1,\dot{\gamma}_1)\right)\; b_i\right) f(\theta_0) = \exp\left(\sum_{i=1}^{n(d)} c_i\left(1, \bar{U}(\gamma_2,\dot{\gamma}_2)\right)\; b_i\right) f(\theta_0)\,.$$

The second result stated in the lemma then implies that

$$\forall i = 1, \dots, n(d), \qquad c_i\left(1, \bar{U}(\gamma_1, \dot{\gamma}_1)\right) = c_i\left(1, \bar{U}(\gamma_2, \dot{\gamma}_2)\right), \tag{37}$$

and it follows, in view of (35), that $z_{\gamma_1}(1) = z_{\gamma_2}(1)$. This in turn establishes that the mapping $(\theta, \gamma) \to z_\gamma(1)$ is a function of θ solely. This is the function $\tilde{f}$ which we were looking for. At this point, it remains only to verify that the function $\tilde{f}$ so defined satisfies the transversality condition (6) for S_{hom}. Recalling that $\tilde{f}(\theta)$ is obtained as the solution of (30) at $t = 1$ and that this solution does not depend on the path γ which passes thru θ at time $t = 1$, one deduces that along any smooth curve $\theta(.)$ the mapping $t \longmapsto \tilde{f}(\theta(t))$ is differentiable with

$$\frac{d}{dt}\tilde{f}(\theta(t)) = \sum_{i=1}^{n(d)} \bar{U}_i(\theta(t), \dot{\theta}(t))\, \tilde{b}_i(\tilde{f}(\theta(t)))\,.$$

This in turn implies that $\tilde{f}$ is smooth and satisfies

$$\forall \theta \in \mathbb{T}^{n(d)-m}, \quad \frac{\partial \tilde{f}}{\partial \theta_j}(\theta) = \sum_{i=1}^{n(d)} u_{i,j}(\theta)\, \tilde{b}_i(\tilde{f}(\theta))\,. \tag{38}$$

This implies that

$$\left(\tilde{b}_1(\tilde{f}(\theta)), \dots, \tilde{b}_m(\tilde{f}(\theta)), \frac{\partial \tilde{f}}{\partial \theta_{m+1}}(\theta), \dots, \frac{\partial \tilde{f}}{\partial \theta_{n(d)}}(\theta)\right) = \left(\tilde{b}_1(\tilde{f}(\theta)), \dots, \tilde{b}_{n(d)}(\tilde{f}(\theta))\right) \begin{pmatrix} I_m & \star \\ 0 & U(\theta) \end{pmatrix},$$

where $I_m \in \mathbb{R}^{m\times m}$ is the identity matrix. Using (29) and the fact that S_{hom} satisfies $LARC(x)$ for $x \in \mathbb{R}^n$—indeed, it satisfies $LARC(0)$ so that, by continuity it satisfies $LARC(x)$ in a neighborhood of the origin and therefore, by homogeneity, in $\mathbb{R}^n$ itself—one easily deduces from the above equality that $\tilde{f}$ satisfies the transversality condition (6) for S_{hom}. □

2.2.4. Proof of Proposition 4. From Remark 4 and property 2 of Lemma 1, it is sufficient to prove the existence of a single function $f \in \mathcal{C}^\infty(\mathbb{T}^{n(d)-m}; \mathbb{R}^{n(d)})$ for which the transversality condition (6) is satisfied. In order to simplify some of the forthcoming analysis, we will use the formalism of differential forms, from which condition (6) can be written as

$$\forall \theta \in \mathbb{T}^{n(d)-m}, \quad (\mathrm{d}x_1 \wedge \cdots \wedge \mathrm{d}x_{n(d)}) \left(b_1, \dots, b_m, \frac{\partial f}{\partial \theta_{m+1}}, \dots, \frac{\partial f}{\partial \theta_{n(d)}}\right)_{|x=f(\theta)} \neq 0.$$

By skew-symmetry of the wedge product, this is equivalent to the condition that

$$\forall \theta \in \mathbb{T}^{n(d)-m}, \tag{39}$$
$$(\mathrm{d}x_1 \wedge \cdots \wedge \mathrm{d}x_m \wedge \omega^x_{m+1} \wedge \cdots \wedge \omega^x_{n(d)}) \left(b_1, \dots, b_m, \frac{\partial f}{\partial \theta_{m+1}}, \dots, \frac{\partial f}{\partial \theta_{n(d)}}\right)_{|x=f(\theta)} \neq 0,$$

where $\omega_i^x = \mathrm{d}x_i - x_{\lambda(i)}\mathrm{d}x_{\rho(i)}$ $(i = m+1, \dots, n(d))$. From (17),

$$\forall j = 1, \dots, m \quad \left\{ \begin{array}{rcll} \mathrm{d}x_i(b_j) & = & \delta_i^j & \text{if } i \in \{1, \dots, m\}, \\ \omega_i^x(b_j) & = & 0 & \text{if } i \in \{m+1, \dots, n(d)\} \end{array} \right.$$

so that one easily rewrites (39) as

$$\forall \theta \in \mathbb{T}^{n(d)-m}, \quad \left(\omega_{m+1} \wedge \cdots \wedge \omega_{n(d)}\right)(\theta) \neq 0, \tag{40}$$

with ω_i the differential one-form on $\mathbb{T}^{n(d)-m}$ defined by

$$\omega_i = \mathbf{d}f_i - f_{\lambda(i)}\mathbf{d}f_{\rho(i)}\,. \tag{41}$$

Design algorithm. The function f is defined by setting $f \triangleq f^{n(d)}$, with the function $f^{n(d)}$ denoting the last function obtained via a recursive construction which starts with some function f^{m+1}. For each $k = m+1, \dots, n(d)$, the function $f^k \in \mathcal{C}^\infty(\mathbb{T}^{k-m}; \mathbb{R}^{n(d)})$ is required to verify the following property:

$$\forall \theta^k = (\theta_{m+1}, \dots, \theta_k) \in \mathbb{T}^{k-m}, \quad \left(\omega_{m+1}^k \wedge \cdots \wedge \omega_k^k\right)(\theta^k) \neq 0\,, \tag{42}$$

with ω_i^k the differential one-form on $\mathbb{T}^{k-m}$ defined by

$$\omega_i^k = \mathbf{d}f_i^k - f_{\lambda(i)}^k \mathbf{d}f_{\rho(i)}^k\,. \tag{43}$$

$\boldsymbol{f^{m+1}}$. A possible choice for f^{m+1} is as follows:

$$f_i^{m+1}(\theta_{m+1}) = \left\{ \begin{array}{ll} \sin\theta_{m+1} & \text{for } i = \lambda(m+1), \\ \cos\theta_{m+1} & \text{for } i = \rho(m+1), \\ \dfrac{1}{4}\sin 2\theta_{m+1} & \text{for } i = m+1, \\ 0 & \text{otherwise}\,. \end{array} \right. \tag{44}$$

Indeed, it readily follows from this definition that

$$\forall \theta^{m+1} \in \mathbb{T}, \quad \omega_{m+1}^{m+1}(\theta^{m+1}) = \frac{1}{2}\,.$$

$\boldsymbol{f^{k-1} \longrightarrow f^k}$. Assume now that, for some $k-1 \in \{m+1, \dots, n(d)-1\}$, a function $f^{k-1} \in \mathcal{C}^\infty(\mathbb{T}^{k-1-m}; \mathbb{R}^{n(d)})$ which verifies the property (42) for $k-1$ has been obtained. We show below how to construct from this function a new function $f^k \in \mathcal{C}^\infty(\mathbb{T}^{k-m}; \mathbb{R}^{n(d)})$ which verifies the property (42).

Let Δ_μ^k $(\mu > 0)$ denote the dilation defined on $\mathbb{R} \times \mathbb{R} \times \mathbb{R}^{n(d)}$ by

$$\Delta_\mu^k(s, c, f) = \left(\mu^{\ell(\lambda(k))}s, \mu^{\ell(\rho(k))}c, \Delta_\mu(f)\right) \text{ with } \Delta_\mu(f) \triangleq \left(\mu^{\ell(1)}f_1, \dots, \mu^{\ell(n(d))}f_{n(d)}\right). \tag{45}$$

Denote also p_i^k $(i = 1, \dots, n(d))$ the functions defined on $\mathbb{R} \times \mathbb{R}$ by

$$p_i^k(s, c) = s\,\delta_i^{\lambda(k)} + c\,\delta_i^{\rho(k)} + \frac{m_k^k}{2} sc\,\delta_i^k \tag{46}$$

with

$$m_i^k = \begin{cases} 0 & \text{if } \ell(i) \leq \ell(\lambda(k)) \text{ or } \lambda(i) \neq \lambda(k), \\ 1 + m_{\rho(i)}^k & \text{otherwise}. \end{cases} \tag{47}$$

The next step consists in finding polynomial functions $q_{i,j}^k \in \mathcal{C}^\infty(\mathbb{R}^{n(d)}; \mathbb{R})$ for $i = 1, \ldots, n(d)$ and $j = 1, \ldots, j_{i,k} \triangleq \max\{j : \ell(i) - j\ell(\lambda(k)) \geq 0\}$ such that the two following properties are verified.

P1(*i*) (for $i = 1, \ldots, n(d)$). Each function $q_{i,j}^k$ is Δ-homogeneous of degree $\ell(i) - j\ell(\lambda(k))$.

P2(*i*) (for $i = m+1, \ldots, k$).

$$\bar{\omega}_i^k = \left(\mathrm{d}f_i - f_{\lambda(i)}\mathrm{d}f_{\rho(i)} + \bar{\gamma}_i^k\right) + \sum_{j=m+1}^{i-1} t_{i,j}(s,f)\left(\mathrm{d}f_j - f_{\lambda(j)}\mathrm{d}f_{\rho(j)} + \bar{\gamma}_j^k\right), \tag{48}$$

where

$$\forall i = m+1, \ldots, k, \qquad \bar{\omega}_i^k \triangleq \mathbf{d}\bar{f}_i^k - \bar{f}_{\lambda(i)}^k \mathbf{d}\bar{f}_{\rho(i)}^k, \tag{49}$$

$$\bar{f}_i^k : (s,c,f) \longmapsto f_i + p_i^k(s,c) + \sum_{j=1}^{j_{i,k}} s^j q_{i,j}^k(f), \tag{50}$$

the $t_{i,j}$'s are smooth functions, and $\bar{\gamma}_i^k$ is a differential one-form on $\mathbb{R} \times \mathbb{R} \times \mathbb{R}^{n(d)}$ such that

$$\bar{\gamma}_i^k = \bar{\gamma}_{i,1}^k \mathrm{d}s + \bar{\gamma}_{i,2}^k \mathrm{d}c$$

with $\bar{\gamma}_{i,1}$, $\bar{\gamma}_{i,2}$, Δ^k-homogeneous of degree $\ell(i) - \ell(\lambda(k))$ and $\ell(i) - \ell(\rho(k))$, respectively, and

(51)

$$\begin{cases} \bar{\gamma}_{i,1}^k \equiv 0 & \text{if } i < \lambda(k), \\ \bar{\gamma}_{i,1}^k \equiv 1 & \text{if } i = \lambda(k), \\ \bar{\gamma}_{i,1}^k(s,c,0) = 0 & \text{if } \lambda(k) < i < k, \\ \bar{\gamma}_{i,1}^k(s,c,0) = \dfrac{m_k^k}{2}\, c & \text{for } i = k, \end{cases} \qquad \begin{cases} \bar{\gamma}_{i,2}^k \equiv 0 & \text{if } i < \rho(k), \\ \bar{\gamma}_{i,2}^k \equiv 1 & \text{if } i = \rho(k), \\ \bar{\gamma}_{i,2}^k(s,c,0) = 0 & \text{if } \rho(k) < i < k, \\ \bar{\gamma}_{i,2}^k(s,c,0) = -\dfrac{m_k^k}{2}\, s & \text{for } i = k\,. \end{cases}$$

LEMMA 3. *There exist functions $q_{i,j}^k$, which are solutions to the problems* **P1(*i*)** *and* **P2(*i*)**. *In particular, one can always choose*

$$\begin{cases} q_{i,j}^k \equiv 0 & \text{if } i \in \{1, \ldots, \mathrm{Max}\{m, \lambda(k)\}\} \cup \{k+1, \ldots, n(d)\} \\ & \text{and} \quad j \in \{1, \ldots, j_{i,k}\}, \\ q_{i,1}^k \equiv 0 & \text{if } \mathrm{Max}\{m, \lambda(k)\} < i \leq k \text{ and} \quad \lambda(i) < \lambda(k), \\ q_{i,1}^k(f) = m_i^k f_{\rho(i)} & \text{if } \mathrm{Max}\{m, \lambda(k)\} < i \leq k \text{ and} \quad \lambda(i) = \lambda(k). \end{cases} \tag{52}$$

Once suitable functions $q_{i,j}^k$ are determined so that the functions $\bar{f}_i^k$ in (50) are also defined, we set

$$f^k \triangleq \bar{f}^k \circ \bar{g}_\eta^k \quad \text{with} \quad \bar{g}_\eta^k(\theta^k) \triangleq \left(\eta^{\ell(\lambda(k))} \sin\theta_k, \eta^{\ell(\rho(k))} \cos\theta_k, f^{k-1}(\theta^{k-1})\right). \tag{53}$$

LEMMA 4. *For η larger than some positive value η_0, (42) is satisfied with the function f^k defined by (53).*

Therefore, Proposition 4 is proved once Lemmas 3 and 4 are proved.

REMARK 5. *It is simple to verify that each function $\bar{f}_i^k$ in (53) is polynomial in its arguments and Δ^k-homogeneous of degree $\ell(i)$ with respect to the dilation defined by (45). The proof of the lemmas much relies on this property.*

Proof of Lemma 3. We distinguish three cases.

Case 1. $1 \leq i \leq \mathrm{Max}\{m, \lambda(k)\}$. We define $q_{i,j}^k$ according to (52) so that **P1(*i*)** is clearly verified for these values of i. If $i \leq m$, **P2(*i*)** is irrelevant. If $m+1 \leq i \leq \lambda(k)$, it readily follows from (46), (49), (50), and (52) that

$$\bar{\omega}_i^k = \mathrm{d}f_i - f_{\lambda(i)}\mathrm{d}f_{\rho(i)} + \bar{\gamma}_i^k, \tag{54}$$

where $\bar{\gamma}_i^k \equiv 0$ if $i < \lambda(k)$ and $\bar{\gamma}_i^k = \mathrm{d}s$ if $i = \lambda(k)$. Therefore, **P2(*i*)** is also verified.

Case 2. $\mathrm{Max}\{m, \lambda(k)\} < i \leq k$. We define $q_{i,1}^k$ according to (52), which is consistent with **P1(*i*)**. To define the other functions $q_{i,j}^k$, we consider a construction which is recursive in the index i. More precisely, let us assume that functions $q_{1,j}^k, \ldots, q_{i-1,j}^k$ have been defined so that **P1(1)**, ..., **P1(*i*-1)** and **P2(1)**, ..., **P2(*i*-1)** are verified. We show below how to obtain functions $q_{i,j}^k$ so that **P1(*i*)** and **P2(*i*)** are also verified.

We first note that

$$\lambda(i) < \rho(k)\,. \tag{55}$$

Assume, on the contrary, that $\lambda(i) \geq \rho(k)$. Then, from the definition of a P. Hall basis, $\lambda(i) < \rho(i)$. This implies that

$$\ell(i) = \ell(\lambda(i)) + \ell(\rho(i)) \geq 2\ell(\rho(k)) \geq \ell(k)\,.$$

If $\ell(i) > \ell(k)$, then $i > k$, and this contradicts the assumption. Otherwise, $\ell(i) = \ell(k)$, and we also get $i > k$ because of (11) and the fact that $\lambda(i) \geq \rho(k) > \lambda(k)$.

We introduce the following definitions for the sake of simplifying some aspects of the forthcoming analysis.

DEFINITION 2. *A differential one-form $r = r_s\mathrm{d}s + r_c\mathrm{d}c + \sum_{j=1}^{n(d)} r_j\mathrm{d}f_j$, with r_s, r_c, r_j homogeneous of degree $\ell(i) - \ell(\lambda(k))$, $\ell(i) - \ell(\rho(k))$, and $\ell(i) - \ell(j)$, respectively, is said to be of*

- *type* 1 *if $r_j \equiv 0$ for each j, and both r_s and r_c are identically zero at $f = 0$;*
- *type* 2 *if $r_c \equiv r_j \equiv 0$ for each j, and $r_s = as^\kappa$ with $a \in \mathbb{R}$ and $1 \leq \kappa \in \mathbb{N}$;*
- *type* 3 *if $r_s \equiv r_c \equiv 0$ and, for each j, $r_j(s, c, f)$ is in the form $r_j(s, c, f) = s^{2+\kappa_j} r_j'(f)$ with $\kappa_j \in \mathbb{N}$.*

An upper-left index i for a one-form will indicate its type, e.g., ${}^2r$ indicates that ${}^2r$ is of type 2.

Next, we develop $\bar{\omega}_i^k$ and examine the terms involved in this development. From (49) and (50), we have

$$\begin{aligned}\bar{\omega}_i^k = \;& \mathrm{d}f_i + \mathbf{d}p_i^k + \mathbf{d}\left(\sum_{j=1}^{j_{i,k}} s^j q_{i,j}^k\right) \\ & - \left(f_{\lambda(i)} + p_{\lambda(i)}^k + \sum_{j=1}^{j_{\lambda(i),k}} s^j q_{\lambda(i),j}^k\right)\left(\mathrm{d}f_{\rho(i)} + \mathbf{d}p_{\rho(i)}^k + \mathbf{d}\left(\sum_{j=1}^{j_{\rho(i),k}} s^j q_{\rho(i),j}^k\right)\right)\end{aligned} \tag{56}$$

and, by rearranging the terms in the right-hand side of this equality,

$$\bar{\omega}_i^k = \mathrm{d}f_i - f_{\lambda(i)}\mathrm{d}f_{\rho(i)} + \mathbf{d}\left(\sum_{j=2}^{j_{i,k}} s^j q_{i,j}^k\right) + \alpha_1 + \alpha_2 + \alpha_3 + {}^1r + {}^2r + {}^3r \tag{57}$$

with

$$\begin{cases} \alpha_1 \triangleq \mathbf{d}p_i^k - p_{\lambda(i)}^k \mathbf{d}p_{\rho(i)}^k, \\ \alpha_2 \triangleq s\mathbf{d}q_{i,1}^k - sq_{\lambda(i),1}^k \mathrm{d}f_{\rho(i)} - sf_{\lambda(i)}\mathbf{d}q_{\rho(i),1}^k - p_{\lambda(i)}^k \mathrm{d}f_{\rho(i)}, \\ \alpha_3 \triangleq -\mathbf{d}p_{\rho(i)}^k \sum_{j=2}^{j_{\lambda(i),k}} s^j q_{\lambda(i),j}^k \,. \end{cases} \tag{58}$$

In (57), ${}^1r$, ${}^2r$, and ${}^3r$ just correspond to terms which do not need to be specified further and are of type 1, 2, and 3, following Definition 2. In order to obtain (57), we have used the following two arguments: (i) each function $q_{j,1}^k$ $(j \leq i)$ vanishes at the origin—this follows from (52) if $\lambda(j) \leq \lambda(k)$; otherwise, $\lambda(j) > \lambda(k)$ so that $\ell(j) > \ell(\lambda(k))$, and this follows from the fact that $q_{j,1}^k$ is Δ^k-homogeneous of positive degree; (ii) from (55), $\lambda(i) < \rho(k)$ so that (46) implies that $p_{\lambda(i)}^k(s,c)$ is either s or zero. Note also that the homogeneity properties of the components of ${}^1r$, ${}^2r$, and ${}^3r$ follow directly from the homogeneity of $\bar{f}_i^k$ (see Remark 5).

Let us now focus our attention on the terms α_i which are specified in (58). We first note that

$$\alpha_3 \equiv 0 \,. \tag{59}$$

Indeed, assume on the contrary that α_3 is not the null function. Then, in view of (52), it is necessary that $\lambda(i) > \lambda(k)$. (Otherwise, $q_{\lambda(i),j}^k$, and thus α_3, would be equal to zero.) Since $\lambda(i) < \rho(i)$ (from the definition of a P. Hall basis), we also have $\rho(i) \geq \rho(k)$. (Otherwise, $p_{\rho(i)}^k$, and thus α_3, would be equal to zero.) This implies that $i > k$, which is in contradiction with the assumption.

We now consider the term α_2 in (58). We have

$$\lambda(i) < \lambda(k) \Longrightarrow \alpha_2 \equiv 0 \,. \tag{60}$$

This follows from (46) and (52) after noticing that either $\ell(\rho(i)) = 1$ so that $q_{\rho(i),1}^k \equiv 0$, or $\ell(\rho(i)) > 1$ and $\lambda\rho(i) \leq \lambda(i) < \lambda(k)$ (from the definition of a P. Hall basis), so that we still obtain $q_{\rho(i),1}^k \equiv 0$. Then

$$\lambda(i) = \lambda(k) \quad \text{with} \quad \left.\begin{array}{c} \ell(\rho(i)) = 1 \\ \text{or} \\ \lambda\rho(i) < \lambda(k) \end{array}\right\} \Longrightarrow \alpha_2 \equiv 0 \,. \tag{61}$$

Indeed, if the left-hand side of the above implication holds, then (46), (47), and (52) imply

$$\begin{aligned} \alpha_2 &= s\left(m_i^k \mathrm{d}f_{\rho(i)} - f_{\lambda(i)}\mathbf{d}q_{\rho(i),1}^k - \mathrm{d}f_{\rho(i)}\right) \\ &= s\left(m_i^k \mathrm{d}f_{\rho(i)} - \mathrm{d}f_{\rho(i)}\right) \\ &\equiv 0 \,. \end{aligned} \tag{62}$$

From the definition of a P. Hall basis, $\lambda\rho(i) \leq \lambda(i)$ so that the case where $\lambda(i) = \lambda(k)$ with $\lambda\rho(i) > \lambda(k)$ cannot happen. Therefore, if $\lambda(i) = \lambda(k)$, the last possible case is $\lambda\rho(i) = \lambda(k)$. We have

$$\left.\begin{array}{l} \lambda(i) = \lambda(k) \text{ and } \lambda\rho(i) = \lambda(k) \end{array}\right\} \Longrightarrow \alpha_2 = s\, m^k_{\rho(i)} \left(\mathrm{d}f_{\rho(i)} - f_{\lambda\rho(i)}\mathrm{d}f_{\rho^2(i)}\right). \tag{63}$$

Indeed, from (52),

$$\begin{aligned} \alpha_2 &= s\left(m^k_i \mathrm{d}f_{\rho(i)} - f_{\lambda\rho(i)}\mathrm{d}q^k_{\rho(i),1} - \mathrm{d}f_{\rho(i)}\right) \\ &= s\left(m^k_i \mathrm{d}f_{\rho(i)} - f_{\lambda\rho(i)} m^k_{\rho(i)}\mathrm{d}f_{\rho^2(i)} - \mathrm{d}f_{\rho(i)}\right), \end{aligned}$$

and (63) follows from (47). Concerning α_2, there remains only to examine the case where $\lambda(i) > \lambda(k)$. In this case $p^k_{\lambda(i)} \equiv 0$—since, by (55), $\lambda(i) < \rho(k)$—so that

$$\alpha_2 = s\left(\mathbf{d}q^k_{i,1} - q^k_{\lambda(i),1}\mathrm{d}f_{\rho(i)} - f_{\lambda(i)}\mathbf{d}q^k_{\rho(i),1}\right). \tag{64}$$

Each term within the above parentheses is a sum of terms $p_{i,j}(f)\mathrm{d}f_j$, where each $p_{i,j}$ is homogeneous of degree $\ell(i) - \ell(\lambda(k)) - \ell(j)$. By applying property 4 in Lemma 1 to the term $q^k_{\lambda(i),1}\mathrm{d}f_{\rho(i)} + f_{\lambda(i)}\mathbf{d}q^k_{\rho(i),1}$ and by replacing x with f in Lemma 1, we obtain

$$\alpha_2 = s\left(\mathbf{d}q^k_{i,1} - \mathbf{d}h_1 + \sum_{1<\ell(j)\leq\ell(i)-\ell(\lambda(k))} h_{2,j}(f)\left(\mathrm{d}f_j - f_{\lambda(j)}\mathrm{d}f_{\rho(j)}\right)\right)$$

for some functions h_1 and $h_{2,j}$ Δ^k-homogeneous of degree $\ell(i) - \ell(\lambda(k))$ and $\ell(i) - \ell(\lambda(k)) - \ell(j)$, respectively. Furthermore, by choosing

$$q^k_{i,1} = h_1 \tag{65}$$

(this choice is clearly consistent with **P1(*i*)**), we get

$$\alpha_2 = s \sum_{1<\ell(j)\leq\ell(i)-\ell(\lambda(k))} h_{2,j}(f)\left(\mathrm{d}f_j - f_{\lambda(j)}\mathrm{d}f_{\rho(j)}\right). \tag{66}$$

From what precedes, we finally obtain

$$\alpha_2 = \begin{cases} s\displaystyle\sum_{j=m+1}^{\min\{i,\rho(k)\}-1} h_{2,j}(f)\left(\mathrm{d}f_j - f_{\lambda(j)}\mathrm{d}f_{\rho(j)} + \bar\gamma^k_j\right) - s\sum_{j=m+1}^{\min\{i,\rho(k)\}-1} h_{2,j}(f)\bar\gamma^k_j & \text{if } i < k, \\ s\, m^k_{\rho(k)}\left(\mathrm{d}f_{\rho(k)} - f_{\lambda\rho(k)}\mathrm{d}f_{\rho^2(k)} + \bar\gamma^k_{\rho(k)}\right) - s\, m^k_{\rho(k)}\bar\gamma^k_{\rho(k)} & \text{if } i = k. \end{cases} \tag{67}$$

The second equation is a consequence of (63) when $\lambda\rho(k) = \lambda(k)$, and of (47) and (61) otherwise. As for the first equation, we argue as follows. If $\lambda(i) < \lambda(k)$, the result follows directly from (60) with $h_{2,j} \equiv 0$. If $\lambda(i) = \lambda(k)$ so that $\rho(i) < \rho(k)$, the result follows from (61) or (63). Finally, if $\lambda(i) > \lambda(k)$, then, by (11) and the assumption $i < k$, $\ell(i) < \ell(k)$, so that $\ell(i) - \ell(\lambda(k)) < \ell(\rho(k))$, and the result follows from (66).

Let us now consider the term ${}^3r$ in (57). From Definition 2, ${}^3r$ is a sum of one-forms $s^{2+\kappa_j} r'_j \mathrm{d}f_j$, where each r'_j is a polynomial function of f, Δ^k-homogeneous of degree

$$\ell(i) - \ell(j) - (2+\kappa_j)\ell(\lambda(k)) < \min\{\ell(i) - \ell(j), \ell(\rho(k)) - \ell(j)\}.$$

By applying property 4 in Lemma 1 to each one-form $r'_j \mathrm{d}f_j$, we get

$$(68)\quad {}^3r = s^2 \left(\sum_j s^{\kappa_j} \mathbf{d}h_{1,j} + \sum_{j=m+1}^{\min\{i,\rho(k)\}-1} h'_{2,j}(s,f) \left(\mathrm{d}f_j - f_{\lambda(j)} \mathrm{d}f_{\rho(j)} + \bar{\gamma}_j^k \right) - \sum_{j=m+1}^{\min\{i,\rho(k)\}-1} h'_{2,j}(s,f) \bar{\gamma}_j^k \right),$$

where the functions $h_{1,j}$ are Δ^k-homogeneous of positive degree and therefore vanish at the origin.

We can now define the functions $q_{i,j}^k$. Let us note that $q_{i,1}^k$ has already been defined by (52) if $\lambda(i) \leq \lambda(k)$ and by (65) otherwise. For the definition of $q_{i,j}^k$ with $j > 1$, we distinguish two cases according to whether i is smaller than or equal to k.

If $i < k$, by using (59), (67), and (68), relation (57) can be rewritten in the form (48), with

$$(69)\quad \bar{\gamma}_i^k = \mathbf{d} \left(\sum_{j=2}^{j_{i,k}} s^j q_{i,j}^k \right) + \alpha_1 +^1 r +^2 r - s \sum_{j=m+1}^{\min\{i,\rho(k)\}-1} \left(h_{2,j} + s h'_{2,j} \right)(s,f) \bar{\gamma}_j^k + \sum_j s^{2+\kappa_j} \mathbf{d}h_{1,j}$$

and smooth functions $t_{i,j}$ which we do not need to specify further. The functions $h_{2,j}$ and $s h'_{2,j}$, involved in the above expression of $\bar{\gamma}_i^k$, are polynomial in s and f. From the induction hypothesis and (51), the $\bar{\gamma}_j^k$'s in the right-hand side of (69) are such that $\bar{\gamma}_j^k = \bar{\gamma}_{j,1}^k \mathrm{d}s$ because $j < \rho(k)$. Furthermore, $\bar{\gamma}_{j,1}^k$ depends on s and f only because it is homogeneous of degree $\ell(j) - \ell(\lambda(k)) \leq \ell(\rho(k))$, and $\bar{\gamma}_{j,1}^k(s,c,0) = 0$. As a consequence, we have

$$(70)\quad -s \sum_{j=m+1}^{\min\{i,\rho(k)\}-1} \left(h_{2,j} + s h'_{2,j} \right)(s,f) \bar{\gamma}_j^k = s h'(s,f) \mathrm{d}s = a_0 s^{\kappa'} \mathrm{d}s + h'' \mathrm{d}s$$

with $a_0 \in \mathbb{R}$, $1 \leq \kappa' \in \mathbb{N}$, h' and h'' functions of s and f only, and h'' identically zero when $f = 0$. From Definition 2, (70) can be rewritten as

$$(71)\quad -s \sum_{j=m+1}^{\min\{i,\rho(k)\}-1} \left(h_{2,j} + s h'_{2,j} \right)(s,f) \bar{\gamma}_j^k =^1 r' +^2 r'.$$

From (46), (58), and the fact that $i < k$ implies that either $\lambda(i) < \lambda(k)$ or $\lambda(k) \leq \lambda(i) < \rho(i) < \rho(k)$, we deduce that $\alpha_1 = \mathbf{d}p_i^k$. Therefore, by using (71) in (69),

$$(72)\quad \bar{\gamma}_i^k = \mathbf{d} \left(\sum_{j=2}^{j_{i,k}} s^j q_{i,j}^k \right) + \mathbf{d}p_i^k +^1 r'' + \mathbf{d}(a s^{2+\kappa}) + \sum_j s^{2+\kappa_j} \mathbf{d}h_{1,j},$$

where we have used the fact that any function of type 2 is the differential of a polynomial $a s^q$ with $q \geq 2$. From there, the functions $q_{i,j}^k$ $(j > 1)$ are uniquely defined by setting

$$(73)\quad \sum_{j=2}^{j_{i,k}} s^j q_{i,j}^k \triangleq -a s^{2+\kappa} - \sum_j s^{2+\kappa_j} h_{1,j}.$$

It is simple to check that **P1(*i*)** is verified with this choice. This yields, in view of (72),

$$\bar\gamma_i^k = \mathbf{d}p_i^k +^1 r'' - \sum_j h_{1,j}\mathbf{d}\left(s^{2+\kappa_j}\right) = \mathbf{d}p_i^k +^1 r''',$$

where the last equality comes from the fact that $h_{1,j}(0) = 0$, as mentioned after (68). By using the definition of one-forms of type 1, it follows that (51) is satisfied and thus that **P2(*i*)** is verified—note that, if ${}^1r''' = r_s\mathrm{d}s + r_c\mathrm{d}c$ and $i \le \rho(k)$, then r_c is homogeneous of nonpositive degree so that it is necessarily a constant, which in fact is equal to zero since r_c vanishes at $f = 0$.

For the last case, $i = k$, we proceed similarly. By using (59), (67), and (68), relation (57) can again be rewritten in the form (48), this time with

(74)
$$\bar\gamma_k^k = \mathbf{d}\left(\sum_{j=2}^{j_{k,k}} s^j q_{k,j}^k\right) + \alpha_1 - s m_{\rho(k)}^k \bar\gamma_{\rho(k)}^k +^1 r +^2 r - s^2 \sum_{j=m+1}^{\rho(k)-1} h_{2,j}(s,f)\bar\gamma_j^k + \sum_j s^{2+\kappa_j}\mathbf{d}h_{1,j}$$

instead of (69). From (46), (47), (58), and the induction hypothesis **P2($\rho(k)$)** if $\rho(k) > m$,

(75)
$$\begin{aligned}\alpha_1 - s\, m_{\rho(k)}^k \bar\gamma_{\rho(k)}^k &= \alpha_1 - s\, m_{\rho(k)}^k \mathrm{d}c - s\, m_{\rho(k)}^k \bar\gamma_{\rho(k),1}^k \mathrm{d}s \\ &= \frac{m_k^k}{2}(c\mathrm{d}s - s\mathrm{d}c) - s\, m_{\rho(k)}^k \bar\gamma_{\rho(k),1}^k \mathrm{d}s\,.\end{aligned}$$

If $\rho(k) \le m$ so that $\lambda(k) < \rho(k) \le m < k$, these equalities are still valid since (47) implies that $m_{\rho(k)}^k = 0$. Using (75), (74) rewrites as

(76)
$$\begin{aligned}\bar\gamma_k^k = \mathbf{d}\left(\sum_{j=2}^{j_{k,k}} s^j q_{k,j}^k\right) + \frac{m_k^k}{2}(c\mathrm{d}s - s\mathrm{d}c) +^1 r +^2 r - s^2 \sum_{j=m+1}^{\rho(k)-1} h_{2,j}(s,f)\bar\gamma_j^k \\ - s\, m_{\rho(k)}^k \bar\gamma_{\rho(k),1}^k \mathrm{d}s + \sum_j s^{2+\kappa_j}\mathbf{d}h_{1,j}.\end{aligned}$$

From here, we proceed as for the previous case in order to rewrite the above equation as (compare with (72))

$$\bar\gamma_k^k = \mathbf{d}\left(\sum_{j=2}^{j_{k,k}} s^j q_{k,j}^k\right) + \frac{m_k^k}{2}(c\mathrm{d}s - s\mathrm{d}c) +^1 r'' + \mathbf{d}(as^{2+\kappa}) + \sum_j s^{2+\kappa_j}\mathbf{d}h_{1,j}\,. \tag{77}$$

Using again (73) to define the functions $q_{k,j}^k$ $(j > 1)$ yields

$$\bar\gamma_k^k = \frac{m_k^k}{2}(c\mathrm{d}s - s\mathrm{d}c) +^1 r''',$$

and it is simple to check that the one-form $\bar\gamma_k^k$ satisfies (51) so that **P2(*i*)** is verified. This ends the study of Case 2.

Case 3. $k < i \le n(d)$. We define $q_{i,j}^k \equiv 0$ according to (52) so that both **P1(*i*)** and **P2(*i*)** are readily verified. This ends the proof of Lemma 3.

Proof of Lemma 4. Since $f^k = \bar{f}^k \circ \bar{g}^k_\eta$, we deduce from (43), (48), and (49) that, for $i \in \{m+1, \dots, k\}$,

$$\begin{aligned} \omega^k_i &= \bar{\omega}^k_i \circ \mathbf{d}\bar{g}^k_\eta \\ &= \left(\omega^{k-1}_i + \gamma^k_i \mathrm{d}\theta_k\right) + \sum_{j=m+1}^{i-1} t'_{i,j} \left(\omega^{k-1}_j + \gamma^k_j \mathrm{d}\theta_k\right), \end{aligned} \tag{78}$$

where

$$\gamma^k_i(\theta^k) = \bar{\gamma}^k_{i,1}(\bar{g}^k_\eta(\theta^k))\eta^{\ell(\lambda(k))} \cos\theta_k - \bar{\gamma}^k_{i,2}(\bar{g}^k_\eta(\theta^k))\eta^{\ell(\rho(k))} \sin\theta_k \,. \tag{79}$$

By skew-symmetry of the wedge product, it follows from (78) that

$$\omega^k_{m+1} \wedge \cdots \wedge \omega^k_k = \left(\omega^{k-1}_{m+1} + \gamma^k_{m+1}\mathrm{d}\theta_k\right) \wedge \cdots \wedge \left(\omega^{k-1}_k + \gamma^k_k \mathrm{d}\theta_k\right).$$

Since each ω^{k-1}_i is a one-form on $\mathbb{T}^{k-m-1}$, we deduce from the above equation (using multilinearity and skew-symmetry of the wedge product) that

$$\omega^k_{m+1} \wedge \cdots \wedge \omega^k_k = \sum_{i=m+1}^{k} \gamma^k_i \left(\omega^{k-1}_{m+1} \wedge \cdots \wedge \omega^{k-1}_{i-1} \wedge \mathrm{d}\theta_k \wedge \omega^{k-1}_{i+1} \wedge \cdots \wedge \omega^{k-1}_k\right). \tag{80}$$

From (45) and (53),

$$\begin{aligned} \bar{\gamma}^k_{i,1}(\bar{g}^k_\eta(\theta^k)) &= \bar{\gamma}^k_{i,1}(\Delta^k_\eta(\sin\theta_k, \cos\theta_k, \Delta_{1/\eta} f^{k-1}(\theta^{k-1}))) \\ &= \eta^{\ell(i)-\ell(\lambda(k))}\bar{\gamma}^k_{i,1}(\sin\theta_k, \cos\theta_k, \Delta_{1/\eta} f^{k-1}(\theta^{k-1})) \\ &= \eta^{\ell(i)-\ell(\lambda(k))}\bar{\gamma}^k_{i,1}(\sin\theta_k, \cos\theta_k, 0) + \sum_{j<\ell(i)-\ell(\lambda(k))} \eta^j \bar{\beta}_{i,j}(\theta^k), \end{aligned} \tag{81}$$

where the $\bar{\beta}_{i,j}$'s denote smooth functions on $\mathbb{T}^{k-m}$. The second equality in the above equation comes from the fact that $\bar{\gamma}^k_{i,1}$ is Δ^k-homogeneous of degree $\ell(i) - \ell(\lambda(k))$, and the third one from the fact that $\bar{\gamma}^k_{i,1}(s, c, f)$ is polynomial in s, c, and f. A similar calculation yields

$$\bar{\gamma}^k_{i,2}(\bar{g}^k_\eta(\theta^k)) = \eta^{\ell(i)-\ell(\rho(k))}\bar{\gamma}^k_{i,2}(\sin\theta_k, \cos\theta_k, 0) + \sum_{j<\ell(i)-\ell(\rho(k))} \eta^j \bar{\bar{\beta}}_{i,j}(\theta^k). \tag{82}$$

From (51), (79), (81), and (82),

$$\gamma^k_i(\theta^k) = \begin{cases} \eta^{\ell(k)}\dfrac{m^k_k}{2} + \displaystyle\sum_{1<j<\ell(k)} \eta^j \beta_{k,j}(\theta^k) & \text{if } i = k, \\ \displaystyle\sum_{1<j<\ell(k)} \eta^j \beta_{i,j}(\theta^k) & \text{otherwise} \end{cases} \tag{83}$$

for some smooth functions $\beta_{i,j}$ on $\mathbb{T}^{k-m}$. In view of (80) and (83),

$$\left(\omega^k_{m+1} \wedge \cdots \wedge \omega^k_k\right)(\theta^k) = \eta^{\ell(k)}\frac{m^k_k}{2}\left(\omega^{k-1}_{m+1} \wedge \cdots \wedge \omega^{k-1}_{k-1}\right)(\theta^{k-1}) + \sum_{1\le j<\ell(k)} \eta^j \beta'_{k,j}(\theta^k)$$

for some other smooth functions $\beta'_{k,j}$ on $\mathbb{T}^{k-m}$. By the compacity of $\mathbb{T}^{k-m}$ and the induction hypothesis, (42) follows when η is larger than some $\eta_0 > 0$. ∎

3. Example. We illustrate the construction of transverse functions, as specified in the proof of Proposition 4, in the case of the free system $S(2,3)$ on $\mathbb{R}^5$. The associated truncated P. Hall basis is $\mathcal{B}_3 = \{B_1, \dots, B_5\}$, where

$$B_1 \triangleq X_1, B_2 \triangleq X_2, B_3 \triangleq [B_1, B_2] = [X_1, X_2], B_4 \triangleq [B_1, B_3], B_5 \triangleq [B_2, B_3]\,. \tag{84}$$

We have to compute $f = f^{n(d)} = f^5$, starting from $f^{m+1} = f^3$. From (14) and (84), $\lambda(3) = 1$ and $\rho(3) = 2$. Therefore, in view of (44),

$$f^3(\theta_3) = \left(\sin\theta_3, \cos\theta_3, \frac{\sin 2\theta_3}{4}, 0, 0\right)^T . \tag{85}$$

Let us now compute f^4 from f^3. From (14) and (84), $\lambda(4) = 1$ and $\rho(4) = 3$. Then (46), (47), (50), and (52) give

$$\bar{f}^4(s,c,x) = x + \begin{pmatrix} s \\ 0 \\ c \\ s\,c \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ sq^4_{3,1}(x) + s^2 q^4_{3,2}(x) \\ sq^4_{4,1}(x) + s^2 q^4_{4,2}(x) + s^3 q^4_{4,3}(x) \\ 0 \end{pmatrix}. \tag{86}$$

From (52)

$$\begin{cases} q^4_{3,1}(x) = m^4_3\, x_{\rho(3)} = x_{\rho(3)} = x_2, \\ q^4_{4,1}(x) = m^4_4\, x_{\rho(4)} = 2x_{\rho(4)} = 2x_3. \end{cases} \tag{87}$$

Let us now proceed with the determination of $\bar{\omega}^4_3$, as defined by (49). Since $q^4_{3,2}$ is by definition homogeneous of degree $\ell(3) - 2\ell(1) = 0$, it is a constant function. A direct calculation gives

$$\bar{\omega}^4_3 = \mathrm{d}x_3 - x_1\mathrm{d}x_2 + (x_2 + 2sq^4_{3,2})\mathrm{d}s + \mathrm{d}c\,.$$

With the simple choice

$$q^4_{3,2} \equiv 0\,, \tag{88}$$

consistent with **P1(3)**, it follows that (48) is verified with $\bar{\gamma}^4_3 \triangleq x_2\mathrm{d}s + \mathrm{d}c$, a one-form which satisfies the conditions in **P2(3)**. There remains to determine $q^4_{4,2}$ and $q^4_{4,3}$. Again, $q^4_{4,3}$ is homogeneous of degree zero, and thus it is a constant function. A simple calculation gives

$$\begin{aligned} \bar{\omega}^4_4 =& \mathrm{d}x_4 - x_1\mathrm{d}x_3 + s(\mathrm{d}x_3 - x_1\mathrm{d}x_2 + \bar{\gamma}^4_3) + s^2(\mathbf{d}q^4_{4,2} - \mathrm{d}x_2) - (x_1 + s)\mathrm{d}c \\ &+ (c + 2x_3 + 2sq^4_{4,2} + 3s^2 q^4_{4,3} - x_1x_2 - 2sx_2)\mathrm{d}s\,. \end{aligned}$$

The choice

$$q^4_{4,2}(x) = x_2\,, \quad q^4_{4,3} \equiv 0\,, \tag{89}$$

is clearly consistent with **P1(4)** and allows us to rewrite $\bar{\omega}_4^4$ in the form (48), with $\bar{\gamma}_4^4 \triangleq (c + 2x_3 - x_1x_2)\mathrm{d}s - (x_1 + s)\mathrm{d}c$ a one-form which satisfies the conditions in **P2(4)**. We finally obtain the following from (86), (87), (88), and (89):

$$\bar{f}^4(s,c,x) = x + \begin{pmatrix} s \\ 0 \\ c + sx_2 \\ s\,c + 2sx_3 + s^2x_2 \\ 0 \end{pmatrix}. \tag{90}$$

The expression of f^4 is then obtained by application of (53). As for the parameter η_4, it must be chosen large enough so that (42) is satisfied for $k = 4$. By inspection the (conservative) condition $\eta_4 \geq 5/2$ can be derived.

The determination of f^5 from f^4 is performed in the same way. We obtain (details are left to the reader)

$$\bar{f}^5(s,c,x) = x + (0, s, c, 0, s\,c/2 + sx_3)^T. \tag{91}$$

Then, (53) gives the expression of $f = f^5$. One verifies from (85), (90), and (91) that

$$f(\theta^5) = \begin{pmatrix} \sin\theta_3 + \eta_4 \sin\theta_4 \\ \cos\theta_3 + \eta_5 \sin\theta_5 \\ \dfrac{1}{4}\sin 2\theta_3 + \eta_4^2\cos\theta_4 + \eta_4 \sin\theta_4\cos\theta_3 + \eta_5^2\cos\theta_5 \\ \dfrac{\eta_4^3}{2}\sin 2\theta_4 + \dfrac{\eta_4}{2}\sin\theta_4\sin 2\theta_3 + \eta_4^2\sin^2\theta_4\cos\theta_3 \\ \dfrac{\eta_5^3}{4}\sin 2\theta_5 + \eta_5\sin\theta_5(f_3(\theta_5) - \eta_5^2\cos\theta_5) \end{pmatrix}.$$

For practical purposes, adequate values for the parameters η_4 and η_5 must be specified. In this respect, let us mention only that numerical computations tend to indicate that for $\eta_4 = 3$ any value $\eta_5 \geq 7$ guarantees the satisfaction of (42).

Appendix.

Proof of Lemma 1 (*property* 4). We assume that $i \in \{1, \ldots, m\}$, since otherwise a simple algebraic manipulation yields

$$\begin{aligned} \mathrm{d}x_i = (\mathrm{d}x_i - x_{\lambda(i)}\mathrm{d}x_{\rho(i)}) + \sum_{r=1}^{\bar{r}} x_{\lambda(i)}x_{\lambda\rho(i)}\ldots x_{\lambda\rho^{r-1}(i)}(\mathrm{d}x_{\rho^r(i)} - x_{\lambda\rho^r(i)}\mathrm{d}x_{\rho^{r+1}}) \\ + x_{\lambda(i)}x_{\lambda\rho(i)}\ldots x_{\lambda\rho^{\bar{r}}(i)}\mathrm{d}x_{\rho^{\bar{r}+1}(i)}, \end{aligned}$$

where $\bar{r}$ is the smallest integer such that $\rho^{\bar{r}+1}(i) \in \{1, \ldots, m\}$. It is sufficient to specify some functions h_1 and $h_{2,j}$ such that equality (20) holds when each side is applied to any element of the basis $\{b_r, \partial/\partial x_s,\ r = 1, \ldots, m,\ s = m+1, \ldots, n(d)\}$ of the tangent space to $\mathbb{R}^n$. From (17),

$$\forall i = 1, \ldots, m \quad \begin{cases} \mathrm{d}x_j(b_i) = \delta_i^j & \text{if } j \in \{1, \ldots, m\}, \\ \omega_j(b_i) = 0 & \text{if } j \in \{m+1, \ldots, n(d)\}, \end{cases} \tag{92}$$

where $\omega_j = \mathrm{d}x_j - x_{\lambda(j)}\mathrm{d}x_{\rho(j)}$. Therefore, (20) applied to any b_r holds by setting $h_1 = q^i$ defined by (19). Finally, the functions $h_{2,j}$ are defined recursively, for $\ell(j)$ decreasing from d' to 2, by setting

$$\begin{cases} h_{2,j} = -\dfrac{\partial h_1}{\partial x_j}, & \ell(j) = d', \\ h_{2,j} = -\dfrac{\partial h_1}{\partial x_j} + \displaystyle\sum_{\ell(j)<\ell(j')\le d'} h_{2,j'} x_{\lambda(j')} \mathrm{d}x_{\rho(j')}(\partial/\partial x_j), & 1 < \ell(j) < d'. \end{cases} \qquad \square$$

Proof of Lemma 2. Since the set $\{g_1, \dots, g_m\}$ is nilpotent of order $d+1$, it follows from the definition of the P. Hall basis that $\{g_1, \dots, g_{n(d)}\}$ is a basis of $\mathrm{Lie}\{g_1, \dots, g_m\}$. Therefore, it is clearly a basis of $\mathrm{Lie}\{g_1, \dots, g_{n(d)}\}$. Then, (33) follows from the well-known fact that the solution of (32) is an exponential Lie series (see, e.g., [16] for details).

Let us finally prove that the mapping defined by (34) is one-to-one. Consider the system

$$\dot{x} = \sum_{i=1}^{n(d)} c_i g_i(x). \tag{93}$$

From property 2 of Lemma 1, each v.f. g_i is smooth and Δ-homogeneous of strictly negative degree. Therefore, its kth component $g_{i,k}$ can depend only on the components x_j of x such that $j < k$. From this and property 1 of Lemma 1, we deduce that the kth component of (93) can be written as

$$\dot{x}_k = c_k a_k + h_k(x_k^-, c_k^-), \tag{94}$$

where the notation y_k^- for a vector $y \in \mathbb{R}^n$ denotes the subvector $(y_1, \dots, y_{k-1})$, and h_k is some smooth function. Using (94), one easily proves by induction on k that any solution to (93) satisfies

$$\forall k = 1, \dots, n\,, \forall t, \qquad x_k(t) = x_k(0) + t c_k a_k + f_k(x_k^-(0), c_k^-, t)$$

for some smooth function f_k. Therefore,

$$\forall k = 1, \dots, n\,, \qquad \left[\exp\left(\sum_{i=1}^{n(d)} c_i g_i\right) x_0\right]_k = x_{0,k} + c_k a_k + f_k(x_{0,k}^-, c_k^-, 1),$$

and one easily infers from these equalities that

$$(c_1, \dots, c_{n(d)}) \neq (c'_1, \dots, c'_{n(d)}) \Longrightarrow \exp\left(\sum_{i=1}^{n(d)} c_i g_i\right) x_0 \neq \exp\left(\sum_{i=1}^{n(d)} c'_i g_i\right) x_0\,. \qquad \square$$

REFERENCES

[1] R. Abraham, J. Marsden, and T. Ratiu, *Manifolds, Tensor Analysis, and Applications*, 2nd ed., Springer-Verlag, New York, 1988.

[2] W. Chow, *Uber systeme von linearen partiellen differential-gleichungen erster ordnung*, Math. Ann., 117 (1939), pp. 98–105.

[3] W. Dixon, D. Dawson, E. Zergeroglu, and F. Zhang, *Robust tracking and regulation control for mobile robots*, Internat. J. Robust Nonlinear Control, 10 (2000), pp. 199–216.
[4] R. Hermann, *On the accessibility problem in control theory*, in Proceedings of the International Symposium on Nonlinear Differential Equations and Nonlinear Mechanics, Academic Press, New York, 1963, pp. 325–332.
[5] H. Hermes, *Nilpotent and high-order approximations of vector field systems*, SIAM Rev., 33 (1991), pp. 238–264.
[6] M. Kawski, *Homogeneous stabilizing control laws*, Control Theory Adv. Tech., 6 (1990), pp. 497–516.
[7] M. Kawski, *Nonlinear control and combinatorics of words*, in Geometry of Feedback and Optimal Control, B. Jakubczyk and W. Respondek, eds., Dekker, New York, 1998, pp. 305–346.
[8] M. Kawski and H. J. Sussmann, *Noncommutative power series and formal Lie-algebraic techniques in nonlinear control theory*, in Operators, Systems, and Linear Algebra, D. P.-W. U. Helmke and E. Zerz, eds., Teubner, Stuttgart, Germany, 1997, pp. 111–128.
[9] W. Liu, *An approximation algorithm for nonholonomic systems*, SIAM J. Control Optim., 35 (1997), pp. 1328–1365.
[10] C. Lobry, *Controlabilité des systèmes non linéaires*, SIAM J. Control, 8 (1970), pp. 573–605.
[11] P. Morin, J.-B. Pomet, and C. Samson, *Design of homogeneous time-varying stabilizing control laws for driftless controllable systems via oscillatory approximation of Lie brackets in closed loop*, SIAM J. Control Optim., 38 (1999), pp. 22–49.
[12] P. Morin and C. Samson, *A characterization of the Lie algebra rank condition by transverse periodic functions*, in Proceedings of the IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, 2000, pp. 3988–3993.
[13] P. Morin and C. Samson, *Practical stabilization of a class of nonlinear systems. Application to chain systems and mobile robots*, in Proceedings of the IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, 2000, pp. 2989–2994.
[14] J.-P. Serre, *Lie Algebras and Lie Groups*, 2nd ed., Springer-Verlag, New York, 1992.
[15] G. Stefani, *Polynomial approximations to control systems and local controllability*, in Proceedings of the IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, 1985, pp. 33–38.
[16] H. J. Sussmann, *Lie brackets and local controllability: A sufficient condition for scalar-input systems*, SIAM J. Control Optim., 21 (1983), pp. 686–713.
[17] H. J. Sussmann, *A product expansion for the Chen series*, in Theory and Applications of Nonlinear Control Systems, C. Byrnes and A. Lindquist, eds., North Holland, Amsterdam, pp. 323–335.
[18] H. J. Sussmann and V. Jurdjevic, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 4, pp. 1250–1269

A STOCHASTIC CONTROL APPROACH TO PORTFOLIO PROBLEMS WITH STOCHASTIC INTEREST RATES*

RALF KORN† AND HOLGER KRAFT‡

Abstract. We consider investment problems where an investor can invest in a savings account, stocks, and bonds and tries to maximize her utility from terminal wealth. In contrast to the classical Merton problem, we assume a stochastic interest rate. To solve the corresponding control problems it is necessary to prove a verification theorem without the usual Lipschitz assumptions.

Key words. optimal portfolios, stochastic interest rate, verification theorem

AMS subject classifications. 93E20, 91B28, 60H30

PII. S0363012900377791

1. Introduction. The continuous-time portfolio problem has its origin in the pioneering work of Merton (1969, 1971, 1973). It is concerned with finding the optimal investment strategy of an investor. More precisely, the investor looks for an optimal decision on how many shares of which security she should hold at every time instant between now and a time horizon T to maximize her expected utility from wealth at the time horizon. In the classical Merton problem the investor can allocate her money into a riskless savings account and d different risky stocks. By describing the actions of the investor via the portfolio process (i.e., the percentages of wealth invested in the different securities), Merton was able to reduce the portfolio problem to a control problem which could be solved by using standard stochastic control methodology.

A drawback of this approach, however, is the assumption of a deterministic interest rate.[1] Our main objective in the current paper is to overcome this restriction. We assume that the interest rate follows an Ito process and particularly consider the case of the Ho–Lee (see Ho and Lee (1986)) model and the Vasicek (see Vasicek (1977)) model for the short rate. Such problems are treated rarely in the literature.[2] Further, our theory will enable us to consider mixed bond and stock portfolio problems. We give explicit solutions for both the value functions and the optimal strategies in section 2.

On the theoretical side, the introduction of stochastic interest rates into the portfolio problem has the consequence that the SDE describing the wealth process does not satisfy the usual Lipschitz assumptions needed to apply standard verification theorems. However, due to the special structure of this equation, the wealth equation, we are able to prove a suitable verification result in the appendix. This is possible as some assumptions of the standard verification results as, e.g., given in Fleming and Soner (1993) can be weakened substantially via proving some special estimates.

*Received by the editors September 6, 2000; accepted for publication (in revised form) May 10, 2001; published electronically December 7, 2001.
http://www.siam.org/journals/sicon/40-4/37779.html

†Department of Mathematics, University of Kaiserslautern and Head of Department of Finance, Fraunhofer ITWM, 67653 Kaiserslautern, Germany (Korn@mathematik.uni-kl.de).

‡Department of Finance, Fraunhofer ITWM, 67653 Kaiserslautern, Germany (Kraft@itwm.fhg.de).

[1]The other main approach to optimal portfolios, the martingale method, plays no role in this paper. We refer to Korn (1997) for an introduction to it.

[2]For related problems see Klüppelberg and Korn (1998), Canestrelli and Pontini (1998), and Sørensen (1999). In particular, in Sørensen (1999) the martingale approach of portfolio optimization is used in contrast to our stochastic control approach.

Important future research topics are the solution of the problems treated in this paper for other interest rate models such as the Cox–Ingersoll–Ross (see Cox, Ingersoll, and Ross (1985)) or the corresponding Hull–White (see Hull and White (1990)) approach. This is particularly interesting as those models have some desirable features (such as nonnegative interest rates) that the Ho–Lee or Vasicek model do not have.

2. Two portfolio problems. We consider an economy with $d+1$ assets which are continuously traded on a frictionless market. All traders are assumed to be price takers. The uncertainty is modelled by a probability space $(\Omega, \mathcal{F}, P)$. On this space an m-dimensional Brownian motion $\{(W(t), \mathcal{F}_t)\}_{t\geq 0}$ is defined, where $\{\mathcal{F}_t\}_{t\geq 0}$ denotes the Brownian filtration. One of the assets is a savings account following the differential equation

$$dB(t) = B(t)r(t)dt$$

with $B(0) = 1$. Here r denotes the short rate which can be interpreted as the annualized interest for the infinitesimal period $[t, t+dt]$.

In contrast to Merton's classical model,[3] we assume a short rate modelled by the SDE

$$dr(t) = a(t)dt + b\, dW(t),$$

$t \in [0, T^*]$, $b > 0$, with initial data $r(0) = r_0$. As explicit examples we will consider the Ho–Lee model given by $a(t) = \tilde{a}(t) + b\zeta(t)$ and a Vasicek approach with $a(t) = \theta(t) - \alpha r(t) + b\zeta(t)$, $\alpha > 0$, respectively. The risk premium (RP) ζ is assumed to be deterministic and continuous which implies the progressive measurability of ζ. This assumption particularly guarantees that ζ is bounded on each compact interval. Furthermore, let the initial forward rate curve $f^*(0,T)$, $0 \leq T \leq T^*$, be continuously differentiable, which leads to $\tilde{a}(t) = f^*_T(0,t) + b^2 t$ and $\theta(t) = f^*_T(0,t) + \alpha f^*(0,t) + \frac{b^2}{2\alpha}(1 - e^{-2\alpha t})$.[4] The price processes of the remaining d assets which can be stocks and/or (discount) bonds are assumed to follow Ito processes of the form

$$dP_i(t) = P_i(t)\Big[\mu_i(t)dt + \sigma_i(t)dW(t)\Big]$$

with $P_i(0) = p_i > 0$ and where $\mu(\cdot)$ is $\mathbb{R}^d$-valued and $\sigma_i(\cdot)$ denotes the ith row of the $d \times m$-matrix $\sigma(\cdot)$.

We consider an investor who starts with an initial wealth $x_0 > 0$ at time $t = 0$. In the beginning this initial wealth is invested in the different assets, and she is allowed to adjust her holdings continuously up to a fixed planning horizon T. Her investment behavior is modelled by a portfolio process $\pi = (\pi_1, \ldots, \pi_d)$ which is progressively measurable (with respect to $\{\mathcal{F}_t\}_{t\geq 0}$). Here, $\pi_i(t)$, $i = 0, \ldots, d$ denotes the percentage of total wealth invested in the ith asset at time t. Obviously, the percentage invested in the savings account is given by $1 - \pi'\underline{1}$, where $\underline{1} := (1, \ldots, 1)' \in \mathbb{R}^d$.

If we restrict our considerations to self-financing portfolio processes, her wealth process follows the SDE

$$dX(t) = X(t)\Big[\big(\pi(t)'(\mu(t) - r(t)\cdot\underline{1}) + r(t)\big)dt + \pi(t)'\sigma(t)dW(t)\Big] \tag{2.1}$$

with $X(0) = x_0$.[5]

[3] See Merton (1969, 1971, 1990), Fleming and Rishel (1975), pp. 160f, Duffie (1992), pp.145ff, Fleming and Soner (1993), pp. 174ff, and Korn (1997), pp. 48ff.

[4] See, for example, Musiela and Rutkowski (1997), pp. 323f.

[5] See, for example, Korn (1997), pp. 23f.

The wealth equation can be interpreted as a controlled SDE with the control being the portfolio process $\pi(\cdot)$. In this setting the investor chooses a portfolio process to maximize her utility. We assume that her preferences can be represented by the utility function $U(x) = x^\gamma$, $x \geq 0$, $0 < \gamma < 1$. Furthermore, the investor is allowed only to pick out a portfolio process which is admissible in the sense of Definition 3.1 and leads to a *positive* wealth process X^π. Now we are in the position to formulate her optimization problem:[6]

$$\max_{\pi(\cdot)\in\mathcal{A}^*(0,x_0)} \mathrm{E}(X^\pi(T))^\gamma \tag{2.2}$$

with

$$\begin{aligned} dX^\pi(t) &= X^\pi(t)\Big[(\pi(t)'(\mu(t) - r(t)\cdot\underline{1}) + r(t))dt + \pi(t)\sigma(t)dW(t)\Big], \\ X^\pi(0) &= x_0, \end{aligned}$$

and

$$\mathcal{A}^*(0,x_0) := \Big\{\pi(\cdot)\in\mathcal{A}(0,x_0) \,:\, X^\pi(s)\geq 0 \;\; P-\text{a.s. for } s\in[0,T]\Big\}.$$

We emphasize that applying optimal control methods to this problem does not automatically yield a positive state process. However, Corollary 3.1 and the special form of the coefficients in the wealth equation (2.1) will indeed guarantee the positivity of $X^\pi(t)$. Therefore, we obtain $\mathcal{A}^*(0,x_0) = \mathcal{A}(0,x_0)$.

2.1. A bond portfolio problem. We start by considering a portfolio problem where the investor can split up his wealth in a savings account and a (zero) bond with maturity $T_1 > T$. We assume that the asset price processes can be represented by the Ito processes

$$\begin{aligned} dB(t) &= B(t)r(t)dt, \\ dP(t,T_1) &= P(t,T_1)\Big[\underbrace{(r(t)+\zeta(t)\sigma(t))}_{=:\mu(t)}dt + \sigma(t)dW(t)\Big], \end{aligned}$$

where W is a one-dimensional Brownian motion. In the Ho–Lee and the Vasicek models the volatility of the bond is given by $\sigma(t) = -b(T_1 - t)$ and $\sigma(t) = \frac{b}{\alpha}(\exp(-\alpha(T_1 - t)) - 1)$, respectively.[7] Let $\pi(t)$ be the percentage invested in the bond. This leads to a wealth equation of the form

$$\begin{aligned} dX(t) &= X(t)\Big[(\pi(t)\mu(t) + (1-\pi(t))r(t))dt + \pi(t)\sigma(t)dW(t)\Big] \\ &= X(t)\Big[(\pi(t)\zeta(t)\sigma(t) + r(t))dt + \pi(t)\sigma(t)dW(t)\Big] \end{aligned} \tag{2.3}$$

with initial data $X(0) = x_0$.

In contrast to the classical Merton problem, we assume a stochastic short rate; the drift coefficient includes the additional stochastic term $r(t)$. Thus, to solve the

[6] Here $\mathcal{A}(0,x_0)$ denotes the set of all admissible controls corresponding to the initial condition $(0,x_0)$. See Definition 3.1 in the appendix.

[7] See, for example, Musiela and Rutkowski (1997), pp. 323ff.

portfolio problem (2.2) by stochastic control methods, we have to look at a two-dimensional state process $Y = (X, r)$. Note that the second component cannot be controlled via $\pi(\cdot)$. Using the notation of (3.1) in the appendix, we get[8]

$$\begin{aligned}
Y(t) &= (X(t), r(t))', \\
\Lambda(t,x,r,\pi) &= (x(\pi\zeta\sigma + r), a)', \\
\Sigma(t,x,r,\pi) &= (x\pi\sigma, b)', \\
\Sigma^*(t,x,r,\pi) &= \begin{pmatrix} x^2\pi^2\sigma^2 & bx\pi\sigma \\ bx\pi\sigma & b^2 \end{pmatrix}, \\
A^\pi G(t,x,r) &= G_t + 0.5(x^2\pi^2\sigma^2 G_{xx} + 2x\pi b\sigma G_{xr} + b^2 G_{rr}) \\
&\quad + x(\pi\zeta\sigma + r)G_x + aG_r.
\end{aligned}$$

Hence the following Hamilton–Jacobi–Bellman (HJB) equation has to be solved:

$$\begin{aligned}
\sup_{|\pi|\le\delta} A^\pi G(t,x,r) &= 0, \\
G(T,x,r) &= x^\gamma,
\end{aligned}$$

where $\delta > 0$ will be specified later.

Note that due to the presence of the product rx in the above setting usual verification theorems which require Lipschitz conditions are not applicable to our situation as both the wealth process and the short rate are unbounded processes. We therefore give a suitable verification result (Corollary 3.2) in the appendix. This result then allows us to solve the HJB equation with the usual three step procedure. By this, we would like to emphasize our opinion that the third step, verification of all assumptions of both Corollary 3.2 and those made to perform the following calculations, is an essential part of the solution.

We start with the calculation of the optimal bond position $\pi(\cdot)$.

1st step. Assuming $G_{xx} < 0$, we get the following candidate for the optimal bond position:

$$\pi^* = -\frac{\zeta}{\sigma}\frac{G_x}{xG_{xx}} - \frac{b}{\sigma}\frac{G_{xr}}{xG_{xx}}. \tag{2.4}$$

2nd step. Inserting $\pi^*(t,x,r;G)$ into the HJB equation leads to the PDE

$$\begin{aligned}
0 &= G_t G_{xx} - 0.5\zeta^2 G_x^2 - 0.5b^2 G_{xr}^2 + 0.5b^2 G_{rr} G_{xx} \\
&\quad - b\zeta G_x G_{xr} + xrG_x G_{xx} + aG_r G_{xx}
\end{aligned} \tag{2.5}$$

with the terminal condition $G(T,x,r) = x^\gamma$. Note that $\zeta = (\mu - r)/\sigma$.

The form of this condition recommends the following separation ansatz:

$$G(t,x,r) = f(t,r)\cdot x^\gamma \quad \text{with } f(T,r) = 1 \text{ for all } r.$$

This leads to a second-order PDE for f of the form

$$\begin{aligned}
0 &= (\gamma-1)ff_t - 0.5b^2\gamma f_r^2 - 0.5\zeta^2\gamma f^2 + 0.5b^2(\gamma-1)ff_{rr} \\
&\quad - b\zeta\gamma f f_r + r\gamma(\gamma-1)f^2 + a(\gamma-1)ff_r
\end{aligned}$$

[8]For simplicity we often neglect the functional dependencies with respect to t, x, and r.

with terminal condition $f(T,r)=1$. Using the ansatz

$$f(t,r)=g(t)\cdot\exp(\beta(t)\cdot r)$$

with terminal conditions $\beta(T)=0$ and $g(T)=1$ and simplification yields

$$\begin{aligned}0&=(\gamma-1)\cdot g'+(\gamma-1)\,(\gamma+\beta')\cdot rg\\&\quad-\left(0.5\zeta^2\gamma+0.5b^2\beta^2+b\zeta\gamma\beta\right)\cdot g+a(\gamma-1)\beta\cdot g.\end{aligned}\tag{2.6}$$

Our ansatz for f will only be meaningful if we get an ODE for g which does not include the short rate r.

In the Ho–Lee model the drift a of the short rate is a function of t, whereas in the Vasicek model it is a function of t and r. Therefore, we treat the two interest rate models separately.

Ho–Lee model. In our Ho–Lee setting PDE (2.6) has the form

$$\begin{aligned}0&=(\gamma-1)\cdot g'+(\gamma-1)\,(\gamma+\beta')\cdot rg\\&\quad+\underbrace{\left(-0.5\zeta^2\gamma-0.5b^2\beta^2-b\zeta\gamma\beta+a(\gamma-1)\beta\right)}_{=:h_1(t)}\cdot g.\end{aligned}\tag{2.7}$$

Since $a(t)=f_T^*(0,t)+b^2t+b\zeta(t)$ and ζ is assumed to be deterministic and continuous, h_1 is a continuous and deterministic function. Choosing $\beta(t)=\gamma(T-t)$, we infer from (2.7) the following first-order ODE for g:

$$0=(\gamma-1)\cdot g'+h_1(t)\cdot g$$

with $g(T)=1$. Separation of variables leads to

$$g(t)=\exp\left(\tfrac{1}{1-\gamma}(H_1(t)-H_1(T))\right),$$

where H_1 is a primitive of h_1. Hence we obtain

$$G(t,x,r)=x^\gamma\cdot\exp\left(\frac{1}{1-\gamma}(H_1(t)-H_1(T))+\gamma(T-t)r\right)$$

as a candidate for the value function. Inserting this into (2.4) gives the corresponding control

$$\begin{aligned}\pi^*(t)&=\frac{1}{1-\gamma}\cdot\frac{\zeta(t)+b\beta(t)}{-\sigma(t)}\\&=\frac{1}{1-\gamma}\cdot\frac{\zeta(t)+b(T-t)\gamma}{-b(T_1-t)}.\end{aligned}$$

Obviously, $\pi^*(\cdot)$ is continuous, deterministic, and therefore bounded.

Vasicek model. With the Vasicek specification of a the PDE (2.6) has the following form:

$$\begin{aligned}0&=(\gamma-1)\cdot g'+\underbrace{(\gamma-1)(\beta'-\alpha\beta+\gamma)}_{(*)}\cdot rg\\&\quad+(\underbrace{\theta(\gamma-1)\beta-b\zeta\beta-0.5b^2\beta^2-0.5\zeta^2\gamma}_{=:h_2(t)})\cdot g.\end{aligned}$$

Our ansatz for f is only meaningful if β can be calculated so that the factor $(*)$ becomes zero. As a result, we have to solve an inhomogeneous ODE for β which has the following form:

$$\beta'(t) = \alpha\beta(t) - \gamma$$

with $\beta(T) = 0$ leading to

$$\beta(t) = \frac{\gamma}{\alpha}(1 - \exp(\alpha(t - T))).$$

Choosing β as calculated, we again get a first-order homogeneous ODE for g,

$$0 = (\gamma - 1) \cdot g' + h_2(t) \cdot g,$$

with $g(T) = 1$. Hence

$$g(t) = \exp\left(\frac{1}{1-\gamma}(H_2(t) - H_2(T))\right),$$

where H_2 is a primitive of h_2. Therefore,

$$G(t, x, r) = x^\gamma \cdot \exp\left(\frac{1}{1-\gamma}(H_2(t) - H_2(T)) + \frac{\gamma}{\alpha}(1 - \exp(\alpha(t - T)))r\right).$$

The corresponding control reads as follows:

$$\begin{aligned}\pi^*(t) &= \frac{1}{1-\gamma} \cdot \frac{\zeta(t) + b\beta(t)}{\sigma(t)} \\ &= \frac{1}{1-\gamma} \cdot \frac{\zeta(t) + b \cdot \frac{\gamma}{\alpha}(1 - \exp(\alpha(t - T)))}{\frac{b}{\alpha}(\exp(-\alpha(T_1 - t)) - 1)}.\end{aligned}$$

Again, $\pi^*(\cdot)$ is continuous, deterministic, and therefore bounded.

In both cases one can choose δ in an appropriate way so that the optimal bond position fulfills the condition $\pi(\cdot) \le \delta$. Moreover, the respective $\pi^*(\cdot)$ is of the form

$$\pi^*(t) = \underbrace{\frac{1}{1-\gamma} \cdot \frac{\zeta(t)}{\sigma(t)}}_{\text{Merton result}} - \underbrace{\frac{\gamma}{1-\gamma} \cdot \kappa(t)}_{\text{correction term}}$$

with $\kappa(t) = \frac{T-t}{T_1-t}$ in the Ho–Lee model and $\kappa(t) = \frac{1-e^{-\alpha(T-t)}}{1-e^{-\alpha(T_1-t)}}$ in the Vasicek model. The first term coincides with the classical optimal one in Merton (1969, 1971, 1973) when the coefficients are deterministic. The second term can be interpreted as a correction term which is positive and monotonously decreasing to zero up to the terminal date T. Thus we first have a bigger, negative deviation from the classical result which vanishes at the time horizon. Moreover, the correction term increases with the investor's risk aversion because the less risky savings account will become more attractive if her risk aversion increases. This correction results from the fact that the investor tries to hedge his portfolio against the additional interest rate risk.

3rd step. At first we justify our use of Corollary 3.2, although the state process $Y = (X, r)'$ is two-dimensional: Note that the short rate process does not include the control $\pi(\cdot)$. Therefore, one can prove conditions (i) and (iii) in Definition 3.1 independently of a specified control. Consider the SDE

$$dr(t) = a(t)dt + b\,dW(t) \tag{2.8}$$

of the short rate r with $r(0) = r_0$. The coefficients meet the growth and Lipschitz conditions of the existence and uniqueness theorem for the SDE.[9] Hence (2.8) has a unique solution. Using a theorem of Krylov (1980, p. 85), we get

$$\mathrm{E}\Big(\max_{0\le s\le T} |r(s)|^{\rho}\Big) < +\infty \tag{2.9}$$

with $\rho \in \mathbb{N}$. Therefore, independently of the control under consideration, the conditions (i) and (iii) are fulfilled by the second component of the state process Y. As a result we can treat our problem as if the state process consists only of X. Note that then the wealth equation is a linear controlled SDE.

We can apply Corollary 3.2 if we are able to prove the following assumptions:

(1) $\pi^*(\cdot)$ is progressively measurable;
(2) $\pi^*(\cdot)$ meets condition (ii) in definition 3.1;
(3) $\pi^*(\cdot)$ meets condition (iii) in definition 3.1;
(4) G is a $C^{1,2}$-solution of the HJB;
(5) condition (3.12) is met.

Furthermore, the portfolio process has to lead to a positive wealth process, so

(6) $X^{\pi^*} \ge 0$.

Proof of (1). The respective solution $\pi^*(\cdot)$ is continuous and deterministic, hence progressively measurable.

Proof of (2). Property (ii) of an admissible control is met because the respective $\pi^*(\cdot)$ is bounded.

Proof of (3). By Corollary 3.1 the wealth equation (2.3) for $\pi^*(\cdot)$ has the solution

$$\begin{aligned} X^*(t) = x_0 \exp\Bigg(&\int_0^t \pi^*(s)\zeta(s)\sigma(s) + r(s) - 0.5(\pi^*(s)\sigma(s))^2\,ds \\ &+ \int_0^t \pi^*(s)\sigma(s)\,dW(s)\Bigg). \end{aligned} \tag{2.10}$$

Note that (2.9) implies

$$\mathrm{E}\left(\left|\int_0^T r(s)\,ds\right|\right) \le T \cdot \mathrm{E}\Big(\max_{0\le s\le T} |r(s)|\Big) < +\infty$$

and hence

$$\int_0^T r(s)\,ds < +\infty, \quad P - \text{a.s.}$$

The other assumptions of Corollary 3.1 are obviously met.

[9] See Fleming and Soner (1993, pp. 397f).

With an appropriate constant $K > 0$ we obtain the following estimate. (Be aware of the fact that $\pi^*(\cdot)$, $\sigma(\cdot)$, and $\zeta(\cdot)$ are bounded and that $|uv| \le u^2 + v^2$ for $u, v \in \mathbb{R}$.)

$$
\begin{aligned}
(2.11)\qquad X^*(t)^k &= x_0^k \cdot \exp\left(k \int_0^t \pi^*(s)\zeta(s)\sigma(s) + r(s) - 0.5(\pi^*(s)\sigma(s))^2 \, ds \right. \\
&\qquad \left. + k \int_0^t \pi^*(s)\sigma(s) \, dW(s) \right) \\
&\le K \cdot \exp\left(k \int_0^t r(s) \, ds + k \int_0^t \pi^*(s)\sigma(s) \, dW(s) \right) \\
&\le K \cdot \exp\left(2k \int_0^t r(s) \, ds \right) + K \cdot \exp\left(2k \int_0^t \pi^*(s)\sigma(s) \, dW(s) \right).
\end{aligned}
$$

Now consider the integral $\int_0^t r(s) \, ds$. With the form of the short rate process, in the *Ho–Lee model* we get[10]

$$
\begin{aligned}
(2.12)\qquad \int_0^t r(s) \, ds &= \int_0^t \left(r_0 + \int_0^s a(u) \, du + \int_0^s b \, dW(u) \right) ds \\
&= r_0 t + \int_0^t \int_0^s a(u) \, du ds + b \int_0^t \int_0^s dW(u) ds \\
&= \cdots + b \int_0^t (t-u) \, dW(u).
\end{aligned}
$$

The dots represent a term which is deterministic and bounded on $[0, T]$. Using the variation of constants formula for the SDE[11] in the *Vasicek model*, we obtain

$$
r(t) = e^{-\alpha t} \left(r_0 + \int_0^t e^{\alpha u} \Big(\theta(u) + b\zeta(u) \Big) \, du + \int_0^t b e^{\alpha u} \, dW(u) \right).
$$

Hence

$$
\begin{aligned}
(2.13)\qquad \int_0^t r(s) \, ds &= \int_0^t e^{-\alpha s} \left(r_0 + \int_0^s e^{\alpha u} \Big(\theta(u) + b\zeta(u) \Big) \, du \right) ds \\
&\quad + b \int_0^t \int_0^s e^{\alpha(u-s)} \, dW(u) ds \\
&= \cdots + b \int_0^t \int_u^t e^{\alpha(u-s)} \, ds dW(u).
\end{aligned}
$$

The dots represent a term which is deterministic and bounded on $[0, T]$.

In both cases the problem is reduced to find an estimate for terms of the form $\exp(\int_0^t h(s) \, dW(s))$ with a deterministic and bounded function h, namely,

$$
\begin{aligned}
&\exp\left(\int_0^t h(s) \, dW(s) \right) \\
&\qquad = \underbrace{\exp\left(\int_0^t 0.5 h^2(s) \, ds \right)}_{=const.} \cdot \underbrace{\exp\left(- \int_0^t 0.5 h^2(s) \, ds + \int_0^t h(s) \, dW(s) \right)}_{=:Z(t)}
\end{aligned}
$$

[10] See Ikeda and Watanabe (1981, pp. 117ff) for the interchange of the Lebesgue and the Ito integral.

[11] See Korn (1997, p. 313).

with

$$dZ(t) = Z(t)h(t)dW(t),$$
$$Z(0) = 1.$$

Using Krylov (1980, p. 85), we find that

$$\mathrm{E}\Big(\max_{0\le t\le T} Z(t)\Big) < +\infty.$$

Because of (2.11) and (2.12) or (2.13), respectively, $(X^*)^k$ can be estimated by processes of the same form as Z in both models. Therefore, property (3) is proved.

Proof of (4). Since the condition $G_{xx} < 0$ is met in both models, G is obviously a $C^{1,2}$-solution of the HJB equation.

Proof of (5). It is sufficent to prove that (3.12) is met by all *bounded* admissible bond positions $\pi(\cdot)$. Then the respective $\pi^*(\cdot)$ dominates all admissible bond positions. Let $(t', x', r') \in [0,T] \times \mathbb{R}^2_+ := \{y \in \mathbb{R}^2 : y > 0\}$ and $t' \le t \le T$. We consider the models separately.

Ho–Lee model. The candidate for the value function is

$$G(t,x,r) = x^\gamma \cdot \exp\left(\tfrac{1}{1-\gamma}(H_1(t) - H_1(T)) + \gamma(T-t)r\right),$$

where H_1 denotes a deterministic function which is continuously differentiable. Let K_i, $i = 1,2,3$, be appropriate constants. As H_1, π, ζ, σ, and a are bounded functions, an application of Ito's formula yields

$$\begin{aligned}
&G(t, X(t), r(t))\\
&= X(t)^\gamma \cdot \exp\left(\frac{1}{1-\gamma}(H_1(t) - H_1(T)) + \gamma(T-t)r(t)\right)\\
&= (x')^\gamma \exp\left(\gamma\int_{t'}^t \pi(s)\zeta(s)\sigma(s) + r(s) - 0.5(\pi(s)\sigma(s))^2\,ds\right.\\
&\qquad\qquad \left. +\gamma\int_{t'}^t \pi(s)\sigma(s)\,dW(s)\right)\\
&\quad \cdot \exp\left(\frac{1}{1-\gamma}(H_1(t) - H_1(T))\right) \cdot \exp\left(r(t)\gamma(T-t)\right)\\
&\le K_1 \cdot \exp\left(\gamma\int_{t'}^t r(s)\,ds + \gamma\int_{t'}^t \pi(s)\sigma(s)\,dW(s)\right) \cdot \exp(\gamma T r(t)) \cdot \exp(-\gamma t r(t))\\
&= K_1 \cdot \exp\left(\gamma\int_{t'}^t r(s)\,ds + \gamma\int_{t'}^t \pi(s)\sigma(s)\,dW(s)\right) \cdot \exp\left(\gamma T\int_{t'}^t dr(s)\right)\\
&\quad \cdot \exp\left(-\gamma\int_{t'}^t s\,dr(s) - \gamma\int_{t'}^t r(s)\,ds\right)\\
&= K_1 \cdot \exp\left(\gamma\int_{t'}^t \pi(s)\sigma(s)\,dW(s)\right) \cdot \exp\left(\gamma\int_{t'}^t (T-s)(a(s)\,ds + b\,dW(s))\right)\\
&\le K_2 \cdot \exp\left(\gamma\int_{t'}^t \pi(s)\sigma(s) + b(T-s)\,dW(s)\right)\\
&\le K_3 \cdot \exp\left(\gamma\int_{t'}^t \pi(s)\sigma(s) + b(T-s)dW(s)\right.\\
&\qquad\qquad \left. -0.5\gamma^2\int_{t'}^t \Big(\pi(s)\sigma(s) + b(T-s)\Big)^2 ds\right)\\
&=: K_3 \cdot Z(t),
\end{aligned}$$

where Z is the unique solution of

$$dZ(t) = Z(t)\Big(\gamma(\pi(t)\sigma(t) + b(T-t))\Big)dW(t) \quad \text{mit} \quad Z(t') = 1.$$

Using Krylov (1980, p. 85), we arrive at

$$\mathrm{E}\left(\sup_{t\in[t',T]} |G(t, X(t), r(t))|^2\right) \le K_3 \cdot \mathrm{E}\left(\sup_{t\in[t',T]} |Z(t)|^2\right) < \infty.$$

Hence we have just proved (3.12) in the Ho–Lee model.

Vasicek model. Our candidate for the value function is

$$G(t,x,r) = x^\gamma \cdot \exp\left(\frac{1}{1-\gamma}(H_2(t) - H_2(T)) + \frac{\gamma}{\alpha}(1 - \exp(\alpha(t-T)))r\right),$$

where H_2 is a continuously differentiable and deterministic function. With appropriate constants K_i, $i = 1, \dots, 6$, we find that

$$\begin{aligned}
&G(t, X(t), r(t)) \\
&= X(t)^\gamma \cdot \exp\left(\frac{1}{1-\gamma}(H_2(t) - H_2(T)) + \frac{\gamma}{\alpha}(1 - \exp(\alpha(t-T)))r(t)\right) \\
&\le K_1 \cdot X(t)^\gamma \cdot \exp\left(\frac{\gamma}{\alpha}(1 - \exp(\alpha(t-T)))r(t)\right) \\
&\le K_2 \cdot \exp\left(\gamma \int_{t'}^t \pi(s)\zeta(s)\sigma(s) + r(s) - 0.5(\pi(s)\sigma(s))^2\, ds \right. \\
&\qquad\qquad \left. + \gamma \int_{t'}^t \pi(s)\sigma(s)\, dW(s)\right) \cdot \exp\left(\frac{\gamma}{\alpha}(1 - \exp(\alpha(t-T))) \cdot r(t)\right) \\
&\le K_3 \cdot \exp\left(\gamma \int_{t'}^t r(s)\, ds + \gamma \int_{t'}^t \pi(s)\sigma(s)\, dW(s)\right) \cdot \exp\left(\frac{\gamma}{\alpha} r(t)\right) \\
&\quad \cdot \exp\left(-\frac{\gamma}{\alpha} \exp(\alpha(t-T)) \cdot r(t)\right).
\end{aligned}$$

With the definition $f^h(t,r) := \exp(\alpha(t-T)) \cdot r$ an application of Ito's formula yields

$$\begin{aligned}
&f^h(t, r(t)) \\
&= f^h(t', r') + \int_{t'}^t \alpha \exp(\alpha(s-T)) r(s)\, ds + \int_{t'}^t \exp(\alpha(s-T))\, dr(s) \\
&= f^h(t', r') + \int_{t'}^t \exp(\alpha(s-T)) \cdot (\theta(s) + b\zeta(s))\, ds + \int_{t'}^t b\, \exp(\alpha(s-T))\, dW(s).
\end{aligned}$$

Hence, by virtue of the stochastic integral equation of the short rate, we have

$$\begin{aligned}
&G(t, X(t), r(t)) \\
&\qquad \le K_4 \cdot \exp\left(\gamma \int_{t'}^t r(s)\, ds + \gamma \int_{t'}^t \pi(s)\sigma(s)\, dW(s)\right) \cdot \exp\left(\frac{\gamma}{\alpha} r(t)\right) \\
&\qquad\quad \cdot \exp\left(-\frac{\gamma}{\alpha} \int_{t'}^t b\, \exp(\alpha(s-T))\, dW(s)\right)
\end{aligned}$$

$$
\begin{aligned}
&= K_4 \cdot \exp\left(\gamma \int_{t'}^{t} r(s)\,ds + \gamma \int_{t'}^{t} \pi(s)\sigma(s)\,dW(s)\right)\\
&\quad \cdot \exp\left(\frac{\gamma}{\alpha} r' + \frac{\gamma}{\alpha}\int_{t'}^{t} (\theta(s) - \alpha r(s) + b\zeta(s))\,ds + \frac{\gamma}{\alpha}\int_{t'}^{t} b\,dW(s)\right)\\
&\quad \cdot \exp\left(-\frac{\gamma}{\alpha}\int_{t'}^{t} b\,\exp(\alpha(s-T))\,dW(s)\right)\\
&\le K_5 \cdot \exp\left(\int_{t'}^{t} \gamma\pi(s)\sigma(s) + \frac{\gamma}{\alpha} b\Big(1 - \exp(\alpha(s-T))\Big)\,dW(s)\right)\\
&\le K_6 \cdot \exp\left(\int_{t'}^{t} \gamma\pi(s)\sigma(s) + \frac{\gamma}{\alpha} b\Big(1 - \exp(\alpha(s-T))\Big)\,dW(s)\right.\\
&\qquad \left. - \int_{t'}^{t} 0.5\left[\gamma\pi(s)\sigma(s) + \frac{\gamma}{\alpha} b\Big(1 - \exp(\alpha(s-T))\Big)\right]^2 ds\right)\\
&=: K_6 \cdot \tilde{Z}(t).
\end{aligned}
$$

Since the process $\tilde{Z}$ has the same properties as Z in the Ho–Lee model, an analogous argument leads to (3.12).

Proof of (6). By virtue of (2.10), we have $X^* \ge 0$.

The following theorem summarizes our results.

THEOREM 2.1 (bond portfolio problem). *The optimal portfolio processes in the above bond portfolio problems are given by*

$$\pi^*(t) = \frac{1}{1-\gamma} \cdot \frac{\zeta(t)}{\sigma(t)} - \frac{\gamma}{1-\gamma} \cdot \kappa(t)$$

with

(a) *Ho–Lee case:* $\kappa(t) = \frac{T-t}{T_1-t}$,

(b) *Vasicek case:* $\kappa(t) = \frac{1-e^{-\alpha(T-t)}}{1-e^{-\alpha(T_1-t)}}$.

2.2. A mixed stock and bond portfolio problem. In this subsection we assume that the investor can put his money on a savings account, in a stock, or in a bond with maturity $T_1 > T$. The dynamics of these assets are given by

$$
\begin{aligned}
dB(t) &= B(t)r(t)dt,\\
dS(t) &= S(t)\Big[\mu_S(t)dt + \sigma_S(t)dW_S(t) + \sigma_{SB}(t)dW_B(t)\Big],\\
dP(t) &= P(t)\Big[\underbrace{(r(t) + \zeta_B(t)\sigma_B(t))}_{=:\mu_B(t)}dt + \sigma_B(t)dW_B(t)\Big],
\end{aligned}
$$

where (W_S, W_B) is a two-dimensional Brownian motion and where, for ease of notation, we write $P(t)$ instead of $P(t, T_1)$. In our model the stock price depends on two risk factors: The first factor W_S contains the specific risk of the stock, and the second W_B comes from the stochastic interest rate model.

In Merton's portfolio problem we can split up the (deterministic) drift μ_S of the stock into a liquidity premium (LP) and an excess return, which should be interpreted as RP in this context:[12]

$$\mu_S = \underbrace{r}_{\text{LP}} + \underbrace{\mu_S - r}_{\text{RP}}.$$

[12]There is no uniform use of the words excess return, RP, and market price of risk. Apart from the above interpretation of the drift, throughout the paper we denote $\lambda = \mu - r$ as excess return, $\frac{\lambda}{\sigma}$ as RP, and $\frac{\lambda}{\sigma^2}$ as market price of risk.

The drift of the stock S under consideration can also be

$$\mu_S(t) = r(t) + \underbrace{\mu_S(t) - r(t)}_{=:\lambda_S(t)},$$

where λ_S denotes the RP of the stock

In the following, we assume that the excess return $\lambda_S(\cdot)$ of the stock is deterministic and continuous. This implies that $\lambda_S(\cdot)$ is progressively measurable and bounded on $[0,T]$. Furthermore, assume that the coefficients $\sigma_S(\cdot)$, $\sigma_{SB}(\cdot)$, and $\sigma_B(\cdot)$ are deterministic and continuous. In addition, let $\sigma_S(\cdot)$ and $\sigma_B(\cdot)$ be bounded away from zero.

As before, we consider both a Ho–Lee and a Vasicek model:

$$dr(t) = a(t)dt + bdW_B(t)$$

with $a(t) = \tilde{a}(t) + b\zeta(t)$ in the Ho–Lee model and $a(t) = \theta(t) - \alpha r(t) + b\zeta(t)$ in the Vasicek model.

Moreover, we have $\sigma_B(t) = -b(T_1 - t)$ in the Ho–Lee model and $\sigma_B(t) = \frac{b}{\alpha}(\exp(-\alpha(T_1 - t)) - 1)$ in the Vasicek model.

In this framework the wealth equation (2.1) has the following form:

$$\begin{aligned} dX(t) = X(t)\Big[&(\pi_S(t)\lambda_S(t) + \pi_B(t)\lambda_B(t) + r(t))dt \\ &+\pi_S(t)\sigma_S(t)dW_S(t) + (\pi_S(t)\sigma_{SB}(t) + \pi_B(t)\sigma_B(t))dW_B(t)\Big], \end{aligned}$$

where $\lambda_B(t) := \mu_B(t) - r(t)$ and $\pi := (\pi_S, \pi_B)$.

Using the notations of (3.1) in the appendix, we have

$$\begin{aligned} Y(t) &= (X(t), r(t))', \\ \Lambda(t,x,r,\pi) &= (x(\pi_S\lambda_S + \pi_B\lambda_B + r), a)', \\ \Sigma(t,x,r,\pi) &= \begin{pmatrix} x\pi_S\sigma_S & x(\pi_S\sigma_{SB} + \pi_B\sigma_B) \\ 0 & b \end{pmatrix}, \\ \Sigma^*(t,x,r,\pi) &= \begin{pmatrix} x^2(\pi_S^2\sigma_S^2 + (\pi_S\sigma_{SB} + \pi_B\sigma_B)^2) & bx(\pi_S\sigma_{SB} + \pi_B\sigma_B) \\ bx(\pi_S\sigma_{SB} + \pi_B\sigma_B) & b^2 \end{pmatrix}, \\ A^\pi G(t,x,r) &= G_t + 0.5x^2(\pi_S^2\sigma_S^2 + (\pi_S\sigma_{SB} + \pi_B\sigma_B)^2)G_{xx} + 0.5b^2G_{rr} \\ &\quad + bx(\pi_S\sigma_{SB} + \pi_B\sigma_B)G_{xr} + x(\pi_S\lambda_S + \pi_B\lambda_B + r)G_x + aG_r. \end{aligned}$$

Hence we have to solve the following HJB equation:

$$\begin{aligned} \sup_{|\pi| \le \delta} A^\pi G(t,x,r) &= 0, \\ G(T,x,r) &= x^\gamma. \end{aligned}$$

This will again be done by the three-step-algorithm.

1*st step.* Assuming $G_{xx} < 0$, we calculate the candidates for the optimal portfolio positions

$$\pi_S^* = -\underbrace{(\eta_S - \tfrac{\sigma_{SB}}{\sigma_B}\eta_{BS})}_{=:\hat{\eta}_S} \cdot \frac{G_x}{xG_{xx}}, \tag{2.14}$$

$$\pi_B^* = -\underbrace{\left((1+\tfrac{\sigma_{SB}^2}{\sigma_S^2})\eta_B - \tfrac{\sigma_{SB}}{\sigma_B}\eta_S\right)}_{=:\hat{\eta}_B} \cdot \frac{G_x}{xG_{xx}} - \frac{b}{\sigma_B} \cdot \frac{G_{xr}}{xG_{xx}} \tag{2.15}$$

with $\eta_S := \lambda_S/\sigma_S^2$, $\eta_B := \lambda_B/\sigma_B^2$ and $\eta_{BS} := \lambda_B/\sigma_S^2$.

2*nd step.* Inserting $\pi_S^*(t,x,r;G)$ and $\pi_B^*(t,x,r;G)$ in the HJB equation yields the PDE

$$\begin{aligned} 0 = G_t G_{xx} + (\underbrace{0.5\sigma_S^2\hat{\eta}_S^2 + 0.5(\sigma_{SB}\hat{\eta}_S + \sigma_B\hat{\eta}_B)^2 - \lambda_S\hat{\eta}_S - \lambda_B\hat{\eta}_B}_{=:\tilde{\zeta}(t)})G_x^2 \\ -0.5b^2G_{xr}^2 + 0.5b^2G_{rr}G_{xx} - b\tfrac{\lambda_B}{\sigma_B}G_xG_{xr} + xrG_xG_{xx} + aG_rG_{xx} \end{aligned}$$

with $G(T,x,r) = x^\gamma$. This PDE is of the same form as the corresponding PDE (2.5) above.[13] Note that $\tilde{\zeta}$, in analogy to ζ in (2.5), is a continuous and deterministic function. Therefore, in the Ho–Lee model we get

$$G(t,x,r) = x^\gamma \cdot \exp\left(\frac{1}{1-\gamma}(H_3(t) - H_3(T)) + \gamma(T-t)r\right)$$

and in the Vasicek model

$$G(t,x,r) = x^\gamma \cdot \exp\left(\frac{1}{1-\gamma}(H_4(t) - H_4(T)) + \frac{\gamma}{\alpha}(1-\exp(\alpha(t-T)))r\right),$$

with appropriate continuously differentiable functions H_3 and H_4, respectively. Insertion into (2.14) and (2.15) yields in *both* models for the optimal stock and bond position

$$\begin{aligned} \pi_S^*(t) &= \frac{1}{1-\gamma} \cdot \left(\eta_S(t) - \frac{\sigma_{SB}(t)}{\sigma_B(t)}\eta_{BS}(t)\right) \\ &= \frac{1}{1-\gamma} \cdot \hat{\eta}_S(t), \\ \pi_B^*(t) &= \frac{1}{1-\gamma} \cdot \left(\left(1 + \frac{\sigma_{SB}^2(t)}{\sigma_S^2(t)}\right)\eta_B(t) - \frac{\sigma_{SB}(t)}{\sigma_B(t)}\eta_S(t) - \gamma \cdot \kappa(t)\right) \\ &= \frac{1}{1-\gamma} \cdot \left(\hat{\eta}_B(t) - \gamma \cdot \kappa(t)\right), \end{aligned}$$

where $\kappa(t) = \frac{T-t}{T_1-t}$ in the *Ho–Lee model* and $\kappa(t) = \frac{1-e^{-\alpha(T-t)}}{1-e^{-\alpha(T_1-t)}}$ in the *Vasicek model*.

Both positions are continuous and deterministic processes and are hence bounded.

3*rd step.* With the same argument as in subsection 2.1 we can apply Corollary 3.2. Therefore, in both models we must check the following assumptions:

(1) $\pi^*(\cdot)$ is progressively measurable;

(2) $\pi^*(\cdot)$ meets condition (ii) in Definition 3.1;

[13] One will obtain the PDE (2.5) if $\lambda_S \equiv 0$, $\sigma_S \equiv 0$, and $\sigma_{SB} \equiv 0$.

(3) $\pi^*(\cdot)$ meets condition (iii) in Definition 3.1;
(4) G is a $C^{1,2}$-solution of the HJB equation;
(5) condition (3.12) is met;
(6) $X^{\pi^*} \geq 0$.

Note that $\pi^* := (\pi_S^*, \pi_B^*)'$.

Conditions (1) and (2) are met because in both models $\pi^*(\cdot)$ is a continuous and deterministic process. Obviously, (4) is fulfilled. Condition (6) is met since variation of constants leads to

$$\begin{aligned} X(t) = x_0 \exp\Bigg(& \int_0^t \pi_S(s)\lambda_S(s) + \pi_B(s)\lambda_B(s) + r(s) \\ & - 0.5\Big((\pi_S(s)\sigma_S(s))^2 + (\pi_S(s)\sigma_{SB}(s) + \pi_B(s)\sigma_B(s))^2\Big)\, ds \\ & + \int_0^t \pi_S(s)\sigma_S(s)\, dW_S(s) + \int_0^t \pi_S(s)\sigma_{SB}(s) + \pi_B(s)\sigma_B(s)\, dW_B(s)\Bigg) \end{aligned}$$

for an admissable control $\pi(\cdot)$. Furthermore, since the wealth process has the same properties as in subsection 2.1, we can prove (3) and (5) using the analogous arguments.

The following theorem summarizes our results.

THEOREM 2.2 (mixed portfolio problem). *The optimal portfolio processes in the above mixed portfolio problem are given by*

$$\pi_S^*(t) = \frac{1}{1-\gamma} \cdot \left(\underbrace{\eta_S(t) - \frac{\sigma_{SB}(t)}{\sigma_B(t)}\eta_{BS}(t)}_{=:\hat{\eta}_S} \right), \qquad \textit{(stock)}$$

$$\pi_B^*(t) = \frac{1}{1-\gamma} \cdot \left(\underbrace{\left(1 + \frac{\sigma_{SB}^2(t)}{\sigma_S^2(t)}\right)\eta_B(t) - \frac{\sigma_{SB}(t)}{\sigma_B(t)}\eta_S(t)}_{=:\hat{\eta}_B} - \gamma \cdot \kappa(t) \right) \qquad \textit{(bond)}$$

with

(a) *Ho–Lee case:* $\kappa(t) = \frac{T-t}{T_1-t}$,
(b) *Vasicek case:* $\kappa(t) = \frac{1-e^{-\alpha(T-t)}}{1-e^{-\alpha(T_1-t)}}$.

Considering the optimal positions the analogy to the pure bond problem becomes clear: The variables $\hat{\eta}_S$ and $\hat{\eta}_B$ can be interpreted as modified market prices of risk, where both are weighted differences of η_S and η_{BS} or η_B and η_S, respectively. In the optimal stock position the market price of risk of the stock is corrected by η_{BS}, which stands for the market price of risk of the bond with respect to the stock.

Similarly, the market price of risk of the bond contains a correction of the optimal bond position by the market price of risk of the stock. Both these corrections are plausible ones as an increase of the market price of risk of the bond makes stock investment less attractive and vice versa. Apart from this remark, the interpretation of the bond part as given in section 2.1 remains valid.

Furthermore, we will get the optimal bond position of subsection 2.1 if we choose $\sigma_S \equiv 0$ and $\sigma_{SB} \equiv 0$ in $\pi_B(\cdot)$.

However, as the detailed proof of Theorem 2.2 would have been much more complicated than that of Theorem 2.1, we have decided to only present the one for Theorem 2.1 as it contains the main ideas.

3. Appendix. In this appendix we will present the technical results and details which enabled us to solve the foregoing portfolio problems by stochastic control methods. Let, therefore, $(\Omega, \mathcal{F}, P)$ be a complete probability space. Assume that on this space an m-dimensional Brownian motion $\{(W(t), \mathcal{F}_t)\}_{t\in[0,\infty)}$ is defined with $\{\mathcal{F}_t\}_{t\in[0,\infty)}$ being the Brownian filtration. All adapted or progressively measurable processes are adapted or progressively measurable with respect to the Brownian filtration. Let, further, $|\cdot|$ denote the Euclidean norm or the operator norm, respectively.

As usual we will look at a state process given by a controlled SDE of the form

$$dY(t) = \Lambda(t, Y(t), u(t))dt + \Sigma(t, Y(t), u(t))dW(t) \tag{3.1}$$

with initial value of $Y(t_0) = y_0$ and a d-dimensional control process $u(\cdot)$. Let $[t_0, t_1]$ with $0 \le t_0 < t_1 < \infty$ be the relevant time interval. A control strategy $u(\cdot)$ (for short, control) is a progressively measurable process with $u(t) \in U$ for all $t \in [t_0, t_1]$, where the set $U \subset \mathbb{R}^d$, $d \in \mathbb{N}$, is assumed to be closed. Further, let $Q_0 := [t_0, t_1) \times \mathbb{R}^n$, $n \in \mathbb{N}$. The coefficient functions

$$\Lambda : \bar{Q}_0 \times U \to \mathbb{R}^n,$$
$$\Sigma : \bar{Q}_0 \times U \to \mathbb{R}^{n,m},$$

$m \in \mathbb{N}$, are all assumed to be continuous. Further, for all $v \in U$, let $\Lambda(\cdot,\cdot,v)$ and $\Sigma(\cdot,\cdot,v)$ be in $C^1(\bar{Q}_0)$. We then define the following.

DEFINITION 3.1 (admissible control). *A control $\{(u(t), \mathcal{F}_t)\}_{t\in[t_0,t_1]}$ will be called admissible*[14] *if*

(i) *for all $y_0 \in \mathbb{R}^n$ the corresponding controlled SDE* (3.1) *with initial condition $Y(t_0) = y_0$ admits a pathwise unique solution $\{Y^u(t)\}_{t\in[t_0,t_1]}$;*

(ii) *for all $k \in \mathbb{N}$ the integrability condition*

$$\mathrm{E}\left(\int_{t_0}^{t_1} |u(s)|^k \, ds\right) < \infty$$

is satisfied;

(iii) *the corresponding state process Y^u satisfies*

$$\mathrm{E}^{t_0,y_0}\left(\sup_{t\in[t_0,t_1]} |Y^u(t)|^k\right) < \infty.$$

Let $\mathcal{A}(t_0, y_0)$ denote the set of all admissible controls corresponding to the initial condition $(t_0, y_0) \in Q$.

In the following, the above definition will prove to be extremely useful when we have to overcome some technical difficulties which have their origin in the fact that the wealth equation does not satisfy the usual Lipschitz conditions needed in the standard verification theorems of stochastic control.

To ensure existence and uniqueness of the solution of the controlled SDE (3.1), one typically requires the following Lipschitz and growth conditions for the coefficient functions which imply that controls with property (ii) are already admissible (i.e.,

[14] This definition is more restrictive than the usual one as, e.g., given in Fleming and Rishel (1975, p. 156). However, due to the special form of our control problem, all (optimal) controls in this paper will satisfy the more restrictive requirements of our definition.

they also satisfy properties (i) and (iii)).[15] With a constant $C > 0$ these conditions are

$$|\Lambda_t| + |\Lambda_y| \le C, \tag{3.2}$$
$$|\Sigma_t| + |\Sigma_y| \le C,$$
$$|\Lambda(s,y,v)| \le C(1 + |y| + |v|), \tag{3.3}$$
$$|\Sigma(s,y,v)| \le C(1 + |y| + |v|)$$

for all $s \in [t_0, t_1]$, $y \in \mathbb{R}$, and $v \in U$.

Typically, in our applications the conditions (3.2) and (3.3) will not be satisfied. On the other hand we have only to deal with linear controlled SDEs. This will imply that requirement (ii) on an admissible control already ensures requirement (i) too.

COROLLARY 3.1 (variation of constants). *Let $(t_0, y_0) \in Q$, and let $A_1^{(j)}$, $j = 1, \dots, d$, A_2, $B_1^{(i,j)}$, $i = 1, \dots, m$, $j = 1, \dots, d$, $B_2^{(i)}$, $i = 1, \dots, m$ be progressively measurable real-valued processes satisfying the integrability conditions*

$$\int_{t_0}^{t_1} |A_2(s)|\, ds < \infty \quad P\text{-a.s., } t \ge 0,$$
$$\int_{t_0}^{t_1} \left(\sum_{j=1}^{d} A_1^{(j)}(s)^2 + \sum_{i=1}^{m} B_2^{(i)}(s)^2 \right) ds < \infty \quad P\text{-a.s., } t \ge 0,$$
$$\int_{t_0}^{t_1} \left(\sum_{i=1}^{m} \sum_{j=1}^{d} B_1^{(i,j)}(s)^4 \right) ds < \infty \quad P\text{-a.s., } t \ge 0.$$

Further, let $u(\cdot)$ be a control with property (ii) *of Definition* 3.1. *Then the linear controlled SDE*

$$dY^u(t) = Y^u(t)\Big[(A_1(t)'u(t) + A_2(t))dt + (B_1(t)u(t) + B_2(t))'dW(t)\Big] \tag{3.4}$$

admits the Lebesgue $\bigotimes P$ unique solution

$$Y^u(t) = y_0 \cdot \exp\left(\int_{t_0}^{t_1} \Big(A_1(s)'u(s) + A_2(s) - 0,5|B_1(s)u(s) + B_2(s)|^2 \Big) ds + \int_{t_0}^{t_1} \Big(B_1(s)u(s) + B_2(s) \Big)' dW(s) \right).$$

If we consider only bounded admissible controls, then the following conditions are sufficient:

$$\int_{t_0}^{t_1} \left(\sum_{j=1}^{d} |A_1^{(j)}(s)| + |A_2(s)| \right) ds < \infty \quad P\text{-a.s., } t \ge 0,$$
$$\int_{t_0}^{t_1} \left(\sum_{i=1}^{m} \sum_{j=1}^{d} B_1^{(i,j)}(s)^2 + \sum_{i=1}^{m} B_2^{(i)}(s)^2 \right) ds < \infty \quad P\text{-a.s., } t \ge 0.$$

Proof of Corollary 3.1. The integrability assumptions together with property (ii) of an admissible control imply the requirements of the variation of constants formula given in Korn (1997). Applying it implies all assertions of the corollary. □

Consequently, for our applications it will be enough to verify properties (ii) and (iii) to obtain admissibility of a control. From now on, controlled SDEs (3.4) with

[15] See Fleming and Soner (1993, p. 398).

coefficients satisfying the conditions of Corollary (3.1) will be referred to as linear controlled SDEs.

We will now formulate a standard verification theorem and afterwards derive a version suitable for our applications by modifying the relevant parts of the proof of the standard theorem. Therefore, we look at the following setting: Let $\mathcal{O} \subset \mathbb{R}^n$ be an open subset of $\mathbb{R}^n$. In the case of $\mathcal{O} \neq \mathbb{R}^n$ we additionally assume that its boundary $\partial\mathcal{O}$ is a compact $(n-1)$-dimensional C^3-manifold. In analogy to Q_0 we define $Q := [t_0, t_1) \times \mathcal{O}$. Further, let

$$\tau := \inf\{t \in [t_0, t_1] : (t, Y(t)) \notin Q\}$$

denote the exit time of Y from $\mathcal{O}$. Hence we have

$$(\tau, Y(\tau)) \in \partial^* Q := ([t_0, t_1) \times \partial\mathcal{O}) \cup (\{t_1\} \times \bar{\mathcal{O}}).$$

We now consider continuous, real-valued functions L and Ψ that satisfy the polynomial growth conditions

$$|L(t, y, v)| \leq C(1 + |y|^k + |v|^k), \tag{3.5}$$

$$|\Psi(t, y)| \leq C(1 + |y|^k) \tag{3.6}$$

on $\bar{Q} \times U$ and $\bar{Q}$ for suitable constants $k \in \mathbb{N}$ and $C > 0$. Here L and Ψ model the running and the terminal utility resulting from the control and the position of the controlled process, respectively. It will be our goal to determine an admissible control $u(\cdot)$ such that for each initial value (t_0, y_0) the utility functional

$$J(t_0, y_0; u) := \mathrm{E}^{t_0, y_0}\left(\int_{t_0}^{\tau} L(s, Y^u(s), u(s))\, dt + \Psi(\tau, Y^u(\tau))\right)$$

will be maximized; i.e., we want to solve $\max_{u \in \mathcal{A}(t_0, y_0)} J(t_0, y_0; u)$.

Therefore, define the value function

$$V(t, y) := \sup_{u \in \mathcal{A}(t,y)} J(t, y; u), \quad (t, y) \in Q.$$

For each function $G \in C^{1,2}(Q)$ and $(t, y) \in Q$, $v \in U$, we consider the following differential operator:

$$A^v G(t, y) := G_t(t, y) + 0,5 \sum_{i,j=1}^{n} \Sigma^*_{ij}(t, y, v) \cdot G_{y_i y_j}(t, y) + \sum_{i=1}^{n} \Lambda_i(t, y, v) \cdot G_{y_i}(t, y)$$

with $\Sigma^* := \Sigma\Sigma'$. Then, we have the following theorem.[16]

THEOREM 3.1 (verification theorem). *Let the conditions* (3.2) *and* (3.3) *on the coefficient functions of the controlled SDE* (3.1) *be satisfied. Further, assume conditions* (3.5) *and* (3.6). *Let G be a function with the following properties:*

(a) *We have*

$$G \in C^{1,2}(Q) \cap C(\bar{Q}), \tag{3.7}$$

$$|G(t, y)| \leq K(1 + |y|^k) \tag{3.8}$$

for suitable $K > 0$ and $k \in \mathbb{N}$.

[16] See Fleming and Soner (1993, p. 163).

(b) *G solves the HJB equation:*

$$\sup_{v\in U}\Big\{A^vG(t,y)+L(t,y,v)\Big\}=0, \qquad (t,y)\in Q, \tag{3.9}$$

$$G(t,y)=\Psi(t,y), \quad (t,y)\in\partial^*Q. \tag{3.10}$$

Then we obtain the following result:

(i) *$G(t,y)\geq J(t,y;u)$ for all $(t,y)\in Q$ and $u(\cdot)\in\mathcal{A}(t,y)$.*

(ii) *If for $(t,y)\in Q$ there exists a control $u^*(\cdot)\in\mathcal{A}(t,y)$ with*

$$u^*(s)\in\arg\max_{v\in U}\Big(A^vG(s,Y^*(s))+L(s,Y^*(s),v)\Big) \tag{3.11}$$

for all $s\in[t,\tau]$, where Y^* is the solution of the controlled SDE corresponding to $u^*(\cdot)$, then we have

$$G(t,y)=V(t,y)=J(t,y;u^*),$$

i.e., $u^*(\cdot)$ is an optimal control and G coincides with the value function.

Besides conditions (3.2) and (3.3), the growth condition (3.8) is not satisfied in our applications either. Thus we need to modify the above verification result in a suitable way.

Corollary 3.2 (to the verification theorem). *Consider a linear controlled SDE with coefficients satisfying the assumptions of Corollary* 3.1. *Assume, further, that the functions L and ψ satisfy the conditions* (3.5) *and* (3.6). *Finally, let the function $G\in C^{1,2}(Q)\cap C(\bar{Q})$ be a solution to the HJB equation* (3.9) *with boundary condition* (3.10). *Assume that for all $(t,y)\in Q$ and all admissible controls $u(\cdot)\in\mathcal{A}(t,y)$ there exists a $\rho>1$ such that we have*

$$\mathrm{E}\left(\sup_{s\in[t,t_1]}|G(s,Y(s))|^\rho\right)<\infty. \tag{3.12}$$

Then assertions (i) *and* (ii) *of the verification theorem are valid.*

Proof of Corollary 3.2. Looking at the proof of the verification theorem as given in Fleming and Soner (1993, pp. 163f), we realize the following:

(i) Conditions (3.2) and (3.3) ensure the existence and uniqueness of a solution of the controlled SDE for controls with property (ii) of Definition 3.1. We can then apply the Ito formula to obtain

$$\begin{aligned}G(\theta,Y(\theta))-G(t,y)-\int_t^\theta A^{u(s)}G(s,Y(s))\,ds\\ =\int_t^\theta G_y(s,Y(s))\cdot\Sigma(s,Y(s),u(s))\,dW(s)\end{aligned} \tag{3.13}$$

which corresponds to relation (3.9) in Fleming and Soner (1993).

(ii) The growth condition (3.3) is used to prove the relation

$$\mathrm{E}^{t,y}\left(\int_t^\theta G_y(s,Y(s))\cdot\Sigma(s,Y(s),u(s))\,dW(s)\right)=0 \tag{3.14}$$

for bounded $\mathcal{O}$. (This corresponds to $\mathrm{E}_{tx}M(\theta)=0$ for bounded $\mathcal{O}$ in the notation of Fleming and Soner (1993).)

(iii) The growth condition (3.8) is used to show the uniform integrability of $\{G(\theta_p, Y(\theta_p))\}_p$, where θ_p are stopping times with $t \le \theta_p \le t_1$. (In the notation of Fleming and Soner (1993) this corresponds to the uniform integrability of $\{W(\theta_p, x(\theta_p))\}_p$. There, one also finds the exact definition of the stopping times θ_p, which is irrelevant for our argumentation.)

We now demonstrate that we also have these three properties under the assumptions of our corollary:

(i) For admissible controls the linear controlled SDE admits a unique solution which is explicitly given in Corollary 3.1. Of course, we can apply the Ito formula to such solutions. Thus relation (3.13) remains valid.

(ii) To show property (3.14) note that the diffusion coefficient of the linear controlled SDE is $\Sigma(t, y, v) = y(B_1(t)v + B_2(t))$. As in Fleming and Soner (1993), we look at a bounded set $\mathcal{O}$ and obtain the following estimate for an admissible control $u(\cdot)$:

$$\begin{aligned}\int_{t_0}^{t_1} |\Sigma(s, Y(s), u(s))|^2 \, ds &= \int_{t_0}^{t_1} |Y(s)(B_1(s)u(s) + B_2(s))|^2 \, ds \\ &\le \sup_{s\in[t,t_1]} |Y(s)|^2 \int_{t_0}^{t_1} (|B_1(s)u(s)| + |B_2(s)|)^2 \, ds \\ &\le 2\,\mathrm{diam}(\mathcal{O}) \int_{t_0}^{t_1} |B_1(s)u(s)|^2 + |B_2(s)|^2 \, ds \\ &\le 2\,\mathrm{diam}(\mathcal{O}) \int_{t_0}^{t_1} |B_1(s)|^4 + |u(s)|^4 + |B_2(s)|^2 \, ds.\end{aligned}$$

Here we have made multiple uses of $2|vw| \le v^2 + w^2$ for $v, w \in \mathbb{R}$. Due to property (ii) of an admissible control and the integrability conditions of the coefficients of the linear controlled SDE, we obtain

$$\mathrm{E}^{t,y}\left(\int_t^\theta |G_y(s, Y(s)) \cdot \Sigma(s, Y(s), u(s))|^2 \, ds\right) < \infty$$

and thus (3.14).

(iii) Condition (3.12) implies uniform integrability of $\{G(\theta_p, Y(\theta_p))\}_p$. □

REFERENCES

E. Canestrelli and S. Pontini (1998), *Inquiries on the Applications of Multidimensional Stochastic Processes to Financial Investments*, Working paper, Dipartimento di Matematica Applicata ed Informatica, Université Ca' Foscari, Venezia, Italy.

J. C. Cox, J. E. Ingersoll, Jr., and S. A. Ross (1985), *A theory of the term structure of interest rates*, Econometrica, 53, pp. 385–407.

D. Duffie (1992), *Dynamic Asset Pricing Theory*, Princeton University Press, Princeton.

W. H. Fleming and R. W. Rishel (1975), *Deterministic and Stochastic Optimal Control*, Springer, New York.

W. H. Fleming and H. M. Soner (1993), *Controlled Markov Processes and Viscosity Solutions*, Springer, New York.

T. S. Y. Ho and S.-B. Lee (1986), *Term structure movements and pricing interest contingent claims*, J. Finance, 41, pp. 1011–1029.

J. Hull and A. White (1990), *Pricing interest-rate-derivative securities*, Rev. Financial Studies, 3, pp. 573–592.

N. Ikeda and S. Watanabe (1981), *Stochastic Differential Equations and Diffusion Processes*, North Holland, New York.

C. Klüppelberg and R. Korn (1998), *Optimal portfolios with bounded value-at-risk*, in Berichte zur Stochastik und verwandten Gebieten, Johannes Gutenberg-Universität Mainz.
R. Korn (1997), *Optimal Portfolios*, World Scientific, Singapore.
N. Krylov (1980), *Controlled Diffusion Processes*, Springer, Berlin.
R. C. Merton (1969), *Lifetime portfolio selection under uncertainty: The continuous case*, Rev. Econom. Statistics, 51, pp. 247–257.
R. C. Merton (1971), *Optimal consumption and portfolio rules in a continuous-time model*, J. Econom. Theory, 3, pp. 373–413.
R. C. Merton (1973), *Erratum*, J. Econom. Theory, 6, pp. 213–214.
R. C. Merton (1990), *Continuous-Time Finance*, Basil Blackwell, Cambridge MA.
M. Musiela and M. Rutkowski (1997), *Martingale Methods in Financial Modelling*, Springer, Berlin.
C. Sørensen (1999), *Dynamic asset allocation and fixed income management*, J. Financial and Quantitative Analysis, 34, pp. 513–531.
O. Vasicek (1977), *An equilibrium characterisation of the term structure*, J. Financial Economics, 5, pp. 177–188.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 4, pp. 1270–1279

ON THE BOUNDED SLOPE CONDITION AND THE VALIDITY OF THE EULER LAGRANGE EQUATION*

ARRIGO CELLINA†

Abstract. Under the bounded slope condition on the boundary values of a minimization problem for a functional of the gradient of u, we show that a continuous minimizer w is, in fact, Lipschitzian. An application of this result to prove the validity of the Euler Lagrange equation for w is presented.

Key words. bounded slope condition, weak maximum principle, Euler Lagrange equation

AMS subject classification. 49K20

PII. S0363012999354661

1. Introduction. The bounded slope condition was introduced by Hartman and Nirenberg [5] and, in a variational context, by Stampacchia [9], with the purpose of obtaining pointwise bounds a.e. on the norm of the gradient $\nabla u(x)$ of a solution u to a minimum problem of the form

$$(P) \qquad \text{minimize} \int_\Omega f(\nabla u(x))\,dx \quad \text{on} \quad u - u^0 \in W_0^{1,1}(\Omega).$$

The purpose of this paper is to extend the applicability of this condition and to use the result so obtained to prove the validity of the Euler Lagrange equation for the minimizer, without assuming growth conditions from above for the integrand f. More precisely, our Theorem 4.1 below extends Stampacchia's theorem to a wider class of integrands f, while requiring less regularity on the solutions. Stampacchia's result is based on the a priori assumption that the solution is Lipschitzian, and it yields an estimate on the value of the Lipschitz constant. Our result requires that the solution be continuous, and it derives that it is, in fact, Lipschitzian. This step demands a different proof: Stampacchia's proof was based on the fact that the minimizer satisfies the Euler Lagrange equation; however, without the a priori assumption of Lipschitzianity, proving the validity of the Euler Lagrange equation under the conditions on f required by Stampacchia's theorem is still an open and challenging problem. As a consequence of our Theorem 4.1, we provide in Theorem 4.7 a result on the validity of the Euler Lagrange equation for the minimizer that does not require, as do those commonly used in the literature, growth assumptions *from above* on the integrand f.

For the proof of Theorem 4.1 we use the method of translations. This method has been used in contexts similar to the one here by, e.g., Brezis and Stampacchia [2], Brezis and Sibony [1], and, more recently, Treu and Vornicescu [10]. In all of the above papers the functional considered is

$$\int_\Omega [f(\nabla u(x)) + g(u(x))]\,dx,$$

and the argument used depends on g being strictly monotone; i.e., the case $g = 0$ is excluded. In the proof presented here, we develop an argument that allows us to

*Received by the editors April 9, 1999; accepted for publication (in revised form) June 8, 2001; published electronically December 7, 2001.
http://www.siam.org/journals/sicon/40-4/35466.html
†Dipartimento di Matematica e Applicazioni, Università di Milano Bicocca, Viale Sarca 202, 20126 Milano, Italy (cellina@matapp.unimib.it).

extend the method of translations to the case $g = 0$, the case of interest in this paper. Theorem 4.2 below, instrumental to the proof of our main result, is a weak maximum principle under rather general assumptions on f.

2. Notations and preliminary results. The closed ball of radius ρ about the origin is B_ρ. The subgradient of a convex function f is denoted by ∂f and its domain by $Dom(\partial f)$. The closure of its domain, $cl(Dom(\partial f))$, is a convex set [8]. A (possibly extended valued) convex function f is called *strictly convex* if it is strictly convex on its effective domain. A *face* of a convex set is a convex extremal subset. The collection of the relative interiors of the faces of a convex set is a partition of the convex set. We say that a set Ω has the *segment property* if, given $x^0 \in \partial\Omega$, there exist a neighborhood U^0 containing x^0 and a nonzero vector k such that $x + tk \in \Omega$ whenever $x \in \overline{\Omega} \cap U^0$ and $t \in (0, 1]$. Every convex set has the segment property. Let Ω be bounded and open. We say that $u \in W^{1,1}(\Omega)$ satisfies $u \leq 0$ on $\partial\Omega$ in the sense of $W^{1,1}(\Omega)$ if $u^+ \in W_0^{1,1}(\Omega)$.

3. The bounded slope condition $(BSC)_K$.

DEFINITION 3.1. *Let K be a positive real, Ω a bounded convex set. The boundary datum u^0 satisfies $(BSC)_K$ if for every $x^0 \in \partial\Omega$ there exist vectors $k^+(x^0)$ and $k^-(x^0)$, $\|k^+(x^0)\| \leq K, \|k^-(x^0)\| \leq K$, such that for every $x \in \partial\Omega$ we have*

$$u^0(x) - u^0(x^0) \leq \langle k^+(x^0), x - x^0 \rangle$$

and

$$u^0(x) - u^0(x^0) \geq \langle k^-(x^0), x - x^0 \rangle.$$

The validity of $(BSC)_K$ for some K depends on the smoothness of $\partial\Omega$ and of u^0, as can be seen from the classical results of Miranda [7] and of Hartman [6].

Following Stampacchia, let us call an integrand f *regular* if $f \in C^2(\Re^N)$ and the $N \times N$ matrix of partial derivatives is positive definite at every point. Stampacchia's theorem [9] is as follows.

THEOREM 3.2. *Let f be a regular integrand. Let $u(x)$ be a minimizing function for problem (P) among all Lipschitz functions which have the same boundary values $u^0(x)$ satisfying $(BSC)_K$. If, moreover, $u \in C^1(\Omega) \cap H^2(\Omega)$, then*

$$\max_{x \in \overline{\Omega}} |u_{x_i}| \leq K.$$

4. Main results. It is our purpose to prove the following theorem, our main result.

THEOREM 4.1. *Let Ω be open, bounded, and convex; let f be a (possibly extended valued) lower semicontinuous strictly convex function. Let $u^0 : \Omega \to \Re$ be Lipschitzian and let it satisfy $(BSC)_K$. Let w in $C(\Omega) \cap W^{1,1}(\Omega)$ be a solution to problem (P):*

$$\text{minimize} \int_\Omega f(\nabla u(x))\, dx \,:\, u - u^0 \in W_0^{1,1}(\Omega).$$

Then w is Lipschitzian and, for almost every x in Ω, $\|\nabla w(x)\| \leq K$.

Under the conditions of Stampacchia's theorem on f and Ω and assuming that the boundary datum satisfies $(BSC)_K$, the fact that the solution is *Lipschiztian* implies that the Lipschitz constant of the solution is K. Our Theorem 4.1 says that under the conditions of Theorem 4.1 on f and Ω and assuming that the boundary datum satisfies

$(BSC)_K$ for *some* constant K, knowing that the solution is *continuous* implies that the solution is *Lipschitzian*.

The following theorem is a generalized version of the weak maximum principle, to be used in the proof of Theorem 4.1.

THEOREM 4.2. *Let Ω in $\Re^N$ be open and bounded with the segment property; let f be a (possibly extended valued) lower semicontinuous, convex function. Let $u^0(x)$ in $W^{1,1}(\Omega)$ and $\ell(x) = \langle a, x\rangle + b$ be given such that in $\partial\Omega$, $u^0(x) \le \ell(x)$ in the sense of $W^{1,1}(\Omega)$. If the infimum in problem (P) is finite and attained by some function w, then the inequality*

$$w(x) \le \ell(x) \quad \text{for almost every } x \in \Omega$$

follows from either (i) *or* (ii) *below:*

(i) $a \notin Int(Dom(\partial f))$.

(ii) $a \in Int(Dom(\partial f))$ *and the face of $epi(f)$ whose relative interior contains $(a, f(a))$ has dimension less than N.*

(The latter condition is immediate when f is strictly convex.)

Remark. Let f be the indicator function of the unit disk $D \subset \Re^2$; i.e., $f(\xi) = 0$ when $\|\xi\| \le 1, f(\xi) = +\infty$ otherwise. Let $\overline{\Omega}$ be D and u^0 be $= 0$. Finally, let $\ell(x) = 0$. Then $a = 0 \in Int(Dom(f))$ and for $\|x\| = 1$, $\ell(x) \ge u^0(x)$. However, the function $w : \Omega \to \Re$ defined by $w(x) = 1 - \|x\|$, for which $\|\nabla w\| = 1$ a.e. in Ω, is a solution to the minimization problem (P) for the given f and u^0, but it is not true that $\ell(x) \ge w(x)$ a.e. in Ω. The face of $epi(f)$ containing $(0, 0)$ in its relative interior is of dimension $N = 2$.

For the proof of Theorem 4.2 we shall need the following lemma.

LEMMA 4.3. *Let f, Ω, $u^0(x)$, and ℓ be as in Theorem* 4.2. *Let $w - u^0$ be in $W_0^{1,1}(\Omega)$. Let $E^+ = \{x \in \Omega : w(x) > \ell(x)\}$. Then, for every $l \in \Re^N$,*

$$\int_{E^+} \langle l, \nabla w(x) - a\rangle \, dx = 0.$$

Proof. Since $u^0(x) \le \ell(x)$ on $\partial\Omega$ in the sense of $W^{1,1}(\Omega)$, we have

$$0 \le (w - \ell)^+ = [(u^0 - \ell) + (w - u^0)]^+ \le (u^0 - \ell)^+ + (w - u^0)^+,$$

so that $(w - \ell)^+ \in W_0^{1,1}(\Omega)$, i.e., $(w - \ell) \le 0$ on $\partial\Omega$ in the sense of $W^{1,1}(\Omega)$. Hence [11, Lemma 1.59] there exists a sequence (ψ_n), $\psi_n \in C^\infty(\overline{\Omega})$ and $\psi_n(x) \ge 0$ for x in $\partial\Omega$, converging to $(\ell - w)$ in $W^{1,1}(\Omega)$. Let $w_n = \ell - \psi_n$ and assume we have selected a subsequence of the sequence (w_n) converging to w pointwise as well as in $W^{1,1}(\Omega)$. Let $E^- = \{x \in \Omega : w(x) < \ell(x)\}$, $E_0 = \{x \in \Omega : w(x) = \ell(x)\}$, $E_n = \{x \in \Omega : w_n(x) - \ell(x) > 0\}$. Then $\chi_{E_n}(x) \to 1$ for almost every x in E^+, and $\chi_{E_n}(x) \to 0$ for almost every x in E^-.

We have

$$\int_{E^+} \langle l, \nabla w(x) - a\rangle \, dx = \int_\Omega \langle l, \nabla w(x) - a\rangle \chi_{E_n} \, dx + \int_\Omega \langle l, \nabla w(x) - a\rangle (\chi_{E^+} - \chi_{E_n}) \, dx.$$

The last integral is the sum of the same integral over E^+, over E^-, and over E^0. The first two integrals tend to zero from an application of the dominated convergence theorem; the third is zero since, on E^0, $\nabla w(x) = a$ a.e. Hence

$$\int_\Omega \langle l, \nabla w(x) - a\rangle (\chi_{E^+} - \chi_{E_n}) \, dx \to 0.$$

Moreover,

$$\int_\Omega \langle l, \nabla w(x) - a\rangle \chi_{E_n}\, dx = \int_{E_n} \langle l, \nabla w_n(x) - a\rangle\, dx + \int_{E_n} \langle l, \nabla w(x) - \nabla w_n(x)\rangle\, dx.$$

The second integral tends to zero since $w_n \to w$ in $W^{1,1}(\Omega)$. To prove the lemma it suffices to show that $\int_{E_n} \langle l, \nabla w_n(x) - a\rangle\, dx = 0$.

Let $\mathbf{l} = l/\|l\|$. Let P^l be the plane through the origin orthogonal to $\mathbf{l}$, O^l the projection of Ω on P^l, and $L^l(x'), x' \in O^l$, the line $\{x' + \mathbf{l}\tau; \tau \in \Re\}$. The intersection of a line $L^l(x')$ with the open set E_n can be described as $\{x' + \mathbf{l}\tau : \tau \in \cup_i(\alpha_i(x'), \beta_i(x'))\}$, where some or all of the points of $x' + \mathbf{l}\alpha_i(x')$ and of $x' + \mathbf{l}\beta_i(x')$ can belong to $\partial\Omega$. Then

$$int_{E_n} \langle l, \nabla w_n(x) - a\rangle\, dx = \int_{O^l} \left(\int_{E_n \cap L^l(x')} \langle l, \nabla w_n(x' + \mathbf{l}\tau) - a\rangle\, d\tau \right) dx'$$

$$= \int_{O^l} \left(\Sigma_i \int_{\alpha_i(x')}^{\beta_i(x')} \langle l, \nabla w(x' + \mathbf{l}\tau) - a\rangle\, d\tau \right)\, dx'.$$

We have

$$\int_{\alpha_i(x')}^{\beta_i(x')} \langle l, \nabla w_n(x' + \mathbf{l}\tau) - a\rangle\, d\tau = \int_{\alpha_i(x')}^{\beta_i(x')} \|l\| \frac{d}{d\tau}[w_n(x' + \mathbf{l}\tau) - \ell(x' + \mathbf{l}\tau)] d\tau$$

$$= \|l\|\{[w_n(x' + \mathbf{l}\beta_i(x')) - \ell(x' + \mathbf{l}\beta_i(x'))] - [w_n(x' + \mathbf{l}\alpha_i(x')) - \ell(x' + \mathbf{l}\alpha_i(x'))]\}.$$

For each i, when $x' + \mathbf{l}\alpha_i(x')$ is in Ω, w_n and ℓ coincide, and the same is true for $x' + \mathbf{l}\beta_i(x')$. Since at $\partial\Omega$, $w_n(x) \le \ell(x)$ for all $x' + \mathbf{l}\alpha_i(x')$ while $x' + \mathbf{l}\alpha_i(x')$ is the limit of points where $w_n(x) > \ell(x)$, and the same is true for $x' + \mathbf{l}\beta_i(x')$, we have that the last integral is zero. This ends the proof that $\int_{E^+} \langle l, \nabla w(x) - a\rangle\, dx = 0$. □

Proof of Theorem 4.2. We wish to prove that E^+ has measure zero. We assume that it is not so and will show that this leads to a contradiction in either case (i) or (ii).

We must have that $\nabla w(x)$ is a.e. in $Dom(f)$, hence in $cl(Dom(\partial f))$; otherwise the integral would not be finite.

(a) Assume (i), i.e., $a \notin Int(Dom(\partial f))$. Then a can be separated by a hyperplane from the convex and closed set $cl(Dom(\partial f))$, i.e., there exists $h \neq 0$ such that $\langle h, a\rangle \ge \sup_{d \in Dom(\partial f)} \langle h, d\rangle$. Hence for almost every $x \in \Omega$, in particular for almost every $x \in E^+$, we have the following inequality:

$$\langle h, \nabla w(x) - a\rangle \le 0.$$

The proof of case (i) continues in step (e) below.

(b) Assume (ii). Fix k in $\partial f(a)$. Let $\eta^+ = (w - \ell)^+$; since

$$0 \le (w - \ell)^+ = (w - u^0 + u^0\ell)^+ \le (w - u^0)^+ (u^0 - \ell)^+,$$

and both maps at the right-hand side are in $W_0^{1,1}(\Omega)$, so is $(w-\ell)^+$. We thus have

$$(w-\eta^+)(x) = \begin{cases} w(x) & \text{if } w(x) \le \ell(x), \\ \ell(x) & \text{otherwise,} \end{cases}$$

$$\nabla(w-\eta^+)(x) = \begin{cases} \nabla w(x) & \text{if } w(x) \le \ell(x), \\ a & \text{otherwise.} \end{cases}$$

Hence, by the convexity of f and applying Lemma 4.3, we obtain

$$\int_\Omega (f(\nabla w(x)) - f(\nabla(w-\eta^+)(x)))\,dx = \int_{E^+} (f(\nabla w(x)) - f(a))\,dx$$

$$\ge \int_{E^+} \langle k, \nabla w(x) - a\rangle\,dx = 0.$$

(c) Since w is a minimizer, we also have

$$0 \ge \int_\Omega (f(\nabla w(x)) - f(\nabla(w-\eta^+)(x)) = \int_{E^+} (f(\nabla w(x)) - f(a))\,dx \ge 0;$$

hence, from the conclusion of (b),

$$\int_{E^+} \{f(\nabla w(x)) - [f(a) + \langle k, \nabla w(x) - a\rangle]\}dx = 0.$$

The integrand above is nonnegative, so we obtain that, a.e. in E^+, $f(\nabla w(x)) = f(a) + \langle k, \nabla w(x) - a\rangle$, i.e., the $N+1$-dimensional vector $(f(\nabla w(x)), \nabla w(x))$ belongs, for almost every x, to H, the intersection of the epigraph of f with the hyperplane $z = f(a) + \langle k, \xi - a\rangle$. H is a face of $epi(f)$: it is either of dimension less than N or its dimension is N. Let $H^N = \{\xi : f(\xi) = f(a) + \langle k, \xi - a\rangle\}$ be its projection on $\Re^N$.

(d) By assumption, the face of $epi(f)$, containing $(a, f(a))$ in its relative interior, has dimension less than N, and so does F_a^N, its projection on $\Re^N$. F_a^N is a face of H^N: there is a (nonzero) N-vector τ that properly separates F_a^N from H^N, $\langle \tau, \xi - a\rangle = 0$, $\xi \in F_a^N$, $\langle \tau, \xi - a\rangle \le 0$, $\xi \in H^N$, and there is some $z \in (H^N \setminus F_a^N)$ such that $\langle \tau, z - a\rangle < 0$. Hence, for $\xi \in H^N$,

$$f(\xi) \ge f(a) + \langle k, \xi - a\rangle \ge f(a) + \langle k + \tau, \xi - a\rangle.$$

Since, for $x \in E^+$, we have $\nabla w(x) \in H^N$, in particular we have, for $x \in E^+$,

$$f(\nabla w(x)) \ge f(a) + \langle k + \tau, \nabla w(x) - a\rangle.$$

Again, since w is a minimizer, we have

$$0 \ge \int_\Omega (f(\nabla w(x)) - f((\nabla(w-\eta^+)(x))) = \int_{E^+} (f(\nabla w(x)) - f(a))\,dx$$

$$\ge \int_{E^+} \langle k + \tau, \nabla w(x) - a\rangle\,dx = 0,$$

where the last equality follows from Lemma 4.3. Hence

$$\int_{E^+} \{f(\nabla w(x)) - [f(a) + \langle k + \tau, \nabla w(x) - a\rangle]\}dx = 0,$$

and, since $f(\nabla w(x)) \geq f(a)+\langle k+\tau, \nabla w(x)-a\rangle$ a.e. in E^+, it follows that $f(\nabla w(x)) = f(a) + \langle k + \tau, \nabla w(x) - a\rangle$. From the conclusion of (c) we obtain that, a.e. in E^+,

$$\langle \tau, \nabla w(x) - a\rangle = 0.$$

(e) By the conclusion of (a) in case (i) and by the above construction in case (ii), there is a nonzero N-dimensional vector, $k^\perp$, such that $\langle k^\perp, \nabla w(x) - a\rangle \leq 0$ a.e. in E^+. Choose a line $\{x' + k^\perp t\}$ intersecting E^+ on a set of positive measure and such that $t \to w(x' + ht)$ is absolutely continuous. Let $T^+ = \{t : x' + ht \in E^+\}$ and let t^+ be in T^+, i.e., such that for $x^+ = x' + k^\perp t^+$, $\eta^+(x^+)$ is positive. Since the gradient of η^+ is

$$\nabla \eta^+(x) = \begin{cases} 0 & \text{on } \Omega \setminus E^+, \\ \nabla w(x) - a & \text{otherwise,} \end{cases}$$

we have

$$0 < (w - \ell)^+(x' + t^+h) = \int_{-\infty}^{t^+} \left(\frac{d}{dt}(w - \ell)^+(x' + th)\right) dt$$

$$= \int_{(-\infty, t^+]\cap T^+} \langle h, \nabla w(x' + th) - a\rangle \, dt$$

$$\leq \int_{(-\infty, t^+]\cap T^+} \left(\sup_{d \in Dom(\partial f)} \{\langle h, d - a\rangle\}\right) dt \leq 0,$$

a contradiction. So E^+ has measure zero. □

For the proof of Theorem 4.1 we shall need the following preliminary results.

LEMMA 4.4. *Let $\Omega^i, i = 1, 2$, be open and let g^i be in $W_0^{1,1}(\Omega^i)$ and such that for almost every x in Ω^i, $g^i(x) \geq 0$. Then* $\min(g^1(x), g^2(x)) \in W_0^{1,1}(\Omega^1 \cap \Omega^2)$.

Proof. Let $g_n^i : \Omega^i \to \Re, i = 1, 2$, be two sequences of Lipschitzian maps with compact support in Ω^i, g_n^i converging to g^i in $W_0^{1,1}(\Omega^i)$ and pointwise a.e. Set $G(x) = \min(g^1(x), g^2(x))$, $E^1 = \{x \in (\Omega^1 \cap \Omega^2) : g^1(x) < g^2(x)\}$, $E^2 = \{x \in (\Omega^1 \cap \Omega^2) : g^2(x) < g^1(x)\}$, $E^0 = \{x \in (\Omega^1 \cap \Omega^2) : g^1(x) = g^2(x)\}$; set also $G_n(x) = \min((g_n^1(x), g_n^2(x))$: the maps G_n are Lipschitzian with compact support contained in $(\Omega^1 \cap \Omega^2)$. One has

$$\int_{\Omega^1\cap\Omega^2} |G - G_n| = \int_{E^1} |G - G_n| + \int_{E^2} |G - G_n| + \int_{E^0} |G - G_n|.$$

To evaluate the first integral, set $E_n^{1,1} = E^1 \cap \{x : g_n^1 < g_n^2\}, E_n^{1,2} = E^1 \cap \{x : g_n^2 < g_n^1\}, E_n^{1,0} = E^1 \cap \{x : g_n^1 = g_n^2\}$. Then

$$\int_{E^1} |G - G_n| = \int_{E_n^{1,1}} |g^1 - g_n^1| + \int_{E_n^{1,0}} |g^1 - g_n^1| + \int_{E_n^{1,2}} |g^1 - g_n^2|,$$

and the first two integrals converge to zero since $g_n^1 \to g^1$ in $W^{1,1}(\Omega^1)$. Also, g_n^2 converges pointwise to g^2; hence $\chi_{E_n^{1,2}} \to 0$ pointwise a.e. The sequence $(|g^1 - g_n^2|)$ is equiintegrable, since g_n^2 converges in $L^1(\Omega^2)$; by Egoroff's theorem $\int_{E_n^{1,2}} |g^1 - g_n^2| \to 0$. Similarly for the other cases and for $\int_{\Omega^1\cap\Omega^2} \|\nabla G - \nabla G_n\|$. □

LEMMA 4.5. *Let f be strictly convex on its effective domain. For every pair (a, b) in its effective domain, $a \neq b$, for every $\lambda, 0 < \lambda < 1$, we have*

$$f(a + \lambda(b-a)) - f(a) + f(b - \lambda(b-a)) - f(b) < 0.$$

Proof. Consider the restriction of f to the line oriented from a to b. Under the conditions of the Lemma, the map $c \to \frac{f(c+\lambda(b-a))-f(c)}{\lambda(b-a)}$ is strictly monotonic. □

Proof of Theorem 4.1. It is convenient to set $\Psi^+(x) = \inf_{x^0 \in \partial\Omega} \langle k^+(x^0), x - x^0\rangle + u^0(x^0)$ and $\Psi^-(x) = \sup_{x^0 \in \partial\Omega} \langle k^-(x^0), x - x^0\rangle + u^0(x^0)$; the maps Ψ^+ and Ψ^- are Lipschitzian with Lipschitz constant K. Applying Theorem 4.2 to each of the affine maps $\langle k^+(x^0), x - x^0\rangle + u^0(x^0)$, we infer that the solution w satisfies $w \leq \Psi^+$. Applying the same theorem to the problem $\tilde{P}$ whose data are $\tilde{f}(\xi) = f(-\xi)$ and $\tilde{u}^0 = -u^0$, we obtain $\Psi^- \leq w$.

To prove the theorem it is enough to show that there cannot exist a unit vector $\mathbf{v}$, a scalar $M > K$, and a set $E \subset \Omega$ with $\mu(E) > 0$ such that, for x in E, $\langle \nabla w(x), \mathbf{v}\rangle > M$. Let us assume that $M, \mathbf{v}, E$ exist and derive a contradiction.

(a) There exists a representative of w that is absolutely continuous on almost every line parallel to $\mathbf{v}$. Since it coincides with w a.e. in Ω, on almost every such line $\{x = t\mathbf{v} + a : t \in \Re\}$, it coincides with w for almost every t; by continuity, they coincide for all t on every such line. Hence w is absolutely continuous on almost every line parallel to $\mathbf{v}$. On a plane orthogonal to $\mathbf{v}$ there exists a set of points of positive $(N-1)$ measure, such that lines parallel to $\mathbf{v}$ through these points meet E in a set of positive one-dimensional measure. Let us fix one such line; let x^* be a point on it that is at once in E and such that the map $t \to w(x^* + t\mathbf{v})$ is differentiable at $t = 0$ with derivative

$$\frac{d}{dt}[w(x^* + t\mathbf{v})]|_{t=0} = \langle \nabla w(x^*), \mathbf{v}\rangle = M + \zeta, \qquad \zeta > 0.$$

Then there exists $h^* > 0$ such that for every $0 < h \leq h^*$

$$w(x^* + h\mathbf{v}) - w(x^*) - Mh > 0.$$

(b) We wish to prove the following claim. Let x^{**} be a point in Ω such that $t \to w(x^{**} + t\mathbf{v})$ is differentiable at $t = 0$ with derivative $D^{**} > M$, and let $h^{**} > 0$ be such that for every $0 < h \leq h^{**}$, $x^{**} + h\mathbf{v}$ is in Ω and

$$w(x^{**} + h\mathbf{v}) - w(x^{**}) - Mh > 0.$$

Then $t \to w(x^{**} + t\mathbf{v})$ is affine on $[0, h^{**}]$ with derivative D^{**}.

Proof of the claim. Fix any $h \in (0, h^{**}]$. On the convex set $\Omega_h = \Omega \cap (\Omega - h\mathbf{v})$ both $x \to w(x)$ and $x \to w(x + h\mathbf{v})$ are defined. By assumption, the set

$$E_h^+ = \{x \in \Omega_h; w(x + h\mathbf{v}) > w(x) + hM\}$$

is an open subset of Ω_h containing x^{**}. For $x \in E_h^+$, we have that $y = x + h\mathbf{v}$ is such that $y - h\mathbf{v}$ is in Ω and $w(y - h\mathbf{v}) < w(y) - hM$. The set

$$E_h^- = \{y \in \Omega_{-h}; w(y - h\mathbf{v}) < w(y) - hM\}$$

is a translate of $E_h^+ : E_h^- - h\mathbf{v} = E_h^+$. Let $\eta_h^+(x)$ on Ω_h be $(w(x+h\mathbf{v}) - w(x) - hM)^+$ and $\eta_h^-(x)$ on Ω_{-h} be $(w(x - h\mathbf{v}) - w(x) + hM)^-$. We wish to show that η_h^+ and η_h^-

are admissible variations, i.e., that they are in $W_0^{1,1}(\Omega)$. From the Lipschitzianity of Ψ^+ and of Ψ^- we obtain

$$\eta^+(x) \le \Psi^+(x+h\mathbf{v}) - Mh - w(x) \le \Psi^+(x) - w(x),$$

$$\eta^+(x) \le w(x+h\mathbf{v}) - Mh - \Psi^-(x+h\mathbf{v}) + Kh \le w(x+h\mathbf{v}) - \Psi^-(x+h\mathbf{v});$$

i.e., $\eta_h^+ \le \min(\Psi^+(x) - w(x), w(x+h\mathbf{v}) - \Psi^-(x+h\mathbf{v}))$.

Apply Lemma 4.4 with $\Omega^1 = \Omega, \Omega^2 = \Omega_h, g^1(x) = \Psi^+(x) - w(x), g^2(x) = w(x+h\mathbf{v}) - \Psi^-(x+h\mathbf{v})$ to infer that η_h^+ is an admissible variation, and the same is true for η_h^-. Since w is a minimum, we must have that for all λ

$$\int_\Omega f(\nabla w(x) + \lambda \nabla \eta_h^+(x))\, dx \ge \int_\Omega f(\nabla w(x))\, dx,$$

$$\int_\Omega f(\nabla w(x) + \lambda \nabla \eta_h^-(x))\, dx \ge \int_\Omega f(\nabla w(x))\, dx.$$

We have

$$\nabla \eta_h^+ = \begin{cases} \nabla w(x+h\mathbf{v}) - \nabla w(x) & \text{if } x \in E_h^+, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\nabla \eta_h^- = \begin{cases} \nabla w(x-h\mathbf{v}) - \nabla w(x) & \text{if } x \in E_h^-, \\ 0 & \text{otherwise,} \end{cases}$$

so that the above inequalities yield

$$\int_{E_h^+} f(\nabla w(x) + \lambda[\nabla w(x+h\mathbf{v}) - \nabla w(x)]) - f(\nabla w(x))\, dx \ge 0,$$

$$\int_{E_h^-} f(\nabla w(x) + \lambda[\nabla w(x-h\mathbf{v}) - \nabla w(x)]) - f(\nabla w(x))\, dx \ge 0.$$

Making the change of variables $y = x + h\mathbf{v}$ and adding the two inequalities, one obtains

$$\int_{E_h^+} \{f(\nabla w(x) + \lambda[\nabla w(x+h\mathbf{v}) - \nabla w(x)]) - f(\nabla w(x)) + f(\nabla w(x+h\mathbf{v})$$

$$-\lambda[\nabla w(x+h\mathbf{v}) - \nabla w(x)]) - f(\nabla w(x+h\mathbf{v}))\}\, dx \ge 0.$$

From Lemma 4.5 we obtain that, for every x such that $\nabla w(x) \ne \nabla w(x+h\mathbf{v})$, the integrand is negative. Since E_h^+ is a nonempty open set, this is a contradiction unless, a.e. in E_h^+, $\nabla w(x) = \nabla w(x+h\mathbf{v})$.

The set E_h^+ contains a ball B_h about x^{**}; for x in this ball, $\nabla w(x) - \nabla w(x+h\mathbf{v}) = 0$ a.e. By the continuity of w, there exists a constant C such that, on $B_h, w(x) - w(x+h\mathbf{v}) = C$. Then, since the limit

$$\lim_{t\to 0} \frac{1}{t}[w(x^{**} + t\mathbf{v}) - w(x^{**})]$$

exists and equals D^{**}, so does

$$\lim_{t\to 0}\frac{1}{t}[w(x^{**}+t\mathbf{v}+h\mathbf{v})-w(x^{**}+h\mathbf{v})].$$

In particular, the derivative at $t=0$ of the map $t\to w(x^*+h\mathbf{v}+t\mathbf{v})$ exists and equals D^{**}. This reasoning holds for every $0<h\le h^{**}$, thus proving the claim.

(c) The previous claim applies at x^*. Hence the map $t\to w(x^*+t\mathbf{v})$ is affine on $[0,h^*]$ with derivative $M+\zeta$. Let $[0,\Lambda]$ be the maximal interval on which this map is affine. We claim that $x^*+\Lambda\mathbf{v}$ is in $\partial\Omega$. If it is in Ω for some $\varepsilon>0$, then so is $x^*+(\Lambda+\tau)\mathbf{v}$ for $0\le\tau<\varepsilon$. Choose λ in $(0,\Lambda)$. The map $t\to w(x^*+t\mathbf{v})$ is differentiable at λ with derivative $\langle\nabla w(x),\mathbf{v}\rangle$. Moreover we have that $w(x^*+\Lambda\mathbf{v})-w(x^*+\lambda\mathbf{v})=(M+\zeta)(\Lambda-\lambda)$, i.e.,

$$w((x^*+\lambda\mathbf{v})+(\Lambda-\lambda)\mathbf{v})-w(x^*+\lambda\mathbf{v})-(\Lambda-\lambda)M=\zeta M.$$

Hence, by the continuity of w, for all $\tau\le\varepsilon_1<\varepsilon$,

$$w((x^*+\lambda\mathbf{v})+(\Lambda-\lambda+\tau)\mathbf{v})-w(x^*+\lambda\mathbf{v})-(\Lambda-\lambda+\tau)M>0.$$

The point $x^*+\lambda\mathbf{v}$ can be used as x^{**} with $h^{**}=(\Lambda-\lambda+\varepsilon_1)$. Applying the claim of part (b), we have that the map $t\to w(x^*+t\mathbf{v})$ is affine on $[0,\Lambda+\varepsilon_1]$, contradicting the maximality of Λ. Hence $x^*+\Lambda\mathbf{v}$ is in $\partial\Omega$.

(d) Let x^{***} be $x^*+\Lambda\mathbf{v}$; since u^0 is continuous, the conditions $u^0\le\ell$ and $u^0\ge\ell$ on $\partial\Omega$ in $W^{1,1}$ in sense and pointwise coincide. Thus, by Theorem 4.2, for every $x\in\Omega$ (in particular for x^*)

$$u^0(x^{***})+\langle k^-(x^{***}),x-x^{***}\rangle\le w(x)\le u^0(x^{***})+\langle k^+(x^{***}),x-x^{***}\rangle.$$

Hence, from point (c),

$$w(x^{***})=w(x^*)+\|x^*-x^{***}\|(M+\zeta)>w(x^*)-\langle k^-(x^{***}),x^*-x^{***}\rangle\ge u^0(x^{***}),$$

while, for every $x\in\Omega$,

$$w(x)\le u^0(x^{***})+\langle k^+(x^{***}),x-x^{***}\rangle.$$

The above inequalities are incompatible whenever $\|x-x^{***}\|$ is sufficiently small. This is a contradiction. □

COROLLARY 4.6. *Under the same assumptions on Ω, f, and u^0 as in Theorem 4.2, let solutions to problem (P) be continuous. Then problem (P) and problem $(P)_K$,*

$$\textit{minimize}\int_\Omega f(\nabla u(x))\,dx\quad\textit{on}\quad u-u^0\in W_0^{1,1}(\Omega)\ \textit{and}\ \|\nabla u(x)\|\le K,$$

are equivalent, in the sense that they have the same solutions.

Known results on the validity of the Euler Lagrange equation for a minimizer w hold under growth assumptions *from above* on f, i.e., under slow growth assumptions. (An exception to this statement is [4], whose results are for integrands f that tend to $+\infty$ at the boundary of $Dom(f)$, under conditions different from those presented here.)

THEOREM 4.7. *Let $f:\Re^N\to\Re$ be C^1, strictly convex, and such that for some α and $\beta>0$, $f(\xi)\ge\alpha+\beta\|\xi\|^p$, $p>N$. Let Ω be bounded and convex and let u^0*

satisfy $(BSC)_K$ *for some constant* K. *Let* w *be a solution to problem* (P). *Then* w *is Lipschitzian and it satisfies the Euler Lagrange equation in the sense that*

$$\int_\Omega \langle \nabla f(\nabla w(x)), \nabla \eta(x) \rangle \, dx = 0$$

for every Lipschitzian η, $\eta|_{\partial\emptyset} = 0$.

Proof. From the growth assumptions we know that $w \in W^{1,p}(\Omega)$; hence we know that it is continuous. Theorem 4.1 applies and $\|\nabla w(x)\| \leq K$ a.e. in Ω. Fix η and let λ be so small that $\lambda \|\nabla \eta\| \leq 1$. Let $M = \max_{\xi \in B_{K+1}} \{\|\nabla f(\xi)\|\}$. Since w is a minimum, one has

$$0 \leq \left(\frac{1}{\lambda}\right) \int (f(\nabla w + \lambda \nabla \eta) - f(\nabla w)) = \int \langle \nabla f(\nabla (w(x) + \sigma(x) \lambda \nabla \eta(x)), \nabla \eta(x) \rangle \, dx,$$

and the term under the integral sign, that converges pointwise to $\langle \nabla f(\nabla w(x)), \nabla \eta(x) \rangle$ as $\lambda \to 0$, is bounded in norm by M. Hence, applying the dominated convergence theorem, the result follows. □

Acknowledgment. The author is indebted to an anonymous referee for the kind, competent, and acute remarks and suggestions.

REFERENCES

[1] H. Brezis and M. Sibony, *Equivalence de deux Inequations variationnelles et applications*, Arch. Ration. Mech. Anal., 41 (1971), pp. 254–265.
[2] H. Brezis and G. Stampacchia, *Sur la regularité de la solution d'inequations elliptiques*, Bull. Soc. Math. France, 96 (1968), pp. 153–180.
[3] A. Cellina, *On minima of a functional of the gradient: Necessary conditions*, Nonlinear Anal., 20 (1993), pp. 337–341.
[4] H.J. Choe, *On the minimizers of certain singular convex functionals*, J. Korean Math. Soc., 30 (1993) pp. 315–335.
[5] P. Hartman and L. Nirenberg, *On spherical image maps whose Jacobians do not change sign*, Amer. J. Math., 81 (1959), pp. 901–920.
[6] P. Hartman, *On the bounded slope condition*, Pacific J. Math., 18 (1966), pp. 495–511.
[7] M. Miranda, *Un teorema di esistenza e unicità per il problema dell'area minima in n variabili*, Ann. Scuola Norm. Sup. Pisa Cl. Sci., 9 (1965), pp. 233–249.
[8] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1972.
[9] G. Stampacchia, *On some regular multiple integral problems in the calculus of variations*, Comm. Pure Appl. Math., 16 (1963), pp. 383–421.
[10] G. Treu and M. Vornicescu, *On the equivalence of two variational problems*, Calc. Var. Partial Differential Equations, 11 (2000), pp. 307–319.
[11] G.M. Troianiello, *Elliptic Differential Equations and Obstacle Problems*, Plenum, New York, 1987.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 4, pp. 1280–1295

OPTIMALITY CONDITIONS FOR IRREGULAR INEQUALITY-CONSTRAINED PROBLEMS*

A. F. IZMAILOV† AND M. V. SOLODOV‡

Abstract. We consider feasible sets given by conic constraints, where the cone defining the constraints is convex with nonempty interior. We study the case where the feasible set is not assumed to be regular in the classical sense of Robinson and obtain a constructive description of the tangent cone under a certain new second-order regularity condition. This condition contains classical regularity as a special case, while being weaker when constraints are twice differentiable. Assuming that the cone defining the constraints is finitely generated, we also derive a special form of primal-dual optimality conditions for the corresponding constrained optimization problem. Our results subsume optimality conditions for both the classical regular and second-order regular cases, while still being meaningful in the more general setting in the sense that the multiplier associated with the objective function is nonzero.

Key words. tangent cone, regularity, constraint qualification, optimality conditions

AMS subject classifications. 90C30, 46T20, 47J07, 90C33

PII. S0363012999357549

1. Introduction.

Let X and Y be normed linear spaces. We consider the sets given by

$$D = \{x \in X \mid F(x) \in K\}, \tag{1.1}$$

where the constraint mapping $F : X \to Y$ is smooth enough and K is a closed convex cone in Y with nonempty interior. The problem of an accurate and constructive description of the tangent cone to a set at a given point is fundamental for many reasons, one of which is deriving optimality conditions. Recall that a vector $h \in X$ is called *tangent* to a set $D \subset X$ at a point $\bar{x} \in D$ if there exists a mapping $r : \Re_+ \to X$ such that

$$\bar{x} + th + r(t) \in D \quad \forall\, t \in \Re_+, \quad \|r(t)\| = o(t). \tag{1.2}$$

The set of all such vectors h in X is the *tangent cone* to the set D at the point $\bar{x}$, which we shall denote by $T_D(\bar{x})$. As is well known,

$$T_D(\bar{x}) \subset \{h \in X \mid F'(\bar{x})h \in T_K(F(\bar{x}))\}, \tag{1.3}$$

which is the first-order necessary condition for tangency. To obtain a *precise* description of $T_D(\bar{x})$, i.e., a sufficient condition for tangency, some regularity (also called constraint qualification) condition is needed. One classical condition in this setting is Robinson's condition [27]:

$$0 \in \operatorname{int}(F(\bar{x}) + \operatorname{Im} F'(\bar{x}) - K). \tag{1.4}$$

*Received by the editors June 7, 1999; accepted for publication (in revised form) June 8, 2001; published electronically December 7, 2001.

http://www.siam.org/journals/sicon/40-4/35754.html

†Computing Center of the Russian Academy of Sciences, Vavilova Str. 40, Moscow, GSP-1, Russia (izmaf@ccas.ru). The research of this author was supported by Russian Foundation for Basic Research grants 99-01-00472 and 01-01-00810.

‡Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Jardim Botânico, Rio de Janeiro, RJ 22460-320, Brazil (solodov@impa.br). The research of this author was supported in part by CNPq grant 300734/95-6, by PRONEX–Optimization, and by FAPERJ.

Note that in (1.4) cone K is not required to have a nonempty interior. If (1.4) is satisfied, then (1.3) holds as an equality, e.g., [12, Corollary 2.91]. Deriving an accurate constructive description of the tangent cone without assuming (1.4) and, more generally, when (1.3) does not necessarily hold as an equality, is one of the principal goals of this paper. Our approach is based on a certain new notion of second-order regularity, which in the setting of K with nonempty interior is weaker than (1.4); see Definition 2.1 and Remark 2.1. An immediate application of this description is the primal form of necessary optimality conditions for the problem

$$\min\{f(x) \mid x \in D\}, \tag{1.5}$$

where the objective function $f : X \to \Re$ is smooth enough.

Our second goal is to obtain primal-dual optimality conditions for the irregular case, with a nonzero multiplier associated to the objective function. If $\bar{x}$ is a local solution of (1.5), (1.1), then the classical F. John–type first-order necessary optimality conditions (e.g., see [11]) state that there exists a generalized Lagrange multiplier $(y_0, y^*) \in (\Re \times Y^*) \setminus \{0\}$ such that

$$\begin{aligned} &y_0 f'(\bar{x}) - (F'(\bar{x}))^* y^* = 0, \\ &F(\bar{x}) \in K,\ y^* \in K^*,\ \langle y^*, F(\bar{x})\rangle = 0,\ y_0 \geq 0, \end{aligned} \tag{1.6}$$

where Y^* is the dual space of Y, $(F'(\bar{x}))^*$ is the adjoint operator of $F'(\bar{x})$, and K^* is the dual cone of K. If $y_0 = 0$, the F. John conditions hold trivially independently of the objective function and therefore their utility for describing optimality in that case is very limited (at least without some further developments). Assumptions that guarantee the existence of a multiplier (y_0, y^*) with $y_0 \neq 0$ are again constraint qualification conditions, such as (1.4). For problems with a finitely generated cone K, without assuming (1.4) or equality in (1.3), we obtain a special form of primal-dual optimality conditions under our assumption of second-order regularity. Our optimality conditions resemble the structure of (1.6), where $y_0 \neq 0$ and a certain term involving the second derivative of F is added to the standard Lagrangian; see Theorem 3.2. Our optimality conditions subsume those for the classical regular case of (1.4), as well as those for the more general second-order regular case of [7, 8]; see section 4.

In section 4, we compare our results with other approaches relevant for irregular inequality-constrained problems. We also provide an example showing that our results can be used to verify optimality in cases where other known approaches appear not to be applicable. We note that those cases do not seem pathological or exotic; see Example 4.1.

Finally, we note that in the case of the nonlinear programming problem, i.e., when $Y = \Re^m \times \Re^s$ and $K = \Re^m_- \times \{0\}$, Robinson's regularity condition (1.4) reduces to the classical Mangasarian–Fromovitz constraint qualification [23], and with $y_0 \neq 0$ optimality conditions (1.6) become the classical Karush–Kuhn–Tucker conditions.

Our notation is fairly standard. If Σ is a topological linear space, then Σ^* denotes its (topologically) dual space and $\langle\cdot, \cdot\rangle$ is the pairing of elements in Σ^* and Σ, i.e., $\langle\sigma^*, \sigma\rangle$ is the value of the linear functional $\sigma^* \in \Sigma^*$ on $\sigma \in \Sigma$. For a cone C in Σ, the positive dual cone (sometimes also referred to as the polar cone) of C is $C^* := \{\sigma^* \in \Sigma^* \mid \langle\sigma^*, \sigma\rangle \geq 0 \quad \forall\, \sigma \in C\}$. For an arbitrary set Ω in Σ, the set orthogonal to Ω is $\Omega^\perp := \{\sigma^* \in \Sigma^* \mid \langle\sigma^*, \sigma\rangle = 0 \quad \forall\, \sigma \in \Omega\}$. If Υ and Σ are topological linear spaces and $\Lambda : \Upsilon \to \Sigma$ is a continuous linear operator, then $\Lambda^* : \Sigma^* \to \Upsilon^*$ denotes the adjoint operator of Λ. The interior and the closure of a set Ω (in appropriate topology) are denoted by $\operatorname{int}\Omega$ and $\operatorname{cl}\Omega$, respectively, and linear and

conic hulls of this set (in appropriate linear space) by $\operatorname{lin}\Omega$ and $\operatorname{cone}\Omega$, respectively. A cone C in a linear space Σ is referred to as finitely generated if either it is empty or there exists a positive integer s and some elements $\sigma^i \in \Sigma$, $i = 1, \ldots, s$, such that $\operatorname{cl} C = \operatorname{cone}\{\sigma^1, \ldots, \sigma^s\} \cup \{0\}$. When we write that a mapping F is twice Fréchet-differentiable at a point $\bar{x}$, we mean that it is Fréchet-differentiable on a neighborhood of $\bar{x}$, and its derivative is Fréchet-differentiable at $\bar{x}$ (and similarly for higher-order Fréchet-differentiability).

Some auxiliary facts from convex analysis that are used throughout the paper are collected in the appendix.

2. Tangent cone description. As is well known [24], [12, Lemma 2.99], in our setting where $\operatorname{int} K \neq \emptyset$, Robinson's regularity condition (1.4) is equivalent to

$$\exists\, \bar{\xi} \in X \text{ such that } F(\bar{x}) + F'(\bar{x})\bar{\xi} \in \operatorname{int} K. \tag{2.1}$$

This condition implies that for $h \in T_D(\bar{x})$ the inclusion

$$F'(\bar{x})h \in T_K(F(\bar{x})) = \operatorname{cl}(K + \operatorname{lin}\{F(\bar{x})\}) \tag{2.2}$$

is both necessary and sufficient, e.g., [12, Corollary 2.91]. In the irregular case, $T_D(\bar{x})$ can be smaller than the set of $h \in X$ satisfying (2.2), and a more refined description is needed. To this end, it is natural to take into account the second-order information about F at $\bar{x}$. We proceed with a second-order characterization of the tangent cone, starting with the following definition.

DEFINITION 2.1. *We say that conic constraints in* (1.1) *are* second-order regular *at a feasible point* $\bar{x}$ with respect to a direction $h \in X$ *if*

$$\exists\, (\bar{\xi}, \bar{h}) \in X \times X \text{ such that } F(\bar{x}) + F'(\bar{x})\bar{h} \in K,$$
$$F(\bar{x}) + F'(\bar{x})\bar{\xi} + F''(\bar{x})[h, \bar{h}] \in \operatorname{int} K.$$

Remark 2.1. If Robinson's condition (2.1) is satisfied, then second-order regularity holds with respect to *every* $h \in X$, including $h = 0$. (To verify this, just choose $\bar{\xi}$ satisfying (2.1) and $\bar{h} = 0$.)

Observe further that Definition 2.1 is equivalent to saying that

$$\exists\, \bar{h} \in X \text{ such that } F'(\bar{x})\bar{h} \in T_K^r(F(\bar{x})) = K + \operatorname{lin}\{F(\bar{x})\}, \tag{2.3}$$
$$F''(\bar{x})[h, \bar{h}] \in \operatorname{int} K + \operatorname{lin}\{F(\bar{x})\} + \operatorname{Im} F'(\bar{x}), \tag{2.4}$$

where $T_K^r(y)$ stands for the so-called radial tangent cone to K at $y \in K$. This form of second-order regularity will be used in the subsequent analysis. We are now ready to state the main result of this section.

THEOREM 2.2. *Let X and Y be normed linear spaces and let K be a closed convex cone in Y with a nonempty interior. Let set D be given by* (1.1), *where $F : X \to Y$ is twice Fréchet-differentiable at a point $\bar{x} \in D$. Then the following statements hold.*

1. *Every $h \in T_D(\bar{x})$ satisfies* (2.2) *as well as the following condition:*

$$F''(\bar{x})[h]^2 \in \operatorname{cl}(K + \operatorname{lin}\{F(\bar{x})\} + \operatorname{Im} F'(\bar{x})). \tag{2.5}$$

2. *If $h \in X$ satisfies*

$$F'(\bar{x})h \in K + \operatorname{lin}\{F(\bar{x})\} \tag{2.6}$$

and (2.5), *and if constraints in* (1.1) *are second-order regular at $\bar{x}$ with respect to this h, then $h \in T_D(\bar{x})$.*

Proof. Take an arbitrary $h \in T_D(\bar{x})$. Relation (2.2) is standard, so we have to prove only (2.5). By twice differentiability of F, for every $t > 0$ we have that

$$\begin{aligned}\frac{1}{2}F''(\bar{x})[th]^2 &= F(\bar{x} + th + r(t)) - F(\bar{x}) - F'(\bar{x})(th + r(t)) \\ &\quad - \frac{1}{2}F''(\bar{x})[r(t)]^2 - F''(\bar{x})[th, r(t)] + \omega_2(t),\end{aligned}$$

where $\omega_2 : \Re_+ \to Y\,, \|\omega_2(t)\| = o(t^2)$. Observe that the first term in the right-hand side is in K due to (1.2), the second is in $\mathrm{lin}\{F(\bar{x})\}$, and the third is in $\mathrm{Im}F'(\bar{x})$. Dividing by t^2 and passing onto the limit as $t \to 0+$, we obtain (2.5).

Assume now that some $h \in X$ satisfies (2.6) and (2.5). Then there exist $y_1 \in K$ and $\lambda_1 \in \Re$ such that $F'(\bar{x})h = y_1 + \lambda_1 F(\bar{x})$. Consider first the case where

$$F''(\bar{x})[h]^2 \in \mathrm{int}\, K + \mathrm{lin}\{F(\bar{x})\} + \mathrm{Im}\, F'(\bar{x}), \tag{2.7}$$

so that there exist $y_2 \in \mathrm{int}\, K$, $\lambda_2 \in \Re$, and $x \in X$ such that $F''(\bar{x})[h]^2 = y_2 + \lambda_2 F(\bar{x}) + F'(\bar{x})x$. In that case, we obtain that

$$\begin{aligned}F\left(\bar{x} + th - \frac{t^2}{2}x\right) &= F(\bar{x}) + F'(\bar{x})\left(th - \frac{t^2}{2}x\right) \\ &\quad + \frac{1}{2}F''(\bar{x})\left[th - \frac{t^2}{2}x\right]^2 + \omega_2(t) \\ &= F(\bar{x}) + t(y_1 + \lambda_1 F(\bar{x})) - \frac{t^2}{2}F'(\bar{x})x \\ &\quad + \frac{t^2}{2}(y_2 + \lambda_2 F(\bar{x}) + F'(\bar{x})x) + \omega_2(t) \\ &= \left(1 + t\lambda_1 + \frac{t^2}{2}\lambda_2\right)F(\bar{x}) + ty_1 + \frac{t^2}{2}y_2 + \omega_2(t) \\ &\in \mathrm{int}\, K,\end{aligned}$$

where $\omega_2 : \Re_+ \to Y\,, \|\omega_2(t)\| = o(t^2)$, and the inclusion follows from Lemma A.5 for every $t > 0$ sufficiently small. In particular, we conclude that if (2.7) holds, then $h \in T_D(\bar{x})$.

If (2.7) does not hold, but there exists a sequence $\{h^k\} \subset X$ converging to h such that (2.7) is satisfied for every element of this sequence, then again $h \in T_D(\bar{x})$ by the closedness of $T_D(\bar{x})$. We proceed to explicitly construct the desired sequence $\{h^k\}$ under the hypothesis of the theorem that there exists an element $\bar{h} \in X$ for which (2.3), (2.4) are satisfied. Let us take $h^k = (1 - 1/k)h + \bar{h}/k$, $k = 1, 2, \ldots$. For each index k we then obtain

$$F'(\bar{x})h^k = \left(1 - \frac{1}{k}\right)F'(\bar{x})h + \frac{1}{k}F'(\bar{x})\bar{h} \in K + \mathrm{lin}\{F(\bar{x})\},$$

where the inclusion follows from (2.6), (2.3). We further obtain

$$\begin{aligned}F''(\bar{x})[h^k]^2 &= \left(1 - \frac{1}{k}\right)^2 F''(\bar{x})[h]^2 \\ &\quad + \frac{1}{k}\left(2\left(1 - \frac{1}{k}\right)F''(\bar{x})[h, \bar{h}] + \frac{1}{k}F''(\bar{x})[\bar{h}]^2\right) \\ &\in \mathrm{int}\, K + \mathrm{lin}\{F(\bar{x})\} + \mathrm{Im}\, F'(\bar{x}),\end{aligned}$$

where the inclusion holds for all k sufficiently large, due to (2.4), (2.5) and Lemmas A.2 and A.5. This construction completes the proof. □

In section 4, we compare this theorem (as well as the other results of this paper) with related facts and approaches to irregular inequality constraints and provide an illustrative example. Here, we note that in the regular case (1.4) implies that

$$K + \lin\{F(\bar{x})\} + \operatorname{Im} F'(\bar{x}) = Y, \tag{2.8}$$

and thus (2.5) holds trivially for every $h \in X$. This observation together with Remark 2.1 show that Theorem 2.2 subsumes (when K has nonempty interior) the classical result on the tangent cone in the regular case. In the irregular case, the right-hand side of (2.5) does not coincide with Y (again, in our setting of $\operatorname{int} K \neq \emptyset$), and therefore condition (2.5) is nontrivial.

Remark 2.2. If K is a finitely generated cone, then (2.6) is equivalent to (2.2), as the right-hand sides of these relations coincide (this follows from Lemma A.3). But in the general case, one cannot substitute the weaker condition (2.2) into the sufficiency part of the theorem, as illustrated by the following example.

EXAMPLE 2.1. Let $X = \Re$, $Y = \Re^3$, and

$$K = \operatorname{cone}\{y \in \Re^3 \mid y_1 = 1,\ y_3 = |y_2|^{3/2}\},$$
$$F : \Re \to \Re^3, \quad F(x) = (1,\, x,\, x^2).$$

For the point $\bar{x} = 0 \in \Re$, we have $F(0) \in K$, $\operatorname{cl}(K + \lin\{F(\bar{x})\}) = \operatorname{cl}(K + \lin\{F(\bar{x})\} + \operatorname{Im} F'(\bar{x})) = \{y \in \Re^3 \mid y_3 \geq 0\}$, and, as is easy to see, for element $h = 1$ conditions (2.3), (2.4) hold with $\bar{h} = h$. At the same time, 0 is obviously an isolated point of the set D given by (1.1), and hence $T_D(\bar{x}) = \{0\}$.

3. Optimality conditions. We now turn our attention to the optimization problem (1.5), where the feasible set is given by (1.1). We assume that K is a closed convex cone with nonempty interior (for primal-dual optimality conditions, also finitely generated), the objective function f is Fréchet-differentiable at the point $\bar{x} \in D$ under consideration, and the mapping F is twice Fréchet-differentiable at $\bar{x}$.

Following the developments of section 2, we first introduce some relevant cones. Let $H_2(\bar{x})$ be the set of all elements satisfying the second-order necessary conditions of tangency (2.2), (2.5) stated in Theorem 2.2, i.e.,

$$H_2(\bar{x}) := \left\{ h \in X \;\middle|\; \begin{array}{l} F'(\bar{x})h \in T_K(F(\bar{x})) = \operatorname{cl}(K + \lin\{F(\bar{x})\}) \\ F''(\bar{x})[h]^2 \in \operatorname{cl}(K + \lin\{F(\bar{x})\} + \operatorname{Im} F'(\bar{x})) \end{array} \right\},$$

and $\tilde{H}_2(\bar{x})$ be the set of elements satisfying the two relations (2.6) and (2.5), which appear in the sufficiency part:

$$\tilde{H}_2(\bar{x}) := \left\{ h \in X \;\middle|\; \begin{array}{l} F'(\bar{x})h \in T_K^r(F(\bar{x})) = K + \lin\{F(\bar{x})\} \\ F''(\bar{x})[h]^2 \in \operatorname{cl}(K + \lin\{F(\bar{x})\} + \operatorname{Im} F'(\bar{x})) \end{array} \right\}.$$

Finally, let $\bar{H}_2(\bar{x})$ consist of all elements satisfying the sufficient conditions of tangency stated in Theorem 2.2, i.e.,

$$\bar{H}_2(\bar{x}) := \left\{ h \in \tilde{H}_2(\bar{x}) \;\middle|\; \exists\, \bar{h} \in X \;:\; \begin{array}{l} F'(\bar{x})\bar{h} \in K + \lin\{F(\bar{x})\} \\ F''(\bar{x})[h,\, \bar{h}] \in \operatorname{int} K + \lin\{F(\bar{x})\} + \operatorname{Im} F'(\bar{x}) \end{array} \right\}.$$

By these definitions,

$$\bar{H}_2(\bar{x}) \cup \{0\} \subset \tilde{H}_2(\bar{x}) \subset H_2(\bar{x}). \tag{3.1}$$

Note that if the second-order regularity condition holds with respect to all $h \in \tilde{H}_2(\bar{x}) \setminus \{0\}$, then the first inclusion in (3.1) holds as an equality. If cone K is finitely generated, then the second inclusion is also an equality (recall Remark 2.2). By Theorem 2.2, we also have that

$$\bar{H}_2(\bar{x}) \cup \{0\} \subset T_D(\bar{x}) \subset H_2(\bar{x}). \tag{3.2}$$

If K is finitely generated and the second-order regularity condition holds with respect to all $h \in \tilde{H}_2(\bar{x}) \setminus \{0\}$, then we have equalities throughout (3.2).

The left-hand inclusion in (3.2) immediately implies the following primal necessary optimality condition for our problem.

THEOREM 3.1. *Let X and Y be normed linear spaces, and let K be a closed convex cone in Y with a nonempty interior. Assume that $f : X \to \Re$ is Fréchet-differentiable, and $F : X \to Y$ is twice Fréchet-differentiable at a point $\bar{x} \in D$, where D is given by* (1.1). *If $\bar{x}$ is a local solution of* (1.5), (1.1), *then*

$$\langle f'(\bar{x}),\, h\rangle \geq 0 \quad \forall\, h \in \bar{H}_2(\bar{x}). \tag{3.3}$$

If X is finite-dimensional, the right-hand inclusion in (3.2) implies that the following condition is sufficient for $\bar{x}$ to be a strict local solution of our problem:

$$\langle f'(\bar{x}),\, h\rangle > 0 \quad \forall\, h \in H_2(\bar{x}) \setminus \{0\}. \tag{3.4}$$

Dualizing (3.3), we can write that

$$f'(\bar{x}) \in (\bar{H}_2(\bar{x}))^*,$$

which is the primal-dual form of necessary optimality conditions. Explicit evaluation of the dual cone in the right-hand side of the above relation in full generality is an extremely difficult problem. However, we are able to give some meaningful results under additional assumptions. Specifically, if cone K is finitely generated and for some $h \in \bar{H}_2(\bar{x})$ the inequality in (3.3) holds as an equality, we derive an explicit primal-dual form of necessary optimality conditions. Note that further study of such "critical direction" h is of particular importance in view of the violation of the sufficient optimality condition (3.4). Assumptions of this type are quite common in the literature [7, 8, 25].

In the proof below, we shall also need the following generalization of the tangent cone description in the *regular* case. Let, in addition to our standard assumptions, C be a closed finitely generated cone in a normed linear space Z, and let $A : X \to Z$ be a continuous linear operator. Consider the set $\Delta = D \cap E$, where $E = \{x \in X \mid Ax \in C\}$, and a point $\bar{x} \in \Delta$. If there exists $\bar{\xi} \in X$ satisfying $A\bar{\xi} \in T_C(A\bar{x})$ and Robinson's condition (2.1), then

$$T_\Delta(\bar{x}) = \{h \in X \mid Ah \in T_C(A\bar{x}),\, F'(\bar{x})h \in T_K(F(\bar{x}))\}. \tag{3.5}$$

This generalization is essentially based on the well-known fact that linearity of constraints can be regarded as a special regularity-type assumption.

THEOREM 3.2. *Suppose that the assumptions of Theorem* 3.1 *are satisfied. Let K be a finitely generated cone, and let the point $\bar{x}$ be a local minimizer for problem* (1.5), (1.1). *Assume that*

$$\exists\, h \in \bar{H}_2(\bar{x}) \text{ such that } \langle f'(\bar{x}),\, h\rangle = 0. \tag{3.6}$$

Then there exist two functionals

$$y_1^* = y_1^*(h) \in K^* \cap \{F(\bar{x})\}^\perp \cap \{F'(\bar{x})h\}^\perp \tag{3.7}$$

and

$$y_2^* = y_2^*(h) \in K^* \cap \{F(\bar{x})\}^\perp \cap (\operatorname{Im} F'(\bar{x}))^\perp \cap \{F''(\bar{x})[h]^2\}^\perp \tag{3.8}$$

such that

$$f'(\bar{x}) = (F'(\bar{x}))^* y_1^* + (F''(\bar{x})[h])^* y_2^*. \tag{3.9}$$

Proof. It can be easily seen that there exists a neighborhood U of h in X such that

$$H_2(\bar{x}) \cap U \subset \bar{H}_2(\bar{x}).$$

(Just recall that since cone K is finitely generated, the second inclusion in (3.1) holds as an equality, and observe that for a neighborhood U small enough, one can choose the same $\bar{h}$ in the definition of $\bar{H}_2(\bar{x})$ for all $h \in U$.) Hence, by Theorem 3.1, we have that

$$\langle f'(\bar{x}), \xi \rangle \geq 0 \quad \forall\, \xi \in H_2(\bar{x}) \cap U.$$

The latter relation and (3.6) imply that h is a local solution of the optimization problem

$$\min \{ \langle f'(\bar{x}), \xi \rangle \mid \xi \in H_2(\bar{x}) \}.$$

By the classical necessary optimality conditions, it then follows that

$$\langle f'(\bar{x}), \xi \rangle \geq 0 \quad \forall\, \xi \in T_{H_2(\bar{x})}(h),$$

or, equivalently,

$$f'(\bar{x}) \in (T_{H_2(\bar{x})}(h))^*. \tag{3.10}$$

We now have to evaluate the cone $T_{H_2(\bar{x})}(h)$ and its dual. The latter problem is now solvable with the help of Lemma A.4, because our second-order regularity condition with respect to h implies that the cone $T_{H_2(\bar{x})}(h)$ is actually given by the linearized model of constraints defining $H_2(\bar{x})$. Indeed, using the assumption that cone K is closed and finitely generated, and applying Lemma A.3 and relation (3.5) to appropriate data, we obtain

$$T_{H_2(\bar{x})}(h) = \left\{ \xi \in X \,\middle|\, \begin{matrix} F'(\bar{x})\xi \in K + \operatorname{lin}\{F(\bar{x})\} + \operatorname{lin}\{F'(\bar{x})h\} \\ F''(\bar{x})[h,\, \xi] \in \operatorname{cl}(K + \operatorname{lin}\{F(\bar{x})\} + \operatorname{Im} F'(\bar{x}) + \operatorname{lin}\{F''(\bar{x})[h]^2\}) \end{matrix} \right\}. \tag{3.11}$$

Note that cone $K + \operatorname{lin}\{F(\bar{x})\} + \operatorname{lin}\{F'(\bar{x})h\}$ is closed and finitely generated. Also, $\dim Y < \infty$. (This is implied by our assumption that a finitely generated cone K has nonempty interior.) In particular, it follows that $\dim(\operatorname{Im} F'(\bar{x})) < \infty$. Hence, cone $K + \operatorname{lin}\{F(\bar{x})\} + \operatorname{Im} F'(\bar{x}) + \operatorname{lin}\{F''(\bar{x})[h]^2\}$ is also closed and finitely generated. Now applying Lemma A.4 to (3.11), we obtain the equality

$$\begin{aligned} (T_{H_2(\bar{x})}(h))^* = {} & (F'(\bar{x}))^*(K^* \cap \{F(\bar{x})\}^\perp \cap \{F'(\bar{x})h\}^\perp) \\ & + (F''(\bar{x})[h])^*(K^* \cap \{F(\bar{x})\}^\perp \cap (\operatorname{Im} F'(\bar{x}))^\perp \cap \{F''(\bar{x})[h]^2\}^\perp), \end{aligned}$$

from which the conclusion of the theorem follows immediately. ∎

Theorem 3.2 subsumes classical first-order necessary optimality conditions for the regular case. Indeed, suppose that h in the requirements of Theorem 3.2 satisfies (2.7). Note that this will always be so in the regular case because, by (2.8) and Lemma A.2, the right-hand side of (2.7) coincides with the entire space Y. Then, using Lemma A.1, we have that

$$K^* \cap \{F(\bar{x})\}^\perp \cap (\operatorname{Im} F'(\bar{x}))^\perp \cap \{F''(\bar{x})[h]^2\}^\perp = \{0\}. \tag{3.12}$$

Therefore in that case $y_2^* = 0$, and representation (3.7)–(3.9) reduces to

$$f'(\bar{x}) = (F'(\bar{x}))^* y_1^*, \tag{3.13}$$

with y_1^* satisfying (3.7). Furthermore, by Remark 2.1, in the regular case Theorem 3.2 can be applied by choosing $h = 0$. With this choice, (3.7) takes the form

$$y_1^* \in K^* \cap \{F(\bar{x})\}^\perp. \tag{3.14}$$

Combined with feasibility condition $F(\bar{x}) \in K$, relations (3.13), (3.14) coincide with the classical optimality conditions (1.6), where the nonsingular multiplier $y_0 = 1$ is chosen. In terms of the nonlinear programming problem, the inclusion $y_1^* \in K^*$ is the nonnegativity condition for the Lagrange multipliers, and the inclusion $y_1^* \in \{F(\bar{x})\}^\perp$ is the condition of complementary slackness.

As will be shown in section 4, Theorem 3.2 also contains optimality conditions under the second-order regularity of [7, 8] but can be applicable when the latter is not.

4. Comparisons and an example. In this section, we provide a comparison of the results obtained above with known approaches to irregular problems, and illustrate our development by an example.

First, we mention Abadie's and Kuhn–Tucker's constraint qualifications (CQs) for nonlinear programming (see [22]; there are also some other CQs of similar type). These are weaker than the Mangasarian–Fromovitz constraint qualification (MFCQ) but still guarantee that the tangent cone is given by the linearized model of the constraints; e.g., see [23, 22]. From the point of view of the problem data, these CQs are less constructive than MFCQ, which is closer to our development. (MFCQ is subsumed by our framework.) Such CQs of nonalgebraic nature are usually rather difficult to verify directly. Perhaps even more importantly, we deal here with a more general case in which the tangent cone does not necessarily coincide with the linearized cone.

The next issue that deserves to be discussed is reformulating inequality constraints as equalities, with the aim of subsequently using results available for the latter. This technique is known to be useful for regular inequality-constrained problems; e.g., see [9]. Analogously, one might try to apply known optimality conditions for (irregular) equality-constrained problems to reformulations of irregular inequality constraints. For example, the theory of 2-regularity [29, 4, 6, 5, 16, 1, 13, 20, 17, 15] offers optimality conditions for the case in which irregularity of the problem is induced only by equality constraints, with inequality constraints being either absent or regular. We next show that in our context, applicability of this approach is very limited.

For simplicity, let us take $Y = \Re^m$, $K = \Re^m_-$, and $F(\bar{x}) = 0$, and reformulate the inequality-constrained set D by introducing slacks:

$$\Delta = \{(x, u) \in X \times \Re^m \mid F(x) + u = 0,\ u \geq 0\}.$$

The new set Δ is given by equality and "simple" inequality constraints. Clearly, the equality constraint in Δ is regular at every point, but MFCQ is still violated at $(\bar{x}, 0)$. Hence, the classical results for the regular case are not applicable. Results from the theory of 2-regularity are obviously also not useful, as there are simply no irregular equality constraints in Δ.

Another possibility is a purely equality-constrained reformulation:

$$\Delta = \{(x, u) \in X \times \Re^m \mid F(x) + u^2 = 0\},$$

where the square is componentwise. Here, the equality constraint is irregular at $(\bar{x}, 0)$, and 2-regularity theory is applicable, at least formally. However, this application leads to something meaningful only when $\ker F'(\bar{x}) \neq \{0\}$, which is an unnatural requirement for inequality constraints. Our approach is certainly free of this restriction. Moreover, even if $\ker F'(\bar{x}) \neq \{0\}$, for inequality constraints this subspace can have little to do with the tangent cone, as in Example 4.1 below. Without going into detail, we shall mention that there are also some other limitations in the "brute force" approach of applying results known for irregular equality constraints to equation reformulations of irregular inequality constraints. It seems that developing a special approach specifically designed for inequality constraints is really necessary. An initial step in the direction pursued in the present paper was made in [14].

Another known approach to irregular problems consists of second-order necessary and sufficient optimality conditions of Levitin–Milyutin–Osmolovskii type, e.g., [21, 18, 7, 8, 1, 2] (see also recent work in [10, 25]), which employ F. John first-order necessary conditions (with undefined multiplier corresponding to the objective function). This approach is effective when applied to inequality-constrained problems, but it leads to results of a completely different nature, which makes comparison with the present paper difficult. We note that this approach is not principally associated with precise description of the tangent cone, i.e., it does not deal with sufficient conditions for tangency beyond the regular case.

Next, we discuss the well-known second-order CQ [7, 8], which was introduced using second-order parabolic tangent sets, and which is especially relevant for irregular inequality-constrained problems. In our setting, this CQ can be stated as follows:

$$\exists\, h \in X \text{ such that } \langle f'(\bar{x}), h\rangle = 0, \tag{4.1}$$

$$F'(\bar{x})h \in K + \operatorname{lin}\{F(\bar{x})\}, \tag{4.2}$$

$$F''(\bar{x})[h]^2 \in \operatorname{int} K + \operatorname{lin}\{F(\bar{x})\} + \operatorname{Im} F'(\bar{x}). \tag{4.3}$$

This condition is also weaker than Robinson's regularity (in the regular case, (4.1)–(4.3) hold with $h = 0$), yet it guarantees that if $\bar{x}$ is a local solution of (1.5), (1.1), then F. John-necessary conditions are satisfied with a nonzero multiplier corresponding to the objective function. Note that relations (4.2) and (4.3) already appear in Theorem 2.2 (see (2.6) and (2.7)), where they are used to explicitly construct a parabolic feasible arc tangent to h. But observe that in Theorem 2.2 we consider a larger set of directions. Namely, for an element h satisfying second-order necessary conditions of tangency, this theorem gives constructive sufficient conditions for h to be a limit point of elements satisfying (4.2), (4.3). This is important, because it is certainly possible that (4.1) does not hold for any h satisfying (4.2), (4.3), but that it does hold for some limit point of such elements. Moreover, Example 4.1 below illustrates that this situation (i.e., the second-order CQ (4.1)–(4.3) is violated, but our Theorem 3.2 is applicable) is in fact quite likely to occur.

Finally, note that if h is an element satisfying (4.1)–(4.3), then (3.6) also holds, and the assumptions of Theorem 3.2 are satisfied. Moreover, in this case, (3.12) holds. Hence, relation (3.8) in Theorem 3.2 implies that $y_2^* = 0$. We conclude that optimality conditions under the second-order CQ (4.1)–(4.3) are a particular case of Theorem 3.2 (under the additional assumption that K is finitely generated).

To complete this section, we present an example illustrating all the results derived above, and showing that they can be applicable when the F. John-optimality conditions and optimality conditions based on classical (first- and second-order) CQs are not useful. Note that our example is not pathological or exotic.

EXAMPLE 4.1. Let $X = Y = \Re^2$, $K = \Re^2_-$, and consider a family of functions

$$f : \Re^2 \to \Re, \quad f(x) = ax_1 + bx_2 + \omega_1(x)$$

and the mapping

$$F : \Re^2 \to \Re^2, \quad F(x) = \left(-x_1, -\frac{1}{2}(x_1^2 - x_2^2)\right) + \omega_2(x),$$

where $\omega_1 : \Re^2 \to \Re$, $|\omega_1(x)| = o(\|x\|)$, and $\omega_2 : \Re^2 \to \Re^2$, $\|\omega_2(x)\| = o(\|x\|^2)$.

Consider the point $\bar{x} = 0$ in $\Re^2$. We have that $F(0) = 0$, so that $0 \in D$, where D is given by (1.1). It can be easily seen that MFCQ does not hold here, and so classical theory does not apply. By direct computations, we obtain that

$$H_2(0) = \tilde{H}_2(0) = \{h \in \mathbf{R}^2 \mid h_1 \geq 0,\ h_1^2 - h_2^2 \geq 0\},$$
$$\bar{H}_2(0) = \{h \in H_2(0) \mid \exists\, \bar{h} \in \mathbf{R}^2 : \bar{h}_1 \geq 0,\ h_1\bar{h}_1 - h_2\bar{h}_2 > 0\} = H_2(0).$$

Hence, by Theorem 2.2,

$$T_D(0) = H_2(0) = \{h \in \mathbf{R}^2 \mid h_1 \geq |h_2|\},$$

which is actually geometrically obvious. Observe further that the linearized cone is given by

$$\{h \in \mathbf{R}^2 \mid F'(\bar{x})h \in T_K(F(\bar{x}))\} = \{h \in \mathbf{R}^2 \mid h_1 \geq 0\},$$

which is different from $T_D(0)$. Hence, the Kuhn–Tucker, Abadie, and any other CQs guaranteeing that the tangent and linearized cones coincide are violated in this example. Note that in this case, the tangent cone is actually polyhedral, just different from the linearized one. This shows that our description can be useful even when the tangent cone is "simple."

It is easy to see that for all values of parameters a and b, the F. John conditions (1.6) for problem (1.5), (1.1) hold at 0 with $y_0 = 0$. Furthermore, y_0 can be nonzero only if $b = 0$ and $a \leq 0$. For all other values of the parameters, F. John conditions are not meaningful for describing optimality.

As is easy to see, the set of elements satisfying (4.2), (4.3) is $\{h \in \mathbf{R}^2 \mid h_1 > |h_2|\}$. Clearly, if 0 is a local minimizer, conditions (4.1)–(4.3) can hold for some h simultaneously only if $a = b = 0$. Hence, for all other values of the parameters, the classical second-order CQ (4.1)–(4.3) does not hold, and the corresponding results are not applicable.

We next illustrate our approach, considering several characteristic values of the parameters.

If $a = 1$, $b = -1$, then 0 is a (nonisolated) local minimizer for problem (1.5), (1.1). As is easy to see,

$$\langle f'(0), h\rangle \geq 0 \quad \forall\, h \in H_2(0), \tag{4.4}$$

which illustrates Theorem 3.1. Note that for $h = (1, 1) \in H_2(0)$, the latter inequality holds as equality, and our primal-dual optimality conditions (3.7)–(3.9) are satisfied with the multipliers

$$y_1^* = (0, \alpha) \in \Re^2, \quad \alpha \in \Re_-, \quad y_2^* = (0, -1) \in \Re^2\,.$$

This gives an illustration for Theorem 3.2. Note that for $h \in H_2(0) \setminus \mathrm{lin}\{(1, 1)\}$, a similar representation does not hold. The reason is that for such h, strict inequality holds in (4.4).

If $a = 1$, $b = 0$, then (4.4) holds as a strict inequality for every $h \in H_2(0) \setminus \{0\}$, and 0 is an isolated local minimizer. This illustrates sufficient optimality condition (3.4).

Finally, if $a = 0$, $b = 1$, then it is easy to see that (4.4) does not hold for those elements $h \in H_2(0)$ for which $h_2 < 0$. Theorem 3.1 implies that 0 is not a local minimizer in this case. We could similarly use Theorem 3.2 to verify this conclusion. Indeed, for the element $h = (1, 0) \in H_2(0)$, (4.4) holds as an equality, but there exist no multipliers y_1^*, $y_2^* \in \Re^2$ for which (3.9) holds.

5. Some further developments. In conclusion, we present some further developments of the optimality conditions obtained above. The first one has to do with a certain form of second-order (in terms of the objective function) necessary optimality conditions, and the second outlines an extension to mixed equality–inequality-constrained problems.

5.1. Second-order optimality conditions. To derive second-order optimality conditions, we need the following notion. Let X and Σ be normed linear spaces, and let a mapping $\Phi : X \to \Sigma$ be twice Fréchet-differentiable at a point $\bar{x} \in X$. Suppose that $\Sigma_1 = \mathrm{Im}\,\Phi'(\bar{x})$ is closed and has a closed complementary subspace Σ_2 in Σ. Let P be a projector onto Σ_2 parallel to Σ_1 in Σ. (By assumptions above, this projector is continuous.) In this setting, the mapping Φ is referred to as 2-*regular at the point* $\bar{x}$ *with respect to an element* $h \in X$ (see [29, 4, 6, 5, 16, 1, 13, 20, 17]) if

$$\mathrm{Im}(\Phi'(\bar{x}) + P\Phi''(\bar{x})[h]) = \Sigma.$$

We note that the 2-regularity property of Φ does not depend on a choice of the complementary subspace Σ_2.

The following generalization of the classical Lyusternik's theorem can be found in [29, 5, 16, 20, 17].

PROPOSITION 5.1. *Let X and Σ be Banach spaces. Assume that a mapping $\Phi : X \to \Sigma$ is three times Fréchet-differentiable at a point $\bar{x} \in X$ such that $\Phi(\bar{x}) = 0$. Assume further that Φ is 2-regular at $\bar{x}$ with respect to an element $h \in X$ such that*

$$h \in \mathrm{Ker}\,\Phi'(\bar{x}), \quad \Phi''(\bar{x})[h]^2 \in \mathrm{Im}\,\Phi'(\bar{x}).$$

Then there exist a number $\delta > 0$ and a mapping $r : (-\delta, \delta) \to X$ such that

$$\Phi(\bar{x} + th + r(t)) = 0 \ \ \forall\, t \in (-\delta, \delta)\,, \quad \|r(t)\| = O(t^2).$$

We next derive a special form of higher-order necessary optimality conditions using the results obtained in section 3.

THEOREM 5.2. *Let X and Y be Banach spaces, let K be a closed finitely generated cone in Y with a nonempty interior, and let $f : X \to \Re$ be twice and $F : X \to Y$ be three times Fréchet-differentiable at the point $\bar{x}$, which is a local minimizer for problem* (1.5), (1.1). *Assume that* (3.6) *holds, and let $\tilde{\Pi}$ be a (continuous) projector onto some closed complementary subspace $\tilde{Y}$ of* $\lin\{F(\bar{x}),\, F'(\bar{x})h\}$ *in Y. Assume finally that*

$$\tilde{\Pi}F''(\bar{x})[h]^2 \in \tilde{\Pi}\,\mathrm{Im}\,F'(\bar{x}) \tag{5.1}$$

and that the mapping $\Phi : X \to \tilde{Y}$, $\Phi(x) = \tilde{\Pi}F(x)$, is 2-regular at the point $\bar{x}$ with respect to h. Then for every $y_1^*, y_2^* \in Y^*$ satisfying (3.7)–(3.9), *it holds that*

$$f''(\bar{x})[h]^2 - \langle y_1^*,\, F''(\bar{x})[h]^2\rangle - \frac{1}{3}\langle y_2^*,\, F'''(\bar{x})[h]^3\rangle \geq 0. \tag{5.2}$$

Proof. By the definition of $\tilde{\Pi}$, we have

$$\Phi'(\bar{x})h = \tilde{\Pi}F'(\bar{x})h = 0.$$

Hence, taking into account (5.1), Proposition 5.1 is applicable (with $\Sigma = \tilde{Y}$). So for some number $\delta > 0$ and some mapping $r : (-\delta, \delta) \to X$, we have that $\forall\, t \in (-\delta, \delta)$

$$\tilde{\Pi}F(\bar{x} + th + r(t)) = 0, \quad \|r(t)\| = O(t^2),$$

where the first equality means that

$$F(\bar{x} + th + r(t)) \in \lin\{F(\bar{x}),\, F'(\bar{x})h\}. \tag{5.3}$$

By (3.7), $y_1^* \in (\lin\{F(\bar{x}),\, F'(\bar{x})h\})^\perp$. Hence, $\forall\, t \in (-\delta, \delta)$ we have

$$\begin{aligned} 0 &= \langle y_1^*,\, F(\bar{x} + th + r(t))\rangle \\ &= \langle y_1^*,\, F'(\bar{x})r(t)\rangle + \frac{1}{2}\langle y_1^*,\, F''(\bar{x})[th]^2\rangle + o(t^2). \end{aligned} \tag{5.4}$$

Similarly, by (3.8), $y_2^* \in (\lin\{F(\bar{x}),\, F'(\bar{x})h\})^\perp$ and also $y_2^* \in (\mathrm{Im}\,F'(\bar{x}))^\perp$, which implies that

$$\begin{aligned} 0 &= \langle y_2^*,\, F(\bar{x} + th + r(t))\rangle \\ &= \langle y_2^*,\, F''(\bar{x})[th,\, r(t)]\rangle + \frac{1}{6}\langle y_2^*,\, F'''(\bar{x})[th]^3\rangle + o(t^3). \end{aligned} \tag{5.5}$$

By (5.3), there exist $\lambda_1,\, \lambda_2 : (-\delta,\, \delta) \to \Re$ such that

$$F(\bar{x} + th + r(t)) = \lambda_1(t)F(\bar{x}) + \lambda_2(t)F'(\bar{x})h.$$

On the other hand, by differentiability of F,

$$F(\bar{x} + th + r(t)) = F(\bar{x}) + tF'(\bar{x})h + o(t).$$

Therefore, we can take $\lambda_1(t) = 1 + o(t),\ \lambda_2(t) = t + o(t)$. Since $h \in \tilde{H}_2(\bar{x})$, we have that $F'(\bar{x})h = y + \lambda F(\bar{x})$ for some $y \in K$, $\lambda \in \Re$. We further obtain

$$\begin{aligned} F(\bar{x} + th + r(t)) &= (1 + o(t))F(\bar{x}) + (t + o(t))(y + \lambda F(\bar{x})) \\ &= (1 + \lambda t + o(t))F(\bar{x}) + (t + o(t))y. \end{aligned}$$

Taking into account that $F(\bar{x}) \in K$ and $y \in K$, it is clear now that if $\delta > 0$ is small enough, then $\bar{x} + th + r(t) \in D \ \forall\, t \in (0, \delta)$, and since $\bar{x}$ is a local minimizer, by differentiability of f it follows that $\forall\, t \in (0, \delta)$

$$0 \le f(\bar{x} + th + r(t)) - f(\bar{x}) = \langle f'(\bar{x}),\, r(t)\rangle + \frac{1}{2} f''(\bar{x})[th]^2 + o(t^2),$$

where we have also used (3.6). Combining the latter relation with (5.4) and (5.5) (divided by -1 and $-t$, respectively), we obtain

$$\begin{aligned}
0 &\le \langle f'(\bar{x}),\, r(t)\rangle - \langle y_1^*,\, F'(\bar{x})r(t)\rangle - \langle y_2^*,\, F''(\bar{x})[h, r(t)]\rangle \\
&\quad + \frac{1}{2} f''(\bar{x})[th]^2 - \frac{1}{2}\langle y_1^*,\, F''(\bar{x})[th]^2\rangle - \frac{1}{6t}\langle y_2^*,\, F'''(\bar{x})[th]^3\rangle + o(t^2) \\
&= \langle f'(\bar{x}) - (F'(\bar{x}))^* y_1^* - (F''(\bar{x})[h])^* y_2^*,\, r(t)\rangle \\
&\quad + \frac{t^2}{2}\left(f''(\bar{x})[h]^2 - \langle y_1^*,\, F''(\bar{x})[h]^2\rangle - \frac{1}{3}\langle y_2^*,\, F'''(\bar{x})[h]^3\rangle \right) + o(t^2) \\
&= \frac{t^2}{2}\left(f''(\bar{x})[h]^2 - \langle y_1^*,\, F''(\bar{x})[h]^2\rangle - \frac{1}{3}\langle y_2^*,\, F'''(\bar{x})[h]^3\rangle \right) + o(t^2),
\end{aligned}$$

where the last equality follows from (3.9). Dividing by $t^2/2$ and passing onto the limit as $t \to 0$, we obtain (5.2). □

Note that the mapping Φ defined in Theorem 5.2 could be regular (rather than 2-regular) only if the Robinson's regularity condition were to be satisfied at $\bar{x}$.

The next example illustrates that Theorem 5.2 provides additional information that can be used to eliminate candidates for optimality.

EXAMPLE 5.1. Consider the setting of Example 4.1, where $a = 1$, $b = -1$, $\omega_2(\cdot) \equiv 0$ on $\Re^2$, and $\omega_1 : \Re^2 \to \Re$ is a quadratic form negative on $h = (1, 1)$. Then the first-order necessary conditions given by Theorems 3.1 and 3.2 are satisfied at 0 (see Example 4.1), but by direct inspection it can be seen that the second-order necessary optimality conditions given by Theorem 5.2 are violated. Indeed, $F'(\bar{x})h = (-1, 0)$, and so one can take $\tilde{Y} = \mathrm{lin}\{(0, 1)\}$. Then Φ can be considered as a scalar-valued function

$$\Phi : \Re^2 \to \Re, \quad \Phi(x) = -\frac{1}{2}(x_1^2 - x_2^2).$$

This function is certainly 2-regular at 0 with respect to every nonzero element. (For scalar-valued functions, the latter property is equivalent to saying that 0 is a nondegenerate critical point [3].) In particular, Φ is 2-regular at 0 with respect to h, which obviously satisfies (5.1.) We further have that

$$f''(\bar{x})[h]^2 - \langle y_1^*,\, F''(\bar{x})[h]^2\rangle - \frac{1}{3}\langle y_2^*,\, F'''(\bar{x})[h]^3\rangle = 2\omega_1(h) < 0,$$

which is in contradiction with (5.2). We conclude that 0 is not a local minimizer for problem (1.5), (1.1).

5.2. Mixed equality and inequality constraints. In contrast to the regular case, it appears very difficult (if not impossible) to extend the results for irregular equality- or inequality-constrained problems to the case with mixed inequality and equality constraints, except for some special cases. (For a complete modification of this kind, one would have to avoid the condition that cone K has a nonempty

interior.) One special case, specifically where the singularity/irregularity is due to equality-type constraints only, is studied thoroughly in [5, 20] (those results were already mentioned in section 4). Let us consider briefly the opposite case, i.e., where irregularity is induced by inequality constraints, while equality constraints are regular. Let set D now be given by

$$D = \{x \in X \mid F(x) \in K,\ G(x) = 0\}. \tag{5.6}$$

Assume $G : X \to Z$ is three times continuously differentiable, where X and Z are Banach spaces. Suppose G is regular at a point $\bar{x} \in D$, i.e.,

$$\operatorname{Im} G'(\bar{x}) = Z,$$

and there exists a continuous projector Π on $\operatorname{Ker} G'(\bar{x})$ in Z. According to the classical facts of nonlinear analysis (see, e.g., [3, 13]), under those assumptions there exist a neighborhood U of 0 in X and a mapping $\rho : U \to X$ such that $\rho(0) = \bar{x}$, $\rho(U)$ is a neighborhood of $\bar{x}$ in X, ρ is a C^3-diffeomorphism from U onto $\rho(U)$, and

$$\begin{aligned} G(\rho(x)) &= G'(\bar{x})x \quad \forall\, x \in U, \\ \rho'(x) &= (R(x))^{-1}R(0) \quad \forall\, x \in U, \end{aligned} \tag{5.7}$$

where

$$R(x) : X \to Y \times \operatorname{Ker} G'(\bar{x}), \quad R(x)\xi = (G'(\rho(x))\xi,\ \Pi\xi), \quad x \in U.$$

Now instead of a feasible point $\bar{x}$ of problem (1.5), (5.6), we can consider for local analysis the feasible point 0 of the *inequality-constrained* problem

$$\min\{\varphi(x) \mid x \in \Delta\}, \quad \Delta = \{x \in \tilde{X} \cap U \mid \Phi(x) \in K\},$$

where $\tilde{X} = \operatorname{Ker} G'(\bar{x})$,

$$\varphi(x) = f(\rho(x)), \quad \Phi(x) = F(\rho(x)), \quad x \in U.$$

Note that taking advantage of (5.7), it is easy to obtain explicit formulas for the first three derivatives of φ and Φ, and so the analysis developed in this paper is applicable to the derivation of optimality conditions for problem (1.5), (5.6).

Appendix. Auxiliary results. All results in this section can be found in standard books on convex analysis [28, 3, 19, 26] or follow from results contained therein.

LEMMA A.1. *Let Σ be a topological linear space, L be a linear subspace in Σ, and C be a convex cone in Σ such that* $\operatorname{int} C \neq \emptyset$. *Then*

$$\operatorname{int} C \cap L = \emptyset \quad \Leftrightarrow \quad C^* \cap L^\perp \neq \{0\}.$$

LEMMA A.2. *Let Σ be a topological linear space and Ω_1, Ω_2 be convex sets in Σ, with* $\operatorname{int} \Omega_1 \neq \emptyset$. *Then*

$$\operatorname{int}(\Omega_1 + \Omega_2) = \operatorname{int} \Omega_1 + \Omega_2.$$

LEMMA A.3. *Let Σ be a normed linear space and C_1, C_2 be finitely generated cones in Σ. Then*

$$\operatorname{cl}(C_1 + C_2) = \operatorname{cl} C_1 + \operatorname{cl} C_2.$$

LEMMA A.4. *Let Υ and Σ be normed linear spaces,* $\dim \Sigma < \infty$, $\Lambda : \Upsilon \to \Sigma$ *be a continuous linear operator, and C be a nonempty closed finitely generated cone in Σ. Then for a cone $\Gamma = \{\xi \in \Upsilon \mid \Lambda\xi \in C\}$ it holds that*

$$\Gamma^* = \Lambda^* C^*.$$

LEMMA A.5. *Let Σ be a locally convex topological linear space, and C be a convex cone in Σ. Then*

$$\sigma^1 \in \operatorname{cl} C\,,\ \sigma^2 \in \operatorname{int} C \quad \Rightarrow \quad \sigma^1 + \sigma^2 \in \operatorname{int} C.$$

Acknowledgments. We are grateful to the two anonymous referees and the editor for constructive suggestions which led to considerable improvement of the paper.

REFERENCES

[1] A. V. ARUTYUNOV, *Extremum Conditions. Abnormal and Degenerate cases* (in Russian), Factorial, Moscow, Russia, 1997.
[2] A. V. ARUTYUNOV, *Second-order conditions in extremal problems: The abnormal points*, Trans. Amer. Math. Soc., 350 (1998), pp. 4341–4365.
[3] J. P. AUBIN AND I. EKELAND, *Applied Nonlinear Analysis*, Wiley, New York, 1984.
[4] Y. R. AVAKOV, *Extremum conditions for smooth problems with equality-type constraints*, USSR Comput. Math. and Math. Phys., 25 (1995), pp. 24–32.
[5] Y. R. AVAKOV, *Necessary extremum conditions for smooth abnormal problems with equality- and inequality constraints*, J. Math. Notes, 45 (1989), pp. 431–437.
[6] K. N. BELASH AND A. A. TRETYAKOV, *Methods for solving singular problems*, Comput. Math. Math. Phys., 28 (1988), pp. 1097–1102.
[7] J. F. BEN-TAL, *Second-order and related extremality conditions in nonlinear programming*, J. Optim. Theory Appl., 31 (1980), pp. 143–165.
[8] J. F. BEN-TAL AND J. ZOWE, *A unified theory of first and second order optimality conditions for extremum problems in topological vector spaces*, Math. Programming Study, 19 (1982), pp. 39–76.
[9] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.
[10] J. F. BONNANS, R. COMINETTI, AND A. SHAPIRO, *Second order optimality conditions based on parabolic second order tangent sets*, SIAM J. Optim., 9 (1999), pp. 466–492.
[11] J. F. BONNANS AND A. SHAPIRO, *Optimization problems with perturbations: A guided tour*, SIAM Rev., 40 (1998), pp. 228–264.
[12] J. F. BONNANS AND A. SHAPIRO, *Perturbation Analysis of Optimization Problems*, Springer-Verlag, New York, 2000.
[13] A. F. IZMAILOV, *On certain generalizations of Morse's lemma*, Proc. Steklov Inst. Math., 220 (1998), pp. 138-153.
[14] A. F. IZMAILOV, *Optimality conditions for degenerate extremum problems with inequality-type constraints*, Comput. Math. Math. Phys., 34 (1994), pp. 723–736.
[15] A. F. IZMAILOV AND M. V. SOLODOV, *The theory of 2-regularity for mappings with Lipschitzian derivatives and its applications to optimality conditions*, Math. Oper. Res., to appear.
[16] A. F. IZMAILOV AND A. A. TRETYAKOV, *Factor analysis of Nonlinear Mappings* (in Russian), Nauka, Moscow, Russia, 1994.
[17] A. F. IZMAILOV AND A. A. TRETYAKOV, *2-Regular Solutions of Nonlinear Problems. Theory and Numerical Methods* (in Russian), Fizmatlit, Moscow, Russia, 1999.
[18] A. D. IOFFE, *Necessary and sufficient conditions for a local minimum*, 3: *Second order conditions and augmented duality*, SIAM J. Control Optim., 17 (1979), pp. 266-288.
[19] A. D. IOFFE AND V. M. TIKHOMIROV, *Theory of Extremal Problems*, North–Holland, Amsterdam, The Netherlands, 1974.
[20] U. LEDZEWICZ AND H. SCHÄTTLER, *High-order approximations and generalized necessary conditions for optimality*, SIAM J. Control Optim., 37 (1998), pp. 33–53.
[21] E. S. LEVITIN, A. A. MILYUTIN, AND N. P. OSMOLOVSKII, *Conditions of higher order for a local minimum in problems with constraints*, Russian Math. Surveys, 33 (1978), pp. 97–168.
[22] O. L. MANGASARIAN, *Nonlinear Programming*, McGraw–Hill, New York, 1969.

[23] O. L. Mangasarian and S. Fromovitz, *The Fritz John necessary optimality conditions in the presence of equality and inequality constraints*, J. Math. Anal. Appl., 7 (1967), pp. 37–47.
[24] J.-P. Penot, *On regularity conditions in mathematical programming*, Math. Programming Study, 19 (1982), pp. 167–199.
[25] J.-P. Penot, *Second-order conditions for optimization problems with constraints*, SIAM J. Control Optim., 37 (1999), pp. 303–318.
[26] B. N. Pshenichniy, *Necessary Conditions of an Extremum* (in Russian), Nauka, Moscow, Russia, 1982.
[27] S. M. Robinson, *Stability theorems for systems of inequalities, Part* II: *Differentiable nonlinear systems*, SIAM J. Numer. Anal., 13 (1976), pp. 497–513.
[28] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[29] A. A. Tretyakov, *Necessary and sufficient conditions for optimality of p-th order*, Comput. Math. Math. Phys., 24 (1984), pp. 123–127.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 4, pp. 1296–1311

INDEFINITE STOCHASTIC LINEAR QUADRATIC CONTROL AND GENERALIZED DIFFERENTIAL RICCATI EQUATION*

M. AIT RAMI†, J. B. MOORE‡, AND XUN YU ZHOU†

Abstract. A stochastic linear quadratic (LQ) control problem is indefinite when the cost weighting matrices for the state and the control are allowed to be indefinite. Indefinite stochastic LQ theory has been extensively developed and has found interesting applications in finance. However, there remains an outstanding open problem, which is to identify an appropriate Riccati-type equation whose solvability is *equivalent* to the solvability of the indefinite stochastic LQ problem. This paper solves this open problem for LQ control in a finite time horizon. A new type of differential Riccati equation, called the generalized (differential) Riccati equation, is introduced, which involves algebraic equality/inequality constraints and a matrix pseudoinverse. It is then shown that the solvability of the generalized Riccati equation is not only sufficient, but also *necessary*, for the well-posedness of the indefinite LQ problem and the existence of optimal feedback/open-loop controls. Moreover, all of the optimal controls can be identified via the solution to the Riccati equation. An example is presented to illustrate the theory developed.

Key words. stochastic LQ control, indefinite costs, generalized Riccati equation, matrix pseudoinverse, matrix minimum principle, dynamic programming

AMS subject classifications. 93E20, 49K45

PII. S0363012900371083

1. Introduction.

Consider the following stochastic linear quadratic (LQ) optimal control problem in a finite time horizon $[0, T]$:

(1.1)
$$\text{Minimize} \quad J = E\left\{\int_0^T [x(t)'Q(t)x(t) + u(t)'R(t)u(t)]dt + x(T)'Hx(T)\right\},$$
$$\text{subject to} \quad \begin{cases} dx(t) = [A(t)x(t) + B(t)u(t)]dt + [C(t)x(t) + D(t)u(t)]dW(t), \\ x(0) = x_0 \in \mathbf{R}^n. \end{cases}$$

Here $W(t)$ is a standard one-dimensional Brownian motion, and the control $u(\cdot)$ takes value in some Euclidean space.

In optimal LQ control theory, the Riccati equation approach has been used systematically to provide an optimal feedback control (see [14, 20, 4] for the deterministic case, and [23, 6, 11] for the stochastic case). In the literature it is typically assumed that the cost function has a positive definite weighting matrix, R, for the control term, and a positive semidefinite weighting matrix, Q, for the state term. In this case, the solvability of the Riccati equation is both necessary and sufficient for the solvability of the underlying LQ problem.

However, it was found in [7] for the first time that a stochastic LQ problem with *indefinite* Q and R may still be well-posed. This phenomenon has to do with the

*Received by the editors April 19, 2000; accepted for publication June 18, 2001; published electronically December 7, 2001.

http://www.siam.org/journals/sicon/40-4/37108.html

†Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, Hong Kong (aitm@se.cuhk.edu.hk, xyzhou@se.cuhk.edu.hk). The research of the third author was supported by the RGC earmarked grants CUHK 4125/97E and CUHK 4054/98E.

‡Department of Information Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong (jmoore@ie.cuhk.edu.hk).

deep nature of the uncertainty involved; see [7] for a detailed discussion and many examples. Follow-up research on indefinite stochastic LQ control in a finite time horizon has been carried out in [8, 16, 9] to incorporate more complicated features such as random coefficients and integral constraints. The infinite-time-horizon case, in which the stability becomes a crucial issue, was treated in [2, 24] via techniques in linear matrix inequality and semidefinite programming [21]. On the other hand, applications of indefinite LQ control to portfolio selection problems and a contingent claim problem can be found in [25, 17] and [15], respectively. We would also like to mention a recent paper [12] in which the stochastic H^∞ problem is dealt with via a Riccati equation that has a structure similar to the one in [7].

In the first paper [7] on indefinite stochastic LQ control, it is shown that if the following type of Riccati equation, called the stochastic Riccati equation (t is suppressed),

(1.2)
$$\begin{cases} \dot{P} + PA + A'P + C'PC - (PB + C'PD)(R + D'PD)^{-1}(B'P + D'PC) + Q = 0, \\ P(T) = H, \\ R + D'PD > 0, \quad \text{a.e. } t \in [0, T], \end{cases}$$

has a solution $P(\cdot)$, then the original (indefinite) LQ problem is well-posed and an optimal feedback control can be constructed explicitly via $P(\cdot)$. (Note that the third positive definiteness constraint in (1.2) is *part* of that equation and must be satisfied by any solution.) In other words, the solvability of the stochastic Riccati equation (1.2) is *sufficient*, but not *necessary* in general, for the well-posedness as well as the solvability of the LQ problem. A natural question then is what can we say about the indefinite LQ problem if (1.2) does *not* have a solution at all? Note that the positive definiteness constraint of $R + D'PD$ in (1.2) is very restrictive, which likely leads to the nonexistence of its solutions. As a consequence, it may happen that the original indefinite LQ problem is well-posed and there exist optimal controls, while (1.2) still has no solution, in which case (1.2) becomes useless. This is quite different from the deterministic counterpart (as mentioned earlier), which in turn suggests that (1.2) may not be *the* right Riccati equation for indefinite LQ control.

Let us look at an example to illustrate the above discussion.

Example 1.1. Consider LQ problem (1.1) with $T = 1$, $A(t) = B(t) = D(t) = 1$, $C(t) = -1$, $Q(t) = -1$, $R(t) = -\frac{2e^{3(1-t)}+1}{3}$, and $H = 1$. Note that Q and R are both *negative* here. Equation (1.2) in this case specializes to

$$\begin{cases} \dot{P}(t) + 3P(t) - 1 = 0, \\ P(1) = 1, \\ R(t) + P(t) > 0, \quad \text{a.e. } t \in [0, T]. \end{cases} \tag{1.3}$$

The only solution that satisfies the first two constraints of the above equation is $P(t) = \frac{2e^{3(1-t)}+1}{3}$. Hence $R(t) + P(t) \equiv 0$, violating the third constraint. This shows that (1.3) has no solution and the result in [7] fails to apply to this case. However, the original LQ problem is well-posed. To see this, let $P(t) = \frac{2e^{3(1-t)}+1}{3}$, and apply Itô's formula to $P(t)x(t)^2$. We then obtain

$$d[P(t)x(t)^2] = [x(t)^2 + P(t)u(t)^2]dt + \{\cdots\}dW(t).$$

Integrating from 0 to 1, and taking expectation, we have $J \equiv P(0)x_0^2$. This implies that the cost function takes a *constant* value $P(0)x_0^2$ regardless of the control being applied. In particular, the LQ problem is well-posed and does have optimal controls.

The above example suggests that the stochastic Riccati equation (1.3) introduced in [7] may not be able to handle certain indefinite stochastic LQ problems. Finding a more appropriate Riccati-type equation, in the sense that its solvability should be *equivalent* to that of the underlying LQ problem, remains an outstanding open problem. The objective of this paper is to tackle this open problem, thereby enabling us to deal with general indefinite stochastic LQ problems, including pathological situations such as the one in Example 1.1. The key to achieving this goal is the introduction of a new type of differential Riccati equation—called a *generalized Riccati equation*—where the positive definiteness constraint of $R + D'PD$ is relaxed. This equation involves a matrix pseudoinverse and an additional algebraic constraint due to the possible singularity of the term $R+D'PD$. This new Riccati equation turns out to be the right one for studying indefinite LQ problems, as the solvability of this equation is not only sufficient, but also *necessary*, for the well-posedness of the LQ problem as well as the attainability of its optimal controls. Moreover, we are able to derive *all* optimal controls via the solution of the generalized Riccati equation.

It is worth mentioning that even for deterministic singular LQ problems (see [22, 13, 18, 10], among others), which are a special case of the problem treated in this paper, our formulation and results are still new; for details see section 3.

The remainder of this paper is organized as follows. Section 2 formulates the indefinite stochastic LQ problem and gives some preliminaries. The generalized Riccati equation (GRE) is also introduced. Section 3 shows that the solvability of the GRE is sufficient for the well-posedness of the LQ problem and the existence of an optimal control. Moreover, all the optimal controls are identified via the solution of the GRE. Sections 4 and 5 prove that the solvability of the GRE is also necessary for the existence of optimal linear feedback controls and optimal open-loop controls, respectively. An example is presented in section 6 to illustrate the results obtained. Finally, section 7 gives some concluding remarks.

2. Problem formulation and preliminaries.

2.1. Notation. We make use of the following notation in this paper:

$\mathbf{N}$	:	the set of positive integers.
$\mathbf{R}$	:	the set of real numbers.
$\mathbf{R}^n$	:	n-dimensional Euclidean space.
M'	:	the transpose of a matrix M.
$M^\dagger$	:	the Moore–Penrose pseudoinverse of a matrix M.
$\mathbf{Tr}(M)$	:	the sum of diagonal elements of a square matrix M.
$\|x\|$	:	$= \sqrt{\sum x_i^2}$ for a vector $x = (x_1, \ldots, x_n)'$.
$\mathbf{R}^{n\times m}$	:	the space of all $n \times m$ matrices.
$\mathcal{S}^n$	:	the space of all $n \times n$ symmetric matrices.
$\mathcal{S}_+^n$	:	the subspace of all positive semidefinite matrices of $\mathcal{S}^n$.
$\hat{\mathcal{S}}_+^n$	:	the subspace of all positive definite matrices of $\mathcal{S}^n$.

Given a filtered probability space $(\Omega, \mathcal{F}, \mathcal{P}; \mathcal{F}_t)$, where $t \in [0, T]$, and a Hilbert space X with the norm $\|\cdot\|_X$, define the Hilbert space

$$L^2_{\mathcal{F}}(0, T; X) = \Big\{\phi(\cdot)| \quad \phi(\cdot) \text{ is an } \mathcal{F}_t\text{-adapted, } X\text{-valued measurable process on } [0, T], \\ \text{and } E\int_0^T \| \phi(t, \omega) \|_X^2 \, dt < +\infty\Big\},$$

with the norm

$$\| \phi(\cdot) \|_{\mathcal{F},2} = \left(E \int_0^T \| \phi(t,\omega) \|_X^2 \, dt \right)^{\frac{1}{2}}.$$

2.2. Problem formulation. Let $(\Omega, \mathcal{F}, \mathcal{P}; \mathcal{F}_t)$ be a given filtered probability space with a standard one-dimensional Brownian motion $W(t)$ on $[0,T]$ (with $W(0) = 0$). Consider the following linear Itô stochastic differential equation:

$$\begin{cases} dx(t) = [A(t)x(t) + B(t)u(t)]dt + [C(t)x(t) + D(t)u(t)]dW(t), \\ x(s) = y, \end{cases} \tag{2.1}$$

where $(s,y) \in [0,T) \times \mathbf{R}^n$ are the initial time and initial state, respectively, and $u(\cdot)$, the admissible control, is an element in $U_{ad} \equiv L^2_{\mathcal{F}}(0,T;\mathbf{R}^{n_u})$. In order to simplify exposition we assume that the Brownian motion is one-dimensional. There is no essential difficulty with the multidimensional case.

For each (s,y) and $u(\cdot) \in U_{ad}$, the associated cost is

$$J(s,y;u(\cdot)) = E\left\{ \int_s^T [x(t)'Q(t)x(t) + u(t)'R(t)u(t)]dt + x(T)'Hx(T) \right\}. \tag{2.2}$$

The solution $x(\cdot)$ of (2.1) is called the response of the control $u(\cdot) \in U_{ad}$, and $(x(\cdot), u(\cdot))$ is called an *admissible pair*. The objective of the optimal control problem is to minimize the cost function $J(s,y;u(\cdot))$, for a given $(s,y) \in [0,T) \times \mathbf{R}^n$, over all $u(\cdot) \in U_{ad}$. The value function is defined as

$$V(s,y) = \inf_{u(\cdot) \in U_{ad}} J(s,y;u(\cdot)). \tag{2.3}$$

Definition 2.1. *The optimization problem* (2.1)–(2.3) *is called* well-posed *if*

$$V(s,y) > -\infty \quad \forall (s,y) \in [0,T) \times \mathbf{R}^n.$$

An admissible pair $(x^*(\cdot), u^*(\cdot))$ *is called* optimal *(with respect to the initial condition* (s,y)*) if* $u^*(\cdot)$ *achieves the infimum of* $J(s,y;u(\cdot))$.

The following basic assumption will be in force throughout this paper.

Assumption **(A)**. The data appearing in the LQ problem (2.1)–(2.3) satisfy

$$\begin{aligned} &A, C \in L^\infty(0,T;\mathbf{R}^{n\times n}), \\ &B, D \in L^\infty(0,T;\mathbf{R}^{n\times n_u}), \\ &Q \in L^\infty(0,T;\mathcal{S}^n), \\ &R \in L^\infty(0,T;\mathcal{S}^{n_u}), \\ &H \in \mathcal{S}^n. \end{aligned}$$

We emphasize again that we are dealing with an *indefinite* LQ problem, namely that Q, R, and H are all possibly indefinite.

2.3. Generalized (differential) Riccati equation. We start by recalling properties of a pseudo matrix inverse [19].

Proposition 2.2. *Let a matrix* $M \in \mathbf{R}^{m\times n}$ *be given. Then there exists a unique matrix* $M^\dagger \in \mathbf{R}^{n\times m}$ *such that*

$$\begin{cases} MM^\dagger M = M, \quad M^\dagger MM^\dagger = M^\dagger, \\ (MM^\dagger)' = MM^\dagger, \quad (M^\dagger M)' = M^\dagger M. \end{cases} \tag{2.4}$$

The matrix $M^\dagger$ above is called the *Moore–Penrose pseudoinverse* of M.

Now, we introduce a new type of Riccati equation associated with the LQ problem (2.1)–(2.3).

DEFINITION 2.3. *The constrained differential equation (with the time argument t suppressed)*

(2.5)
$$\begin{cases} \dot{P} + PA + A'P + C'PC - (PB + C'PD)(R + D'PD)^\dagger(B'P + D'PC) + Q = 0, \\ P(T) = H, \\ (R + D'PD)(R + D'PD)^\dagger(B'P + D'PC) - (B'P + D'PC) = 0, \\ R + D'PD \geq 0, \quad a.e.\ t \in [0, T], \end{cases}$$

is called a generalized (differential) Riccati equation *(GRE).*

If the term $(R + D'PD)$ is further required to be nonsingular, then the GRE reduces to the stochastic Riccati equation (1.2) that was introduced in [7].

Another interesting special case is that in which $(R + D'PD) \equiv 0$; the GRE reduces to the following linear differential matrix system:

$$\begin{cases} \dot{P} + PA + A'P + C'PC + Q = 0, \\ P(T) = H, \\ B'P + D'PC = 0, \\ R + D'PD = 0, \quad \text{a.e. } t \in [0, T]. \end{cases} \tag{2.6}$$

2.4. Some useful lemmas.

LEMMA 2.4. *Let $M(\cdot)$ be a given continuously differentiable (in t) matrix function taking values in $\mathcal{S}^n$. Then for any admissible pair $(x(\cdot), u(\cdot))$ of the system* (2.1)*, we have*

$$\begin{aligned} &E[x(t)'Mx(T)] - y'M(s)y - E\int_s^T [x'(\dot{M} + A'M + MA + C'MC)x](t)dt \\ &-E\int_s^T [2u'(B'M + D'MC)x + u'D'MDu](t)dt = 0. \end{aligned} \tag{2.7}$$

Proof. Using Itô's formula, we have (t is suppressed)

$$\begin{aligned} &d(x'Mx) - [(Ax + Bu)'Mx + x'\dot{M}x + x'M(Ax + Bu) - (Cx + Du)'M(Cx + Du)]dt \\ &\quad - [x'M(Cx + Du) + (Cx + Du)'Mx]dW(t) = 0. \end{aligned}$$

Taking expectations and integrations we obtain (2.7). □

LEMMA 2.5. *Let a symmetric matrix S be given. Then*

(i) $S^\dagger = S^{\dagger'}$.

(ii) $S \geq 0$ *if and only if* $S^\dagger \geq 0$.

(iii) $SS^\dagger = S^\dagger S$.

Proof. Since S is symmetric, it has a singular value decomposition of the following form:

$$S = V \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} V',$$

where Σ is a nonsingular diagonal matrix and V a matrix such that $VV' = V'V = I$. Now, $S^\dagger$ is given by

$$S^\dagger = V \begin{bmatrix} \Sigma^{-1} & 0 \\ 0 & 0 \end{bmatrix} V'.$$

Using the above expression of $S^\dagger$, it is easy to show that items (i)–(iii) hold. □

LEMMA 2.6 (Extended Schur's lemma [3]). *Let matrices $M = M', N$, and $R = R'$ be given with appropriate sizes. Then the following conditions are equivalent:*

(i) $M - NR^\dagger N' \geq 0$, $R \geq 0$, *and* $N(I - RR^\dagger) = 0$.

(ii) $\begin{bmatrix} M & N \\ N' & R \end{bmatrix} \geq 0$.

(iii) $\begin{bmatrix} R & N' \\ N & M \end{bmatrix} \geq 0$.

The following lemma plays a key technical role in this paper.

LEMMA 2.7. *Let matrices L, M, and N be given with appropriate sizes. Then the matrix equation*

$$LXM = N \tag{2.8}$$

has a solution X if and only if

$$LL^\dagger NM^\dagger M = N. \tag{2.9}$$

Moreover, any solution to (2.8) *is represented by*

$$X = L^\dagger NM^\dagger + S - L^\dagger LSMM^\dagger, \tag{2.10}$$

where S is a matrix with an appropriate size.

Proof. If X satisfies the equation $LXM = N$, then we have

$$N = LXM = LL^\dagger LXMM^\dagger M = LL^\dagger NM^\dagger M.$$

Conversely, if (2.9) is satisfied, then $L^\dagger NM^\dagger$ is a solution of $LXM = N$. This proves the first part of the lemma. Now, let Y be any matrix with appropriate size and define $\tilde{X} = Y - L^\dagger LYMM^\dagger$. Then $\tilde{X}$ satisfies the homogeneous equation $L\tilde{X}M = 0$. Hence $L^\dagger NM^\dagger + \tilde{X}$ must satisfy (2.8). On the other hand, let X be a solution to (2.8). Then by (2.9), one has $LSM = 0$, where $S = X - L^\dagger NM^\dagger$. Hence

$$X = L^\dagger NM^\dagger + S - L^\dagger LSMM^\dagger.$$

This completes the proof. □

3. Sufficiency of the GRE. In this section, we will show that the solvability of the GRE (2.5) is sufficient for the well-posedness of the LQ problem and the existence of an optimal linear state feedback control. Moreover, any optimal control can be obtained via the solution to the GRE.

THEOREM 3.1. *If the GRE* (2.5) *admits a solution $P(\cdot)$, then the stochastic LQ problem* (2.1)–(2.3) *is well-posed. Moreover, the set of all the optimal controls with respect to the initial $(s, y) \in [0, T) \times \mathbf{R}^n$ is determined by the following (parameterized by (Y, z)):*

$$\begin{aligned} u_{(Y,z)}(t) = -\Big\{ & [R(t) + D(t)'P(t)D(t)]^\dagger [B(t)'P(t) + D(t)'P(t)C(t)] + Y(t) \\ & - [R(t) + D(t)'P(t)D(t)]^\dagger [R(t) + D(t)'P(t)D(t)]Y(t) \Big\} x(t) \\ & + z(t) - [R(t) + D(t)'P(t)D(t)]^\dagger [R(t) + D(t)'P(t)D(t)]z(t), \end{aligned} \tag{3.1}$$

where $Y(\cdot) \in L^2_{\mathcal{F}}(s, T; \mathbf{R}^{n_u \times n})$ and $z(\cdot) \in L^2_{\mathcal{F}}(s, T; \mathbf{R}^{n_u})$. Furthermore, the value function is uniquely determined by $P(\cdot)$:

$$V(s, y) \equiv \inf_{u(\cdot) \in U_{ad}} J(s, y; u(\cdot)) = y'P(s)y. \tag{3.2}$$

Proof. Let $P(\cdot)$ be a solution of GRE (2.5). Applying Lemma 2.4, we can express the cost function as follows:

$$J(s,y;u(\cdot)) = y'P(s)y + E\int_s^T \Big[x'(\dot P + PA + A'P + C'PC + Q)x + 2u'(B'P + D'PC)x + u'(D'PD + R)u\Big](t)dt. \tag{3.3}$$

Now, let $Y(\cdot) \in L^2_{\mathcal{F}}(s,T;\mathbf{R}^{n_u\times n})$ and $z(\cdot) \in L^2_{\mathcal{F}}(s,T;\mathbf{R}^{n_u})$ be given. Set

$$\begin{aligned} L_1(t) &= Y(t) - [R(t) + D'(t)P(t)D(t)]^\dagger [R(t) + D'(t)P(t)D(t)]Y(t),\\ L_2(t) &= z(t) - [R(t) + D'(t)P(t)D(t)]^\dagger [R(t) + D'(t)P(t)D(t)]z(t). \end{aligned}$$

Applying Proposition 2.2 and Lemma 2.5(iii), we have

$$[R(t) + D'(t)P(t)D(t)]L_i(t) = [R(t) + D'(t)P(t)D(t)]^\dagger L_i(t) = 0, \qquad i = 1,2, \tag{3.4}$$

and

$$[P(t)B(t) + C'(t)P(t)D(t)]L_i(t) = 0, \qquad i = 1,2. \tag{3.5}$$

Then identity (3.3) can be expressed as

$$\begin{aligned} &J(s,y;u(\cdot))\\ &\quad = y'P(s)y + E\textstyle\int_s^T \Big\{x'[\dot P + PA + A'P + C'PC + Q\\ &\qquad - (PB + C'PD)(R + D'PD)^\dagger (B'P + D'PC)]x\\ &\qquad + \big[u + \big((R + D'PD)^\dagger (B'P + D'PC) + L_1\big)x + L_2\big]'\\ &\qquad \times (R + D'PD)\big[u + \big((R + D'PD)^\dagger (B'P + D'PC) + L_1\big)x + L_2\big]\Big\}(t)dt\\ &\quad = y'P(s)y + E\textstyle\int_s^T \Big\{\big[u + \big((R + D'PD)^\dagger (B'P + D'PC) + L_1\big)x + L_2\big]'\\ &\qquad \times (R + D'PD)\big[u + \big((R + D'PD)^\dagger (B'P + D'PC) + L_1\big)x + L_2\big]\Big\}(t)dt. \end{aligned} \tag{3.6}$$

Hence, $J(s,y;u(\cdot))$ is minimized by the control given by (3.1) with the optimal value being $y'P(s)y$.

What remains to show is that *any* optimal control can be represented by (3.1) for some $Y(\cdot)$ and $z(\cdot)$. To this end, let $u(\cdot)$ be an optimal control. Then by (3.6) the integrand in the right-hand side of (3.6) must be zero almost everywhere in t. This implies (t is suppressed)

$$(R + D'PD)^{1/2}\big[u + \big((R + D'PD)^\dagger (B'P + D'PC) + L_1\big)x + L_2\big] = 0,$$

which leads to

$$(R + D'PD)[u + \big((R + D'PD)^\dagger (B'P + D'PC) + L_1\big)x + L_2] = 0,$$

or, equivalently (noting (3.4)),

$$[R(t) + D(t)'P(t)D(t)]u(t) + [B(t)'P(t) + D(t)'P(t)C(t)]x(t) = 0, \quad \text{a.e. } t \in [s,T]. \tag{3.7}$$

To solve the above equation in $u(t)$, we apply Lemma 2.7 with

$$L = R(t) + D(t)'P(t)D(t), \quad M = I, \quad N = -[B(t)'P(t) + D(t)'P(t)C(t)]x(t).$$

Notice that condition (2.9) in the present case is implied by the third constraint in GRE (2.5); hence the general solution (2.10) with $z(t) = S$ and $Y(t) = 0$ yields that $u(t)$ can be represented by (3.1). □

COROLLARY 3.2. *The optimal controls are obtained in the following special cases:*

(i) *If* $R(t) + D(t)'P(t)D(t) \equiv 0$, a.e. $t \in [s,T]$, *then any admissible control is optimal.*

(ii) *If* $R(t) + D(t)'P(t)D(t) > 0$, a.e. $t \in [s,T]$, *then there is a unique optimal control that is given by the following linear feedback law:*

$$u(t) = -[R(t) + D(t)'P(t)D(t)]^{-1}[B(t)'P(t) + D(t)'P(t)C(t)]x(t). \tag{3.8}$$

Proof. The proofs here are straightforward from Theorem 3.1. □

As an immediate consequence of Theorem 3.1, we have the uniqueness of the solution to GRE (2.5).

COROLLARY 3.3. *If there is a solution to the GRE* (2.5), *then it must be the only solution to* (2.5).

Proof. Let $P_1(\cdot)$ and $P_2(\cdot)$ be two solutions of GRE (2.5). Then Theorem 3.1 implies that

$$y'P_1(s)y = y'P_2(s)y \ \ \forall y \in \mathbf{R}^n \ \ \forall s \in [0,T].$$

Hence $P_1(t) \equiv P_2(t)$. □

It is interesting to see the specialization of our results in the deterministic case (i.e., $C(t) = D(t) \equiv 0$). The control weight $R(t)$ is given as satisfying $R(t) \geq 0$, so it is a possibly singular case. The corresponding GRE is

$$\begin{cases} \dot{P}(t) + P(t)A(t) + A(t)'P(t) - P(t)B(t)R(t)^\dagger B(t)'P(t) + Q(t) = 0, \\ P(T) = H, \\ R(t)R(t)^\dagger B(t)'P(t) - B(t)'P(t) = 0 \ \ \forall t \in [0,T]. \end{cases} \tag{3.9}$$

According to Theorem 3.1, if the above equation admits a solution $P(\cdot)$, then there may be infinitely many optimal controls, and any optimal control has the following form:

$$u_{Y,z}(t) = [-R(t)^\dagger B(t)'P(t) + Y(t) - R(t)^\dagger R(t)Y(t)]x(t) + z(t) - R(t)^\dagger R(t)z(t), \tag{3.10}$$

where $Y(\cdot) \in L^2(s,T;\mathbf{R}^{n_u \times n})$ and $z(\cdot) \in L^2(s,T;\mathbf{R}^{n_u})$.

4. Necessity of the GRE. In the previous section, we proved that the solvability of GRE (2.5) is *sufficient* for the well-posedness of the stochastic LQ problem (2.1)–(2.3), and that optimal feedback control laws can be constructed based on the solution to the Riccati equation. In particular, if (2.5) admits a solution, then there must be an optimal *linear feedback* control, obtained by taking $Y(t) \equiv 0$ and $z(t) \equiv 0$ in (3.1). In this section we shall show that the solvability of the GRE is also *necessary* for there to exist an optimal *linear feedback* control for the LQ problem.

4.1. A linear feedback control formulation. If a linear feedback control is optimal for the LQ problem (2.1)–(2.3), then it must be optimal also in the class of linear feedback controls of the following form:

$$u(t) = K(t)x(t), \qquad K(t) \in \mathbf{R}^{n_u \times n}. \tag{4.1}$$

The corresponding closed-loop system with the initial $(s, y) = (0, x_0)$ is

$$\begin{cases} dx(t) &= [A(t) + B(t)K(t)]x(t)dt + [C(t) + D(t)K(t)]x(t)dW(t), \\ x(0) &= x_0 \in \mathbf{R}^n. \end{cases} \tag{4.2}$$

Now, if $x(\cdot)$ satisfies (4.2), then by Itô's formula the matrix $X(t) \equiv E[x(t)x(t)']$ satisfies the differential matrix equation

$$\begin{cases} \dot{X}(t) = [A(t) + B(t)K(t)]X(t) + X(t)[A(t) + B(t)K(t)]' \\ \qquad\qquad + [C(t) + D(t)K(t)]X(t)[C(t) + D(t)K(t)]', \\ X(0) = X_0 \equiv E[x_0 x_0'] \in \mathcal{S}_n^+, \end{cases} \tag{4.3}$$

with the associated cost function J expressed equivalently as

$$J(K(\cdot)) \equiv \int_0^T \mathbf{Tr}[(Q(t) + K'(t)R(t)K(t))X(t)]dt + \mathbf{Tr}(HX(T)). \tag{4.4}$$

To summarize, if we consider only the class of linear feedback controls for the original LQ problem with the initial $x(0) = x_0$, then the problem reduces to the following deterministic optimal control problem:

$$\begin{cases} \text{Minimize}_{K(\cdot)} & \displaystyle\int_0^T \mathbf{Tr}[(Q + K'RK)X]dt + \mathbf{Tr}(HX(T)), \\ \text{subject to} & (4.3). \end{cases} \tag{4.5}$$

4.2. Matrix minimum principle. For the reader's convenience, let us state the matrix minimum principle (see [5]) here. We start by defining the gradient matrix. Let $f(\cdot)$ be a function from $\mathbf{R}^{p \times q}$ to $\mathbf{R}$. Then the gradient matrix of f is a $p \times q$ matrix, denoted by $\frac{\partial f(X)}{\partial X}$, with the ijth component $(\frac{\partial f(X)}{\partial X})_{ij} = \frac{\partial f(X)}{\partial x_{ij}}$.

Consider an $n \times n$ matrix differential system

$$\begin{cases} \dot{X} = F(X(t), U(t), t), \\ X(t_0) = X_0, \end{cases} \tag{4.6}$$

where the control $U(\cdot)$ is a measurable mapping from $[t_0, T]$ to a prescribed set $\Omega \subseteq \mathbf{R}^{n_u \times n}$. With T fixed, consider the cost function

$$J(X_0, U(\cdot)) = \int_{t_0}^T L(X(t), U(t), t)dt + G(X(T)), \tag{4.7}$$

where G and L are scalar-valued functions, satisfying the usual smooth conditions. Let $P(t) \in \mathbf{R}^{n \times n}$ be the costate (adjoint) matrix. Then, the Hamiltonian function is defined as

$$H(X(t), P(t), U(t), t) = L(X(t), U(t), t) + \mathbf{Tr}(F(X(t), U(t), t)P(t)'). \tag{4.8}$$

Now, we present the matrix minimum principle in the form stated in [5].

PROPOSITION 4.1. *If $(X^*(\cdot), U^*(\cdot))$ is optimal for (4.6)–(4.7), then there exists a costate matrix $P^*(t)$ satisfying the costate (adjoint) equation*

(4.9)
$$\begin{cases} \dot{P}^*(t) = -\dfrac{\partial}{\partial X^*(t)} L(X^*(t), U^*(t), t) - \dfrac{\partial}{\partial X^*(t)} \mathbf{Tr}(F(X^*(t), U^*(t), t) P^*(t)'), \\ P^*(T) = \dfrac{\partial}{\partial X^*(T)} G(X^*(T)) \end{cases}$$

such that

$$H(X^*(t), P^*(t), U^*(t), t) \le H(X^*(t), P^*(t), U, t) \ \ \forall U \in \Omega, \ \text{ a.e. } t \in [t_0, T]. \tag{4.10}$$

Note that if $U(\cdot)$ is unconstrained, then (4.10) is equivalent to $\frac{\partial H}{\partial U^*(t)} = 0$.

4.3. Necessity of the GRE. We are going to use the matrix minimum principle to show that if there exists an optimal linear feedback control for the original LQ problem (2.1)–(2.3), then GRE (2.5) must have a solution. Furthermore, we will show that any optimal linear feedback control law has the form (3.1) with $z(t) \equiv 0$.

First we give the following necessary condition.

THEOREM 4.2. *Assume that the LQ problem (2.1)–(2.3) is well-posed. For any $s \in [0, T)$, if there exists $P(\cdot)$ such that*

(4.11)
$$\begin{cases} \dot{P} + PA + A'P + C'PC - (PB + C'PD)(R + D'PD)^\dagger (B'P + D'PC) + Q = 0, \\ P(T) = H, \\ (R + D'PD)(R + D'PD)^\dagger (B'P + D'PC) - B'P - D'PC = 0, \ \text{ a.e. } t \in [s, T], \end{cases}$$

then P must satisfy

$$R + D'PD \ge 0, \ \text{ a.e. } t \in [s, T].$$

Proof. Let $\lambda(t)$ be any fixed eigenvalue of the matrix $R(t) + D(t)'P(t)D(t)$, $t \in [s, T]$. We will show that $\mathbf{mes}(\{t \in [s, T] | \lambda(t) < 0\}) = 0$, where $\mathbf{mes}$ denotes the Lebesgue measure. Let $v_\lambda(t)$ be a unit eigenvector (i.e., $v_\lambda(t)' v_\lambda(t) = 1$) associated with the eigenvalue $\lambda(t)$. Define $I_n(\cdot)$ as the indicator function of the set $\{t \in [s, T] | \lambda(t) < -\frac{1}{n}\}$, $n = 1, 2, \ldots$. Fix a scalar $\delta \in \mathbf{R}$ and consider the state trajectory $x(\cdot)$ of system (2.1) under the feedback control

(4.12)
$$u(t) = \begin{cases} 0 & \text{if } \lambda(t) = 0, \\ \dfrac{\delta I_n(t)}{|\lambda(t)|^{1/2}} v_\lambda(t) & \\ -[R(t) + D(t)'P(t)D(t)]^\dagger [B(t)'P(t) + D(t)'P(t)C(t)]x(t) & \text{if } \lambda(t) \ne 0. \end{cases}$$

Clearly, $u(\cdot) \in L^2_{\mathcal{F}}(s, T; \mathbf{R}^{n_u})$. Now,

$$\begin{aligned} J(s, y; u(\cdot)) = y'P(s)y + E\int_s^T \big[u + (R + D'PD)^\dagger (B'P + D'PC)x\big]'(R + D'PD) \\ \times \big[u + (R + D'PD)^\dagger (B'P + D'PC)x\big](t)dt. \end{aligned}$$

It follows from $\lambda(t) \neq 0$ that $|\lambda(t)|^{-1} I_n(t)(R(t) + D(t)'P(t)D(t))v_\lambda(t) = -I_n(t)v_\lambda(t)$. Hence

$$
\begin{aligned}
J(s, y; u(\cdot)) &= y'P(s)y - \delta^2 \int_s^T I_n(t)dt \\
&= y'P(s)y - \delta^2 \mathbf{mes}\left(\left\{t \in [s,T] | \lambda(t) < -\frac{1}{n}\right\}\right).
\end{aligned} \tag{4.13}
$$

If $\mathbf{mes}(\{t \in [s,T] | \lambda(t) < -\frac{1}{n}\}) > 0$, then by letting $\delta \to \infty$ in (4.13) we obtain $J(s, y; u(\cdot)) \to -\infty$, which contradicts the well-posedness of the LQ problem. Hence $\mathbf{mes}(\{t \in [s,T] | \lambda(t) < -\frac{1}{n}\}) = 0$. Since

$$
\{t \in [s,T] | \lambda(t) < 0\} = \bigcup_{n \in \mathbf{N}} \left\{t \in [s,T] | \lambda(t) < -\frac{1}{n}\right\},
$$

we conclude that $\mathbf{mes}(\{t \in [s,T] | \lambda(t) < 0\}) = 0$, completing the proof. □

THEOREM 4.3. *If a given linear feedback control $u(t) = K(t)x(t)$ is optimal for the LQ problem (2.1)–(2.3) with respect to the initial $(s, y) = (0, x_0)$, then GRE (2.5) must have a solution $P(\cdot)$. Moreover, the optimal feedback control $u(t) = K(t)x(t)$ can be represented via (3.1) with $z(t) \equiv 0$. In particular, the feedback law $u(t) = K(t)x(t)$ must be optimal with respect to any initial $(s, y) \in [0, T) \times \mathbf{R}^n$.*

Proof. Since the given feedback control $u(t) = K(t)x(t)$ is optimal over the set of all admissible controls, it must in particular be optimal over the class of all linear feedback controls. Therefore, as shown earlier, $K(\cdot)$ must solve the following deterministic optimal control problem:

$$
\begin{cases}
\min\limits_{K(\cdot)} & \displaystyle\int_0^T \mathbf{Tr}[(Q + K'RK)X](t)dt + \mathbf{Tr}(HX(T)), \\
\text{subject to} & \begin{cases} \dot{X}(t) = (A + BK)X + X(A + BK)' + (C + DK)X(C + DK)', \\ X(0) = X_0, \ \ X_0 \in \mathcal{S}_n^+. \end{cases}
\end{cases} \tag{4.14}
$$

By the minimum principle, Proposition 4.1, we conclude that the Hamiltonian

$$
\mathbf{Tr}\Big((Q + K'RK)X + [(A + BK)X + X(A + BK)' + (C + DK)X(C + DK)']P'\Big)
$$

is pointwise (in t) minimized at $K(t)$ over the space of $\mathbf{R}^{n_u \times n}$. This, together with the costate equation (4.9), leads to the following:

$$
\begin{cases}
\dot{P} = -Q - K'RK - (C + DK)'P(C + DK) - P(A + BK) - (A + BK)'P, \\
P(T) = H, \\
0 = RKX' + RKX + B'PX' + B'P'X + D'PDKX' + D'PDKX \\
\qquad + D'PCX' + D'P'CX.
\end{cases} \tag{4.15}
$$

Note that in the above calculation we have used the following formulae:

$$
\frac{\partial}{\partial X}\mathbf{Tr}(AX) = A', \ \ \frac{\partial}{\partial X}\mathbf{Tr}(AX') = A, \ \ \frac{\partial}{\partial X}\mathbf{Tr}(AXBX') = A'XB' + AXB.
$$

Since X and P are symmetric, (4.15) is reduced to

(4.16)
$$\begin{cases} \dot{P} = -Q - K'RK - (C + DK)'P(C + DK) - P(A + BK) - (A + BK)'P, \\ P(T) = H, \\ 0 = (R + D'PD)K + B'P + D'PC. \end{cases}$$

Now, we apply Lemma 2.7 to the equation $(R + D'PD)K + B'P + D'PC = 0$. First of all, we know a priori that it does have a solution K. Thus condition (2.9) must hold, which in the present case specializes to

$$(R + D'PD)(R + D'PD)^{\dagger}(B'P + D'PC) = B'P + D'PC.$$

Moreover, by (2.10), K has the following form:

$$K = -(R + D'PD)^{\dagger}(B'P + D'PC) + Y - (R + D'PD)^{\dagger}(R + D'PD)Y. \tag{4.17}$$

Substituting K into the first equation of (4.16), we can see by a simple calculation that P satisfies

$$\dot{P} + PA + A'P + C'PC - (PB + C'PD)(R + D'PD)^{\dagger}(B'P + D'PC) + Q = 0.$$

Noting that Theorem 4.2 implies that $R + D'PD \geq 0$, we conclude that $P(\cdot)$ solves (2.5). The representation of K is given by (4.17). Finally, the last assertion of the theorem follows from Theorem 3.1. □

5. Open-loop optimal controls. In the previous analysis we have shown that the solvability of the GRE is equivalent to the condition that the LQ problem is solvable by linear feedback controls. In this section we further prove that the solvability of the GRE is also equivalent to the case in which the LQ problem is solvable by continuous *open-loop* controls.

First we need the following lemma.

LEMMA 5.1. *Assume that the LQ problem* (2.1)–(2.3) *is well-posed. Then there exists a symmetric matrix function* $P(\cdot)$ *such that*

$$V(s, y) = y'P(s)y \quad \forall (s, y) \in [0, T] \times \mathbf{R}^n. \tag{5.1}$$

Moreover, assume that $Q(t)$ *and* $R(t)$ *are continuous in* t*, and for any initial* $(s, y) \in [0, T] \times \mathbf{R}^n$ *the LQ problem* (2.1)–(2.3) *has an optimal open-loop control that is continuous in* t*; then the matrix function* $P(\cdot)$ *satisfying* (5.1) *is differentiable on* $[0, T]$.

Proof. First, (5.1) can be shown by a simple adaptation of the well-known result in the deterministic case (see, e.g., [10, 4]). Moreover, since the value function $V(s, y)$ is continuous in s, so is $P(\cdot)$. Next, fix (s, y) and let $(u_*(\cdot), x_*(\cdot))$ be an optimal solution of (2.1)–(2.3) with respect to the initial condition $x(s) = y$ with $u_*(\cdot)$ continuous. Then the dynamic programming optimality principle yields

(5.2)
$$V(s, y) = E\left\{ \int_s^{s+h} [x_*(t)'Q(t)x_*(t) + u_*(t)'R(t)u_*(t)]dt + V(s + h, x_*(s + h)) \right\} \forall h \geq 0.$$

Making use of (5.1)–(5.2), we have

$$\begin{aligned}\tfrac{1}{h}[y'P(s+h)y - y'P(s)y] &= \tfrac{1}{h}E[y'P(s+h)y - x_*(s+h)'P(s+h)y]\\ &\quad + \tfrac{1}{h}E[x_*(s+h)'P(s+h)y - x_*(s+h)'P(s+h)x_*(s+h)]\\ &\quad - \tfrac{1}{h}E\textstyle\int_s^{s+h}[x_*(t)'Q(t)x_*(t) + u_*(t)'R(t)u_*(t)]dt.\end{aligned}$$

Noting that $P(\cdot)$ and $x_*(\cdot)$ are continuous and the integrand above is continuous in t by the assumptions, we can show by a standard argument that the limit of each of the three terms on the right-hand side of the above equation exists as h goes to zero. Therefore $\lim_{h\to 0}\frac{1}{h}[y'P(s+h)y - y'P(s)y]$ exists. Since y is arbitrary, $P(s)$ is differentiable at $s \in [0,T]$. □

The assumption that the optimal control is continuous in t is a rather technical one. From the above proof we can see that only the continuity of the control at the initial time s is actually needed. On the other hand, if we assume that $B(t), C(t), D(t)$, and $R(t)$ are continuous, then by (3.1) the existence of a continuous optimal open-loop control is really *necessary* for the solvability of GRE (2.5).

Consider the following convex set of differentiable symmetric matrix functions on $[0,T]$:

$$\mathcal{P} \triangleq \left\{ P(\cdot)\Big| \quad \left[\begin{array}{c|c} \dot P + A'P + PA + C'PC + Q & PB + C'PD \\ \hline B'P + D'PC & R + D'PD \end{array}\right] \geq 0, \right. \\ \left. \text{a.e. } t \in [0,T],\ P(T) \leq H \right\}. \tag{5.3}$$

The following result provides a sufficient condition for the well-posedness of the LQ problem.

THEOREM 5.2. *The LQ problem* (2.1)–(2.3) *is well-posed if the set* $\mathcal{P}$ *is nonempty.*

Proof. Let $P(\cdot) \in \mathcal{P}$. Applying Lemma 2.4, we have, for any admissible (open-loop) control $u(\cdot)$ and any initial $(s,y) \in [0,T) \times \mathbf{R}^n$,

$$\begin{aligned} J(s,y;u(\cdot)) &= y'P(s)y + E[x(T)(H - P(T))x(T)]\\ &\quad + E\int_s^T \begin{pmatrix} x \\ u \end{pmatrix}' \left[\begin{array}{c|c} \dot P + A'P + PA + C'PC + Q & PB + C'PD \\ \hline B'P + D'PC & R + D'PD \end{array}\right] \begin{pmatrix} x \\ u \end{pmatrix}(t)dt. \end{aligned} \tag{5.4}$$

Thus $J(s,y;u(\cdot)) \geq y'P(s)y$, implying $V(s,y) > -\infty\ \ \forall (s,y) \in [0,T) \times \mathbf{R}^n$. □

The following is the main result of this section.

THEOREM 5.3. *Assume that* $B(t), C(t), D(t), Q(t)$, *and* $R(t)$ *are continuous in* t. *Then the LQ problem* (2.1)–(2.3) *has an continuous optimal open-loop control for any initial* $(s,y) \in [0,T] \times \mathbf{R}^n$ *if and only if GRE* (2.5) *has a solution* $P(\cdot)$.

Proof. The "if" part follows from Theorem 3.1. Let us now show the "only if" part. First, since the LQ problem is well-posed, Lemma 5.1 yields that there exists a symmetric matrix function $P(\cdot)$ such that

$$V(s,y) = y'P(s)y\ \ \forall (s,y) \in [0,T) \times \mathbf{R}^n.$$

Moreover, by the assumption and Lemma 5.1, $P(\cdot)$ is differentiable. On the other hand, the dynamic programming principle yields

$$\begin{aligned} V(s,y) \leq\ & E\left\{ \int_s^{s+h} [x(t)'Q(t)x(t) + u(t)'R(t)u(t)]dt + V(s+h, x(s+h)) \right\} \\ & \forall h \geq 0\ \ \forall u(\cdot) \in U_{ad}. \end{aligned}$$

Applying Itô's formula to $V(t,x(t)) \equiv x(t)'P(t)x(t)$, using the above inequality, and employing Lemma 2.4, we obtain

$$E\int_s^{s+h} \begin{pmatrix} x \\ u \end{pmatrix}' \left[\begin{array}{c|c} \dot{P}+A'P+PA+C'PC+Q & PB+C'PD \\ \hline B'P+D'PC & R+D'PD \end{array}\right] \begin{pmatrix} x \\ u \end{pmatrix}(t)dt \geq 0.$$

Taking $u(t) \equiv \bar{u} \in \mathbf{R}^{n_u}$, and then dividing both sides by h and letting $h \to 0$, we obtain

$$\begin{pmatrix} y \\ \bar{u} \end{pmatrix}' \left[\begin{array}{c|c} \dot{P}+A'P+PA+C'PC+Q & PB+C'PD \\ \hline B'P+D'PC & R+D'PD \end{array}\right](s) \begin{pmatrix} y \\ \bar{u} \end{pmatrix} \geq 0, \quad \text{a.e. } s \in [0,T].$$

Since $y \in \mathbf{R}^n$ and $\bar{u} \in \mathbf{R}^{n_u}$ are arbitrary, we obtain

$$\left[\begin{array}{c|c} \dot{P}+A'P+PA+C'PC+Q & PB+C'PD \\ \hline B'P+D'PC & R+D'PD \end{array}\right] \geq 0, \quad \text{a.e. } t \in [0,T]. \tag{5.5}$$

Applying Lemma 2.6 to (5.5), we have

$$\begin{cases} \dot{P}+PA+A'P+C'PC-(PB+C'PD)(R+D'PD)^\dagger(B'P+D'PC)+Q \geq 0, \\ (R+D'PD)(R+D'PD)^\dagger(B'P+D'PC)-B'P-D'PC=0, \\ R+D'PD \geq 0, \quad \text{a.e. } t \in [0,T]. \end{cases} \tag{5.6}$$

Now, let $(x_*(\cdot),u_*(\cdot))$ be an optimal open-loop control for (2.1)–(2.3) with respect to the initial condition $x(s)=y$. Applying Lemma 2.4 to $P(\cdot)$, we have

$$\begin{aligned} V(s,y) = y'P(s)y \; &+ E\int_s^T [x_*'\big(\dot{P}+PA+A'P+C'PC+Q \\ &\qquad -(PB+C'PD)(R+D'PD)^\dagger(B'P+D'PC)\big)x_*](t)dt \\ &+ E\int_s^T [u_*+(R+D'PD)^\dagger(B'P+D'PC)x_*]' \\ &\qquad \times (R+D'PD)[u_*+(R+D'PD)^\dagger(B'P+D'PC)x_*](t)dt. \end{aligned} \tag{5.7}$$

By virtue of the relation $V(s,y)=y'P(s)y$ and (5.6)–(5.7), we obtain

$$\dot{P}+PA+A'P+C'PC-(PB+C'PD)(R+D'PD)^\dagger(B'P+D'PC)+Q=0.$$

This completes the proof. □

Theorem 5.1 says that the nonemptiness of the set $\mathcal{P}$ is sufficient for the well-posedness of the original LQ problem. The following result stipulates that the nonemptiness of the set $\mathcal{P}$ is also *necessary* for the attainability of the LQ problem.

THEOREM 5.4. *Under the same assumption of Theorem* 5.3, *the LQ problem* (2.1)–(2.3) *has a continuous optimal open-loop control for any initial* $(s,y) \in [0,T] \times \mathbf{R}^n$ *only if the set* $\mathcal{P}$ *is nonempty.*

Proof. This is seen from (5.5). □

6. An example. In this section we give an example in which the singularity of $R+D'PD$ does occur, but the LQ problem is well-posed and attainable. Moreover, the example shows that a stochastic LQ problem can be well-posed even when *both* Q and R are negative.

Consider the following one-dimensional LQ problem:

$$\begin{aligned} &\text{Minimize} \quad J = E\left\{\int_0^1 [qx(t)^2 + r(t)u(t)^2]dt + hx(1)^2\right\}, \\ &\text{subject to} \quad \begin{cases} dx(t) = [ax(t) + bu(t)]dt + [cx(t) + \delta u(t)]dW(t), \\ x(0) = x_0, \end{cases} \end{aligned} \tag{6.1}$$

where the coefficients are chosen such that $\delta \neq 0$, $b+\delta c = 0$, $q < 0$, and $2a+c^2+q > 0$. Take $r(t) = -\delta^2 p(t)$, where

$$p(t) = \frac{e^{(2a+c^2)(1-t)}(2ha + hc^2 + q) - q}{2a + c^2} \tag{6.2}$$

is the solution to the following equation:

$$\dot{p}(t) + (2a + c^2)p(t) + q = 0, \qquad p(1) = h. \tag{6.3}$$

Incidentally, Example 1.1 in section 1 is a special case of this example. It is easy to verify directly that (6.3) is exactly the GRE in the present case. (Note that the singularity arises because $r(t)+\delta^2 p(t) \equiv 0$.) Therefore, by Theorem 3.1 and Corollary 3.1(i), the LQ problem is well-posed, and any admissible control is optimal with an optimal cost $p(s)y^2$.

It is interesting to look at the sign of the solution to the Riccati equation (6.3). First assume that $h < 0$. In this case, since $2a + c^2 > 0$ and $q < 0$, we have from (6.2) that $p(t) \leq h < 0 \ \forall t \in [0, 1]$. Hence, the solution to the Riccati equation could be *negative*, which is quite contrary to the deterministic LQ case. On the other hand, if $h > 0$ is large enough so that $2ha + hc^2 + q > 0$, then $p(t) \geq h > 0$ and $r(t) = -\delta^2 p(t) < 0$. In this case, both q and $r(t)$ are *negative* but the LQ problem is well-posed. Again, this is different from the deterministic situation. The essential reason behind this phenomenon is that the positive terminal cost $hx(1)^2$ *outweighs* the negative running cost.

7. Conclusion. Standard LQ theory, which has proved so useful for control applications in the last decades, has been extended here to signal models with multiplicative noises in both state and control and with quadratic weights that are fundamentally different from those in the literature. Such models better approximate nonlinear stochastic systems and arise naturally in areas of current interest such as in finance. A new Riccati equation is introduced in this paper as an appropriate vehicle for identifying optimal controls and calculating the optimal cost value.

Other properties concerning the existence, uniqueness, and asymptotic behavior of solutions to the GRE associated with an indefinite LQ problem are studied in a companion paper [1], which complements results derived in this paper.

REFERENCES

[1] M. Ait Rami, X. Chen, J. B. Moore, and X. Y. Zhou, *Solvability and asymptotic behavior of generalized Riccati equations arising in indefinite stochastic LQ controls*, IEEE Trans. Automat. Control, 46 (2001), pp. 428–440.

[2] M. Ait Rami and X. Y. Zhou, *Linear matrix inequalities, Riccati equations, and indefinite stochastic linear quadratic control*, IEEE Trans. Automat. Control, 45 (2000), pp. 1131–1143.

[3] A. Albert, *Conditions for positive and nonnegative definiteness in terms of pseudoinverses*, SIAM J. Appl. Math., 17 (1969), pp. 434–440.

[4] B. D. O. Anderson and J. B. Moore, *Optimal Control: Linear Quadratic Methods*, Prentice–Hall, Englewood Cliffs, NJ, 1989.
[5] M. Athans, *The matrix minimum principle*, Inform. and Control, 11 (1968), pp. 592–606.
[6] M. Athens, *Special issues on linear–quadratic–Gaussian problem*, IEEE Trans. Automat. Control, 16 (1971), pp. 527–869.
[7] S. Chen, X. Li, and X. Y. Zhou, *Stochastic linear quadratic regulators with indefinite control weight costs*, SIAM J. Control Optim., 36 (1998), pp. 1685–1702.
[8] S. Chen and X. Y. Zhou, *Stochastic linear quadratic regulators with indefinite control weight costs*. II, SIAM J. Control Optim., 39 (2000), pp. 1065–1081.
[9] S. Chen and J. Yong, *Stochastic linear quadratic optimal control problems*, Appl. Math. Optim., 43 (2001), pp. 21–45.
[10] D. Clements, B. D. O. Anderson, and P. J. Moylan, *Matrix inequality solution to linear-quadratic singular control problems*, IEEE Trans. Automat. Control, 22 (1977), pp. 55–57.
[11] M. H. A. Davis, *Linear Estimation and Stochastic Control*, Chapman and Hall, London, 1977.
[12] D. Hinrichsen and A. J Pritchard, *Stochastic H^∞*, SIAM J. Control Optim., 36 (1998), pp. 1504–1538.
[13] D. H. Jacobson, *Totally singular quadratic minimization problems*, IEEE Trans. Automat. Control, 16 (1971), pp. 651–658.
[14] R. E. Kalman, *Contribution to the theory of optimal control*, Bol. Soc. Mat. Mexican, 5 (1960), pp. 102–119.
[15] M. Kohlmann and X. Y. Zhou, *Relationship between backward stochastic differential equations and stochastic controls: A linear quadratic approach*, SIAM J. Control Optim., 38 (2000), pp. 1392–1407.
[16] A. E. B. Lim and X. Y. Zhou, *Stochastic optimal LQR control with integral quadratic constraints and indefinite control weights*, IEEE Trans. Automat. Control, 44 (1999), pp. 359–369.
[17] A. E. B. Lim and X. Y. Zhou, *Mean–Variance Portfolio Selection with Random Parameters*, preprint.
[18] B. P. Molinari, *Nonnegativity of a quadratic functional*, SIAM J. Control, 13 (1975), pp. 792–806.
[19] R. Penrose, *A generalized inverse of matrices*, Proc. Cambridge Philos. Soc., 52 (1955), pp. 17–19.
[20] V. M. Popov, *Hyperstability and optimality of automatic systems with several control functions*, Rev. Roumaine Sci. Tech. Sér. Electrotech. Energ., 9 (1964), pp. 629–690.
[21] L. Vandenerghe and S. Boyd, *Semidefinite programming*, SIAM Rev., 38 (1996), pp. 49–95.
[22] J. C. Willems, *Least squares stationary control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, 16 (1971), pp. 621–234.
[23] W. M. Wonham, *On the separation theorem of stochastic control*, SIAM J. Control, 6 (1968), pp. 312–326.
[24] D. D. Yao, S. Zhang, and X. Y. Zhou, *Stochastic linear quadratic control via semidefinite programming*, SIAM J. Control Optim., 40 (2001), pp. 801–823.
[25] X. Y. Zhou and D. Li, *Continuous-time mean-variance portfolio selection: A stochastic LQ framework*, Appl. Math. Optim., 42 (2000), pp. 19–33.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 4, pp. 1312–1331

ROBUST CONTROLLER SYNTHESIS FOR UNCERTAIN TIME-VARYING SYSTEMS*

CAROL PIRIE† AND GEIR E. DULLERUD†

Abstract. In this paper we develop and present new convex synthesis conditions for robust performance in linear time-varying systems, subject to time-varying perturbations. In particular, these results are *exact* for sensitivity minimization in the presence of multiplicative perturbations. The methods apply to a number of additional robust performance problems and are always both necessary and sufficient when the system plant is periodic. Otherwise, the conditions provided are *always* sufficient, and a controller can be directly constructed when this condition holds.

Key words. uncertain system, time-varying system, discrete-time system, controller synthesis, convex methods, linear matrix inequality

AMS subject classifications. 93D09, 93C55, 93B35

PII. S0363012900369630

1. Introduction.

The focus of this paper is the development of synthesis tools for robust performance of linear time-varying (LTV) systems. The primary engineering motivation for this work is its application to nonlinear trajectory tracking, where the LTV model arises from linearization of the nonlinear model around a nominal trajectory. This is the generalization of linearizing a nonlinear model around an equilibrium point to obtain a linear time-invariant (LTI) model. The work is also applicable to systems which naturally have time-varying dynamics, such as periodic multirate or sampled-data systems [2, 4].

The main result of the paper is a convex synthesis condition for the existence of a controller, which provides robust *performance* to either additive or multiplicative LTV perturbations; this condition is both necessary and sufficient. When the above existence condition is satisfied we indicate how an explicit controller can be computed for implementation via a convex program. As far as we are aware, these are the first synthesis results for *exact* robust performance of LTV systems; namely, the condition obtained is both necessary and sufficient and convex. The methods developed apply to a more general class of robust synthesis problems and are always both necessary and sufficient when the nominal LTV system is either periodic or satisfies a partitioning condition.

The approach used in the paper combines the LTI results reported in [7, 8] with the LTV framework developed in [9]; for independent, closely related LTV tools see [1, 12, 13]. This work builds on work which considered synthesis of LTI discrete-time systems subject to spatial constraints on the inputs and outputs. As shown in [7], some robust control problems, such as sensitivity minimization with multiplicative uncertainty, may be cast in terms of analysis of systems with spatial constraints. We show that a similar theory holds for LTV systems. We remark that a related robustness problem is solved in [22] for LTI systems, with LTI perturbations, and the obtained results were convex but infinite dimensional in general. The general

*Received by the editors March 9, 2000; accepted for publication (in revised form) June 5, 2001; published electronically January 4, 2002. This work was supported by NSF grant ECS-9875244 CAREER and AFOSR grant F49620-98-1-0416.
http://www.siam.org/journals/sicon/40-4/36963.html
†Department of Mechanical and Industrial Engineering, University of Illinois, Urbana, IL 61801 (dullerud@uiuc.edu).

machinery used here has its roots in the related papers [11, 15]. The work presented here is based on [18].

The results are given in terms of linear operator inequalities, which always have a convex solution space, but in general are infinite dimensional. When the initial LTV plant is periodic, or the synthesis problem is only considered over a finite horizon, these conditions reduce to linear matrix inequalities (LMIs). Namely, in these two important cases the results become finite dimensional and convex and thus readily computable.

The paper is organized as follows: section 2 introduces the notation used and the basic LTV machinery; in section 3 we provide a solution to the square ℓ_2 optimization problem, which is a generalization of the H_∞ control problem; in section 4 we show how this synthesis result can be used for robust synthesis.

2. Background.

2.1. Notation and definitions. Given a normed linear space $\mathcal{X}$, we denote the closed unit ball by $\mathcal{B}\mathcal{X}$. If $\mathcal{W}$ is also a normed linear space, the space of bounded linear operators mapping $\mathcal{X}$ to $\mathcal{W}$ will be denoted by $\mathcal{L}(\mathcal{X}, \mathcal{W})$. The induced norm of an operator A in this space will be denoted $\|A\|_{\mathcal{X}\to\mathcal{W}}$; we suppress the subscripts when the spaces are clear. We use A^* to denote the adjoint operator.

We will use the abbreviation $\mathcal{L}(\mathcal{X})$ for $\mathcal{L}(\mathcal{X}, \mathcal{X})$, and we say that an operator in this set is invertible if the algebraic inverse exists and is a bounded operator on $\mathcal{X}$. We state the following standard small-gain result for later reference.

PROPOSITION 2.1. *Suppose $\mathcal{X}$ is a Banach space and $A \in \mathcal{L}(\mathcal{X})$. If $\|A\|_{\mathcal{X}\to\mathcal{X}} < 1$, then $I - A$ is invertible, and further $\|(I-A)^{-1}\|_{\mathcal{X}\to\mathcal{X}} \le \frac{1}{1-\|A\|_{\mathcal{X}\to\mathcal{X}}}$.*

If $\mathcal{X}$ is a Hilbert space and an operator A is self-adjoint, we use $A > 0$ to mean that there exists an $\epsilon > 0$ such that

$$\langle x, Ax\rangle_{\mathcal{X}} \ge \epsilon \|x\|_{\mathcal{X}}^2 \qquad \text{for every } x \in \mathcal{X}.$$

A useful result is the following.

PROPOSITION 2.2 (Schur complement formula). *Suppose $X \in \mathcal{L}(\mathcal{X})$, $Y \in \mathcal{L}(\mathcal{Y})$, $W \in \mathcal{L}(\mathcal{Y}, \mathcal{X})$, and X and Y are self-adjoint. Then*

$$\begin{bmatrix} X & W \\ W^* & Y \end{bmatrix} < 0$$

if and only if $Y < 0$ and $X - WY^{-1}W^* < 0$.

Given a Hilbert space $\mathcal{X}$, we will denote by $\ell_2(\mathcal{X})$ the Hilbert space of square summable sequences of elements of $\mathcal{X}$ with the usual inner product; that is, $x, y \in \ell_2(\mathcal{X})$; then

$$\langle x, y\rangle = \sum_{k=1}^{\infty} \langle x_k, y_k\rangle_{\mathcal{X}},$$

where $x_k, y_k \in \mathcal{X}$. For simplicity we will denote $\ell_2(\mathbb{C}^n)$ by ℓ_2 regardless of the spatial dimension n. Additionally, we will use $\|x\|$ to denote $\sqrt{\langle x, x\rangle}$, the standard norm on this space.

The upper and lower linear fractional transformations of operators in $\mathcal{L}(\ell_2)$ are $\mathcal{F}_u(M, N) = M_{22} + M_{21}N(I - M_{11}N)^{-1}M_{12}$ and $\mathcal{F}_\ell(M, N) = M_{11} + M_{12}N(I - M_{22}N)^{-1}M_{21}$, respectively, where M is partitioned compatibly with N.

2.2. Memoryless operators. Our goal in this section is to define a particular type of operator and an important associated operation which makes working with LTV state space systems nearly identical to operations on LTI state space systems. See [9] for a complete treatment of memoryless operators and their use in LTV systems.

We make the following definition.

DEFINITION 2.3. *A bounded operator Q on $\ell_2(\mathcal{X})$ is* memoryless *if there exists a sequence of operators Q_k in $\mathcal{L}(\mathcal{X})$ such that, for all w, z, if $z = Qw$, then $z_k = Q_k w_k$. Then Q has the representation*

$$\begin{bmatrix} Q_0 & & & 0 \\ & Q_1 & & \\ & & Q_2 & \\ 0 & & & \ddots \end{bmatrix}.$$

Further, if $P_k \in \mathcal{L}(\mathcal{X})$ is a uniformly bounded sequence of operators, we say $P = \operatorname{diag}(P_0, P_1, \dots)$ is the memoryless operator for P_k, and conversely, given that P is a memoryless operator, the blocks are denoted by P_k for $k \in \mathbb{N}_0$.

Suppose F, G, R, and S are memoryless operators, and let A be a *partitioned operator*, each of whose elements is a memoryless operator such as

$$A = \begin{bmatrix} F & G \\ R & S \end{bmatrix}.$$

We now define the following notation:

$$\left[\!\!\left[\begin{array}{cc} F & G \\ R & S \end{array} \right]\!\!\right] = \operatorname{diag}\left(\begin{bmatrix} F_0 & G_0 \\ R_0 & S_0 \end{bmatrix}, \begin{bmatrix} F_1 & G_1 \\ R_1 & S_1 \end{bmatrix}, \dots, \right),$$

which we call the *memoryless realization* of A. Clearly, for any given operator A of this particular structure, $[\![A]\!]$ is simply A with the rows and columns permuted appropriately so that

$$\left[\!\!\left[\begin{array}{cc} F & G \\ R & S \end{array} \right]\!\!\right]_k = \begin{bmatrix} F_k & G_k \\ R_k & S_k \end{bmatrix}.$$

The following is immediate.

PROPOSITION 2.4. *For any real number β and any partitioned operator A consisting of elements which are memoryless, $A < \beta I$ holds if and only if $[\![A]\!] < \beta I$. That is, positivity is preserved under permutation.*

Two further useful facts for the above permutations are the following.

PROPOSITION 2.5.

(i) *Suppose that A and B are partitioned operators consisting of memoryless elements and that their structures are the same. Then*

$$[\![A + B]\!] = [\![A]\!] + [\![B]\!].$$

(ii) *Suppose that A and C are partitioned operators, each of which consists of elements which are memoryless. Further suppose that the block structures are compatible so that the product $\hat{A}\hat{C}$ is memoryless for any operators $\hat{A}$ and $\hat{C}$ with the same block structures as A and C. Then*

$$[\![AC]\!] = [\![A]\!]\,[\![C]\!].$$

2.3. LTV systems. Although the analysis results presented in this paper will apply to any LTV system, the synthesis result (Theorem 3.10) applies only to state space systems. For this reason we briefly review LTV state space systems; see [9] for an in-depth treatment of the theory. Suppose we are considering the time-varying difference equation

$$\begin{aligned} x_{k+1} &= A_k x_k + B_k u_k, \qquad x_0 = 0, \\ y_k &= C_k x_k + D_k u_k, \end{aligned}$$

where A_k, B_k, C_k, and D_k are bounded *real* matrix sequences. Then clearly these sequences define memoryless operators A, B, C, and D, and therefore the above system may be written more compactly in operator form as

$$\begin{aligned} x &= ZAx + ZBu, \\ y &= Cx + Du, \end{aligned}$$

where Z is the shift, or delay, operator on ℓ_2. Thus we can write the map from u to y formally as

$$u \mapsto y = C(I - ZA)^{-1}ZB + D.$$

It is possible to show that the operator $I - ZA$ is invertible exactly when the state space LTV system is exponentially stable. Throughout the paper we will say an open- or closed-loop LTV state space system is *stable* when its A-operator satisfies the above invertibility condition.

A special case of LTV systems which will be of interest to us are the periodic systems. An operator $M \in \mathcal{L}(\ell_2)$ is said to be n-periodic if $Z^n M = M Z^n$.

3. Time-varying square ℓ_2 synthesis. The square ℓ_2 control problem involves the system depicted in Figure 3.1, where the input signal w and the output signal z are partitioned spatially into n_w and n_z *vector-valued* channels, respectively. Let P_i be the operator which projects onto the ith vector-valued channel of w, and similarly, let Q_j be the operator which projects onto the jth vector-valued channel of z. The standard problem of minimizing the ℓ_2-induced norm of the system, considered in [9], does not directly address such additional spatial system structure. Considering such structure will allow us to formulate a number of robust synthesis problems and may also be more representative of physical situations to place constraints on the norm of each channel of w independently. With this objective in mind, we define the square ℓ_2 problem.

PROBLEM 3.1. *Define the square ℓ_2 induced norm of a system M to be*

$$\|M\|_{sq} = \sup_{\|P_i w\| \le 1} \sum_{j=1}^{n_z} \|Q_j M w\|,$$

and let $\gamma_{opt} = \inf_K \|\mathcal{F}_\ell(G,K)\|_{sq}$. Given $\gamma > 0$, the suboptimal square ℓ_2 problem is to determine whether $\gamma > \gamma_{opt}$ and then, where possible, to find a controller K such that

$$\|\mathcal{F}_\ell(G,K)\|_{sq} < \gamma.$$

The choice of the norm $\sum_{j=1}^{n_z} \|Q_j z\|$ on the output z allows the robust control problem of section 4 to be incorporated into the square ℓ_2 framework.

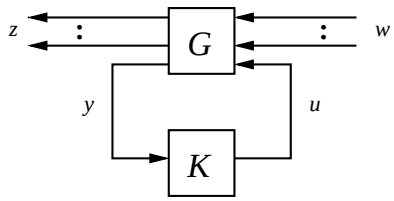


FIG. 3.1. *Square ℓ_2 synthesis arrangement.*

The square ℓ_2 problem is treated for time-invariant systems in [7, 8]. Here we extend its treatment to time-varying systems. The first task is to provide for analysis of a given stable system M; i.e., given γ, is $\|M\|_{sq} < \gamma$? A sufficient condition for M to be contractive with respect to the square norm is given as a scaled small-gain condition on the system, and this condition is further shown to be exact (necessary as well as sufficient) for systems which are *either* periodic or satisfy $n_w \cdot n_z \leq 3$. For a larger class of systems, a related small-gain condition is shown to be necessary for the system to be contractive; however, this condition is not sufficient. A sufficient synthesis which is convex but in general infinite dimensional is then considered for Problem 3.1.

To work effectively with the square norm performance criterion, we introduce a new set of norms on the space ℓ_2. We define the input space $\mathcal{W}_p$ to consist of the elements of ℓ_2, equipped with the norm

$$\|w\|_{\mathcal{W}_p} = |(\|P_1 w\|, \ldots, \|P_{n_w} w\|)|_p,$$

where $|\cdot|_p$ denotes the standard p-norm on $\mathbb{C}^{n_w}$. We similarly define the output space $\mathcal{Z}_p$ via the projections Q_j. For later reference we state the following elementary facts about these new spaces.

PROPOSITION 3.2. *Suppose $1 \leq p \leq \infty$. Then the following hold:*

(a) *The norm $\|\cdot\|_{\mathcal{W}_p}$ is equivalent to the ℓ_2-norm; in particular, $\|w\| \leq \sqrt{n_w}\|w\|_{\mathcal{W}_\infty}$.*
(b) *The mapping A is in $\mathcal{L}(\mathcal{W}_p)$ if and only if $A \in \mathcal{L}(\ell_2)$.*
(c) *If $1 < q \leq \infty$ satisfies $q^{-1} + p^{-1} = 1$, then $\mathcal{W}_q$ represents the normed dual space of $\mathcal{W}_p$ via the ℓ_2 inner product. That is, if f is a linear functional in $\mathcal{L}(\mathcal{W}_p, \mathbb{C})$ with norm α, then there exists $v \in \mathcal{W}_q$ with norm α such that*

$$f(w) = \langle v, w \rangle \qquad \text{for each } w \in \mathcal{W}_p.$$

The analogous statements hold for $\mathcal{Z}_p$.

The above follows from standard analysis, and the proof is omitted.

Having defined these new spaces it is routine to verify that

$$\|M\|_{sq} = \|M\|_{\mathcal{W}_\infty \to \mathcal{Z}_1}. \tag{3.1}$$

That is, the square performance norm is an induced norm.

To simplify the presentation, we define the properties required for the scaling operators in the scaled small-gain conditions of this section.

DEFINITION 3.3. *For a given γ, the time-invariant memoryless operators S, $T \in \mathcal{L}(\ell_2)$ are* γ-admissible scales *if they are of the form*

$$[\![S]\!]_k = \begin{bmatrix} s_1 I & & \\ & \ddots & \\ & & s_{n_w} I \end{bmatrix}, \quad [\![T]\!]_k = \begin{bmatrix} t_1 I & & \\ & \ddots & \\ & & t_{n_z} I \end{bmatrix}$$

and satisfy the condition

$$\sum_{i=1}^{n_w} s_i + \sum_{j=1}^{n_z} t_j < 2\gamma, \text{ where every } s_i > 0 \text{ and } t_j > 0.$$

Note, therefore, that $S = \sum_{i=1}^{n_w} s_i P_i^* P_i$ and $T = \sum_{j=1}^{n_z} t_j Q_j^* Q_j$. The following provides sufficiency in the form of a scaled small-gain condition for contractiveness in the square norm.

THEOREM 3.4. *Suppose that $M \in \mathcal{L}(\ell_2)$ and γ is given. Then if there exist γ-admissible scales S and T such that*

$$\left\| T^{-\frac{1}{2}} M S^{-\frac{1}{2}} \right\|_{\ell_2 \to \ell_2} < 1,$$

then

$$\|M\|_{sq} < \gamma.$$

Proof. The result will be proven for $\gamma = 1$ since in the general case the γ may be absorbed into the plant M. Since the sum of the s_i and t_j is less than 2 by hypothesis, it is routine to see that there must exist an $\alpha > 0$ such that

$$\alpha \sum_{i=1}^{n_w} s_i < 1 \text{ and } \alpha^{-1} \sum_{j=1}^{n_z} t_j < 1.$$

To begin, recall that $\mathcal{Z}_\infty$ represents the normed dual space of $\mathcal{Z}_1$ via the inner product on ℓ_2. Thus we have that the square norm satisfies

$$\|M\|_{\mathcal{W}_\infty \to \mathcal{Z}_1} = \sup_{v \in \mathcal{B}\mathcal{Z}_\infty,\, w \in \mathcal{B}\mathcal{W}_\infty} \operatorname{Re} \langle v, Mw \rangle .$$

We now show that $\mathcal{B}\mathcal{W}_\infty \subset \alpha^{-1} S^{-\frac{1}{2}} \mathcal{B}\ell_2$. To see this, choose $w \in \mathcal{B}\mathcal{W}_\infty$, and set $u = \alpha S^{\frac{1}{2}} w$. Then clearly $w = \alpha^{-1} S^{-\frac{1}{2}} u$ and

$$\|u\|^2 = \alpha s_1 \|P_1 w\|^2 + \cdots + \alpha s_{n_w} \|P_{n_w} w\|^2 < 1,$$

since by definition each $\|P_j w\| \le 1$. An identical argument shows $\mathcal{B}\mathcal{Z}_\infty \subset \alpha T^{-\frac{1}{2}} \mathcal{B}\ell_2$.

Now, returning to the above supremum condition, we get

$$\begin{aligned} \sup_{v \in \mathcal{B}\mathcal{Z}_\infty,\, w \in \mathcal{B}\mathcal{W}_\infty} \operatorname{Re} \langle v, Mw \rangle &\le \sup_{v \in \mathcal{B}\ell_2,\, w \in \mathcal{B}\ell_2} \operatorname{Re} \left\langle \alpha T^{-\frac{1}{2}} v, \alpha^{-1} M S^{-\frac{1}{2}} w \right\rangle \\ &= \sup_{v \in \mathcal{B}\ell_2,\, w \in \mathcal{B}\ell_2} \operatorname{Re} \left\langle v, T^{-\frac{1}{2}} M S^{-\frac{1}{2}} w \right\rangle . \end{aligned}$$

We conclude by observing that the right-hand side is equal to $\|T^{-\frac{1}{2}} M S^{-\frac{1}{2}}\|_{\ell_2 \to \ell_2}$. ∎

Theorem 3.4 provides sufficiency of the scaled small-gain condition $\|T^{-\frac{1}{2}} M S^{-\frac{1}{2}}\|_{\ell_2 \to \ell_2} < 1$ for any operator, and we now investigate necessity. For either periodic systems or systems which satisfy $n_w \cdot n_z \le 3$, we obtain exactness of the scaled small-gain condition through the following result.

THEOREM 3.5. *Suppose $M \in \mathcal{L}(\ell_2)$ and satisfies $\|M\|_{sq} < \gamma$. If either*
(a) *M is periodic, or*
(b) *the product $n_w \cdot n_z \leq 3$,*
then there exist γ-admissible scales S and T such that $\|T^{-\frac{1}{2}} M S^{-\frac{1}{2}}\|_{\ell_2 \to \ell_2} < 1$.

The proof of this theorem is lengthy and is provided in Appendix A. A proof of the time-invariant case is outlined in [7, Theorem 2].

REMARK 3.6. *Here we are working with the complex space $\ell_2(\mathbb{C}^n)$. If instead sequences in the real space $\ell_2(\mathbb{R}^n)$ are considered, a version of Theorem* 3.5 *still holds; however, the inequality in condition* (b) *must be replaced by the more restrictive condition $n_w \cdot n_z \leq 2$. See Remark* A.5.

A more general necessary condition may be developed for time-varying systems which fail to be periodic but which demonstrate a periodic behavior after an initial transient period.

DEFINITION 3.7. *A system M is* eventually periodic *if there is a $k \in \mathbb{N}$ such that the system $Z^{*k} M Z^k$ is periodic.*

With this observation the following generalization of Theorem 3.5 is almost immediate.

PROPOSITION 3.8. *Suppose M is eventually periodic and $\|M\|_{sq} < \gamma$. Then there exist γ-admissible scales S and T such that*

$$\left\| T^{-\frac{1}{2}} \left\{ Z^{*k} M Z^k \right\} S^{-\frac{1}{2}} \right\|_{\ell_2 \to \ell_2} < 1.$$

Proof. We employ the submultiplicative inequality to get

$$\|Z^{*k} M Z^k\|_{\mathcal{W}_\infty \to \mathcal{Z}_1} \leq \|Z^{*k}\|_{\mathcal{Z}_1 \to \mathcal{Z}_1} \|M\|_{\mathcal{W}_\infty \to \mathcal{Z}_1} \|Z^k\|_{\mathcal{W}_\infty \to \mathcal{W}_\infty} = \|M\|_{\mathcal{W}_\infty \to \mathcal{Z}_1} < \gamma.$$

By hypothesis, $Z^{*k} M Z^k$ is periodic; thus the necessity of the scaled small-gain condition follows directly from Theorem 3.5. ∎

Given a system which is eventually periodic but not periodic, this proposition allows a lower bound on the square norm of M to be found. If the scaled small-gain condition of Theorem 3.4 fails for a given value of γ, no conclusion may be made as to whether $\|M\|_{sq} < \gamma$; however, if there are no γ-admissible scales such that $\|T^{-\frac{1}{2}} \{Z^{*k} M Z^k\} S^{-\frac{1}{2}}\|_{\ell_2 \to \ell_2} < 1$, then necessarily $\|M\|_{sq} \geq \gamma$.

To address Problem 3.1, we apply the analysis result of Theorem 3.4 to the closed loop. If, for a given controller K, there exist γ-admissible scales such that

$$\left\| T^{-\frac{1}{2}} \mathcal{F}_\ell(G, K) S^{-\frac{1}{2}} \right\|_{\ell_2 \to \ell_2} < 1, \tag{3.2}$$

then K provides a synthesis such that the closed loop is contractive in the square norm. The key point in solving the synthesis problem is that for fixed scales S and T, finding a K satisfying (3.2) is exactly the problem of finding a controller which makes the closed loop contractive in the induced ℓ_2-norm for the system

$$\hat{G} = \left[\begin{array}{c|cc} \hat{A} & \hat{B}_1 & \hat{B}_2 \\ \hline \hat{C}_1 & \hat{D}_{11} & \hat{D}_{12} \\ \hat{C}_2 & \hat{D}_{21} & 0 \end{array}\right] \equiv \left[\begin{array}{c|cc} A & B_1 S^{-\frac{1}{2}} & B_2 \\ \hline T^{-\frac{1}{2}} C_1 & T^{-\frac{1}{2}} D_{11} S^{-\frac{1}{2}} & T^{-\frac{1}{2}} D_{12} \\ C_2 & D_{21} S^{-\frac{1}{2}} & 0 \end{array}\right],$$

where the notation

$$\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right]$$

signifies the LTV system defined by the memoryless operators A, B, C, D. The proof of the next result uses this idea and the solution from [9] to the induced ℓ_2 synthesis problem and uses the LTI proof in [8] as a formal template.

LEMMA 3.9. *Suppose S and T are γ-admissible scales. There exists a controller K such that the closed loop is internally stable and $\|T^{-\frac{1}{2}}\mathcal{F}_\ell(G,K)S^{-\frac{1}{2}}\|_{\ell_2\to\ell_2} < 1$ if and only if there exist memoryless operators P, Q, $X > 0$, and $Y > 0$ satisfying*

$$N_X^*\left\{\begin{bmatrix} A & B_1 \\ C_1 & D_{11}\end{bmatrix}\begin{bmatrix} X & 0 \\ 0 & P\end{bmatrix}\begin{bmatrix} A & B_1 \\ C_1 & D_{11}\end{bmatrix}^* - \begin{bmatrix} Z^*XZ & 0 \\ 0 & T\end{bmatrix}\right\}N_X < 0, \tag{3.3}$$

$$N_Y^*\left\{\begin{bmatrix} ZA & ZB_1 \\ C_1 & D_{11}\end{bmatrix}^*\begin{bmatrix} Y & 0 \\ 0 & Q\end{bmatrix}\begin{bmatrix} ZA & ZB_1 \\ C_1 & D_{11}\end{bmatrix} - \begin{bmatrix} Y & 0 \\ 0 & S\end{bmatrix}\right\}N_Y < 0, \tag{3.4}$$

$$\begin{bmatrix} X & I \\ I & Y\end{bmatrix} \geq 0, \quad \begin{bmatrix} P & I \\ I & S\end{bmatrix} \geq 0, \quad \begin{bmatrix} Q & I \\ I & T\end{bmatrix} \geq 0, \tag{3.5}$$

where

$$\mathrm{Im}N_X = \ker\begin{bmatrix} B_2^* & D_{12}^*\end{bmatrix}, \quad N_X^*N_X = I,$$
$$\mathrm{Im}N_Y = \ker\begin{bmatrix} C_2 & D_{21}\end{bmatrix}, \quad N_Y^*N_Y = I.$$

Proof. Suppose there is a solution, K, for the scaled synthesis problem. Applying the time-varying induced ℓ_2 synthesis to the system $\hat{G}$ defined above, $\|\mathcal{F}_\ell(\hat{G},K)\|_{\ell_2\to\ell_2} < 1$ if and only if there are memoryless operators $X > 0$ and $Y > 0$ which solve the three operator inequalities

$$\hat{N}_X^*\left\{\begin{bmatrix} \hat{A} & \hat{B}_1 \\ \hat{C}_1 & \hat{D}_{11}\end{bmatrix}\begin{bmatrix} X & 0 \\ 0 & I\end{bmatrix}\begin{bmatrix} \hat{A} & \hat{B}_1 \\ \hat{C}_1 & \hat{D}_{11}\end{bmatrix}^* - \begin{bmatrix} Z^*XZ & 0 \\ 0 & I\end{bmatrix}\right\}\hat{N}_X < 0, \tag{3.6}$$

$$\hat{N}_Y^*\left\{\begin{bmatrix} Z\hat{A} & Z\hat{B}_1 \\ \hat{C}_1 & \hat{D}_{11}\end{bmatrix}^*\begin{bmatrix} Y & 0 \\ 0 & I\end{bmatrix}\begin{bmatrix} Z\hat{A} & Z\hat{B}_1 \\ \hat{C}_1 & \hat{D}_{11}\end{bmatrix} - \begin{bmatrix} Y & 0 \\ 0 & I\end{bmatrix}\right\}\hat{N}_Y < 0, \tag{3.7}$$

$$\begin{bmatrix} X & I \\ I & Y\end{bmatrix} \geq 0, \tag{3.8}$$

where $\mathrm{Im}\hat{N}_X = \ker\begin{bmatrix} B_2^* & D_{12}^*T^{-\frac{1}{2}}\end{bmatrix}$ and $\mathrm{Im}\hat{N}_Y = \ker\begin{bmatrix} C_2 & D_{21}S^{-\frac{1}{2}}\end{bmatrix}$. Concentrating on the inequality (3.6) and factoring out the scales S and T, we may rewrite it as

$$\hat{N}_X^*\begin{bmatrix} I & 0 \\ 0 & T^{-\frac{1}{2}}\end{bmatrix}\left\{\begin{bmatrix} A & B_1 \\ C_1 & D_{11}\end{bmatrix}\begin{bmatrix} X & 0 \\ 0 & S^{-1}\end{bmatrix}\begin{bmatrix} A & B_1 \\ C_1 & D_{11}\end{bmatrix}^* - \begin{bmatrix} Z^*XZ & 0 \\ 0 & T\end{bmatrix}\right\}\begin{bmatrix} I & 0 \\ 0 & T^{-\frac{1}{2}}\end{bmatrix}\hat{N}_X < 0.$$

Noting that

$$\mathrm{Im}\left(\hat{N}_X\begin{bmatrix} I & 0 \\ 0 & T^{-\frac{1}{2}}\end{bmatrix}\right) = \ker\begin{bmatrix} B_2^* & D_{12}^*\end{bmatrix} = \mathrm{Im}N_X,$$

this inequality has the equivalent form

$$N_X^*\left\{\begin{bmatrix} A & B_1 \\ C_1 & D_{11}\end{bmatrix}\begin{bmatrix} X & 0 \\ 0 & S^{-1}\end{bmatrix}\begin{bmatrix} A & B_1 \\ C_1 & D_{11}\end{bmatrix}^* - \begin{bmatrix} Z^*XZ & 0 \\ 0 & T\end{bmatrix}\right\}N_X < 0.$$

A similar procedure allows (3.7) to be expressed as

$$N_Y^* \left\{ \begin{bmatrix} ZA & ZB_1 \\ C_1 & D_{11} \end{bmatrix}^* \begin{bmatrix} Y & 0 \\ 0 & T^{-1} \end{bmatrix} \begin{bmatrix} ZA & ZB_1 \\ C_1 & D_{11} \end{bmatrix} - \begin{bmatrix} Y & 0 \\ 0 & S \end{bmatrix} \right\} N_Y < 0.$$

Taking $P = S^{-1}$ and $Q = T^{-1}$, immediately X, Y, P, and Q satisfy (3.3)–(3.5) for the given values of S and T.

Conversely, suppose there is a solution to the inequalities (3.3)–(3.5). Applying Schur complements to (3.5) results in $P \geq S^{-1}$ and $Q \geq T^{-1}$. Thus

$$\begin{bmatrix} X & 0 \\ 0 & S^{-1} \end{bmatrix} \leq \begin{bmatrix} X & 0 \\ 0 & P \end{bmatrix},$$

so

$$\begin{bmatrix} A & B_1 \\ C_1 & D_{11} \end{bmatrix} \begin{bmatrix} X & 0 \\ 0 & S^{-1} \end{bmatrix} \begin{bmatrix} A & B_1 \\ C_1 & D_{11} \end{bmatrix}^* \leq \begin{bmatrix} A & B_1 \\ C_1 & D_{11} \end{bmatrix} \begin{bmatrix} X & 0 \\ 0 & P \end{bmatrix} \begin{bmatrix} A & B_1 \\ C_1 & D_{11} \end{bmatrix}^*,$$

and hence

$$\begin{aligned} &N_X^* \left\{ \begin{bmatrix} A & B_1 \\ C_1 & D_{11} \end{bmatrix} \begin{bmatrix} X & 0 \\ 0 & S^{-1} \end{bmatrix} \begin{bmatrix} A & B_1 \\ C_1 & D_{11} \end{bmatrix}^* - \begin{bmatrix} Z^*XZ & 0 \\ 0 & T \end{bmatrix} \right\} N_X \\ &\leq N_X^* \left\{ \begin{bmatrix} A & B_1 \\ C_1 & D_{11} \end{bmatrix} \begin{bmatrix} X & 0 \\ 0 & P \end{bmatrix} \begin{bmatrix} A & B_1 \\ C_1 & D_{11} \end{bmatrix}^* - \begin{bmatrix} Z^*XZ & 0 \\ 0 & T \end{bmatrix} \right\} N_X < 0. \end{aligned}$$

Similarly, since $Q \geq T^{-1}$,

$$N_Y^* \left\{ \begin{bmatrix} ZA & ZB_1 \\ C_1 & D_{11} \end{bmatrix}^* \begin{bmatrix} Y & 0 \\ 0 & T^{-1} \end{bmatrix} \begin{bmatrix} ZA & ZB_1 \\ C_1 & D_{11} \end{bmatrix} - \begin{bmatrix} Y & 0 \\ 0 & S \end{bmatrix} \right\} N_Y < 0.$$

The ℓ_2 synthesis may be applied again to show that

$$\left\| \mathcal{F}_\ell \left(\hat{G}, K \right) \right\|_{\ell_2 \to \ell_2} = \left\| T^{-\frac{1}{2}} \mathcal{F}_\ell \left(G, K \right) S^{-\frac{1}{2}} \right\|_{\ell_2 \to \ell_2} < 1. \qquad \square$$

A synthesis which is sufficient for $\|\mathcal{F}_\ell(G,K)\|_{sq} < \gamma$ is now immediate.

THEOREM 3.10. *If there exist γ-admissible scales S and T and memoryless operators P, Q, $X > 0$, and $Y > 0$ satisfying inequalities* (3.3)–(3.5)*, then there exists a stabilizing controller K such that $\|\mathcal{F}_\ell(G,K)\|_{sq} < \gamma$.*

Proof. By Lemma 3.9 and the existence of a solution to the inequalities (3.3)–(3.5),

$$\left\| T^{-\frac{1}{2}} \mathcal{F}_\ell \left(G, K \right) S^{-\frac{1}{2}} \right\|_{\ell_2 \to \ell_2} < \gamma.$$

Applying Theorem 3.4, $\|\mathcal{F}_\ell(G,K)\|_{sq} < 1$. $\square$

Theorem 3.10 provides a sufficient test to determine the existence of a controller synthesis such that $\|\mathcal{F}_\ell(G,K)\|_{sq} < \gamma$. To construct such a controller, in the case that the operator inequalities (3.3)–(3.5) have a solution X, Y, P, Q, S, and T, a controller realization may be obtained by forming the system $\hat{G}$ and proceeding to solve the synthesis operator inequalities for the time-varying ℓ_2 problem; see [9]. It follows that determining feasibility and constructing an admissible controller are both convex, though infinite dimensional, problems. As Theorem 3.10 provides (in general) only a sufficient test, infeasibility of the inequalities (3.3)–(3.5) does not allow any

conclusion to be made as to the existence of an admissible controller. This test is, however, exact for systems which are either periodic or $n_w \cdot n_z \leq 3$ since the analysis condition was shown in Theorem 3.5 to be exact.

It should be noted that the operator inequalities of Theorem 3.10 reduce to finite dimensional LMIs when the system is either periodic or the problem is considered only over a finite horizon. For finite horizon problems the reduction in dimension is immediate since the system realization may be taken to be zero at all time steps after the horizon of the problem. The operator inequalities may be written as a countable collection of coupled LMIs by considering the memoryless realization of the inequalities, and, due to the system realization being zero after the horizon all but a finite number of these LMIs are trivial. Denoting the horizon by k_o the problem thus reduces to the LMIs

$$1.\quad N_{X,k}^* \left\{ \begin{bmatrix} A_k & B_{1,k} \\ C_{1,k} & D_{11,k} \end{bmatrix} \begin{bmatrix} X_k & 0 \\ 0 & P_k \end{bmatrix} \begin{bmatrix} A_k & B_{1,k} \\ C_{1,k} & D_{11,k} \end{bmatrix}^* - \begin{bmatrix} X_{k+1} & 0 \\ 0 & T \end{bmatrix} \right\} N_{X,k} < 0 \quad \text{for } 0 \leq k \leq k_o,$$

$$2.\quad N_{Y,k}^* \left\{ \begin{bmatrix} A_k & B_{1,k} \\ C_{1,k} & D_{11,k} \end{bmatrix}^* \begin{bmatrix} Y_{k+1} & 0 \\ 0 & Q_k \end{bmatrix} \begin{bmatrix} A_k & B_{1,k} \\ C_{1,k} & D_{11,k} \end{bmatrix} - \begin{bmatrix} Y_k & 0 \\ 0 & S \end{bmatrix} \right\} N_{Y,k} < 0 \quad \text{for } 0 \leq k \leq k_o,$$

$$3.\quad \begin{bmatrix} X_k & I \\ I & Y_k \end{bmatrix} \geq 0 \quad \text{for } 0 \leq k \leq k_o + 1,$$

$$4.\quad \begin{bmatrix} P_k & I \\ I & S \end{bmatrix} \geq 0, \begin{bmatrix} Q_k & I \\ I & T \end{bmatrix} \geq 0 \quad \text{for } 0 \leq k \leq k_o,$$

where S and T are now *constant* matrices satisfying the γ-admissible scale properties, and the following conditions on the matrix variables are satisfied:

$$\begin{aligned} X_k > 0, \quad Y_k > 0 \quad & \text{for } 0 \leq k \leq k_o + 1, \\ P_k > 0, \quad Q_k > 0 \quad & \text{for } 0 \leq k \leq k_o. \end{aligned}$$

Note that due to the presence of the shift operator in the original inequalities, the $k_o + 1$ blocks of X and Y are constrained in these LMIs.

For periodic systems the reduction to finite dimensions is a result of an LMI solution to the ℓ_2 problem for periodic systems given in [9]. Applying the same methodology as in Lemma 3.9 and Theorem 3.10 and using the LMI solution in place of the operator inequalities will immediately produce an LMI solution which is exact for the periodic square ℓ_2 synthesis problem. Define the *cyclic shift operator*

$$\tilde{Z} = \begin{bmatrix} 0 & \cdots & 0 & I \\ I & \ddots & & 0 \\ & \ddots & & \vdots \\ & & I & 0 \end{bmatrix},$$

and for a periodic operator F define the first period truncation to be

$$\tilde{F} = \begin{bmatrix} F_0 & & & \\ & \ddots & & \\ & & & F_{n-1} \end{bmatrix}.$$

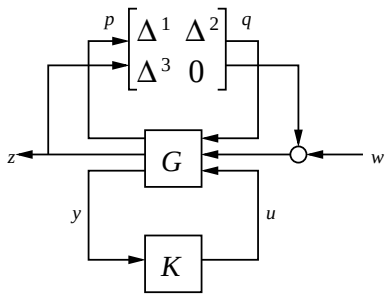


FIG. 4.1. *Structured uncertainty system interconnection.*

COROLLARY 3.11. *There exist γ-admissible scales $\tilde{S}$ and $\tilde{T}$ and matrices $\tilde{P}$, $\tilde{Q}$, $\tilde{X} > 0$, and $\tilde{Y} > 0$ satisfying the LMIs obtained from* (3.3)–(3.5) *by formally replacing the shift operator with the cyclic shift and the other operators with their first period truncations if and only if there exists a stabilizing controller K such that $\|\mathcal{F}_\ell(G,K)\|_{sq} < 1$. Furthermore, this controller may be chosen to be periodic.*

4. Robust synthesis for structured uncertainty. The structured uncertainty control problem of interest involves the system interconnection shown in Figure 4.1, where the input signal w, the output z, and the signals p and q are partitioned into n_w, n_z, n_p, and n_q vector-valued channels, respectively. The uncertainty Δ is assumed to be in $\bar{\mathbf{\Delta}}$, where

$$\bar{\mathbf{\Delta}} = \left\{ \begin{bmatrix} \Delta^1 & \Delta^2 \\ \Delta^3 & 0 \end{bmatrix} : \Delta^1 \in \mathbf{\Delta}_{n_q,n_p}, \Delta^2 \in \mathbf{\Delta}_{n_q,n_z}, \Delta^3 \in \mathbf{\Delta}_{n_w,n_p} \right\}$$

and

$$\mathbf{\Delta}_{n,m} = \left\{ \Delta \in \mathcal{L}(\ell_2) : \|\Delta_{ij}\|_{\ell_2 \to \ell_2} \le 1, i = 1, \dots, n, j = 1, \dots, m \right\}.$$

This uncertainty class encompasses a number of common uncertainty arrangements, and we illustrate this with two specific examples in what follows; see also [7]. Most of the results presented in this section appear in [6, 7]; here we provide alternative proofs in our current context that are concise and complete.

We begin with several definitions.

DEFINITION 4.1. *A linear operator $M \in \mathcal{L}(\ell_2)$ is* robustly stable *to an uncertainty set $\mathbf{\Delta}$ if $I - M\Delta$ is invertible on ℓ_2 for every $\Delta \in \mathbf{\Delta}$. An operator M is* uniformly robustly stable *to $\mathbf{\Delta}$ if, in addition to being robustly stable,*

$$\sup_{\Delta \in \mathbf{\Delta}} \left\| (I - M\Delta)^{-1} \right\|_{\ell_2 \to \ell_2} < \infty.$$

Given operators G and K, we say that K uniformly robustly stabilizes G if $\mathcal{F}_\ell(G,K)$ is uniformly robustly stable. Note that the above definitions remain unchanged if the ℓ_2-norm is replaced by the $\mathcal{Z}_p$-norm since all these norms are equivalent for all p.

The suboptimal robust control problem follows.

PROBLEM 4.2. *Given the structured uncertainty set $\bar{\mathbf{\Delta}}$, a linear system G on ℓ_2, and a desired performance level γ, construct a controller K such that the nominal system $\mathcal{F}_\ell(G,K)$ is uniformly robustly stable and*

$$\sup_{\Delta \in \bar{\mathbf{\Delta}}} \|T_{wz}\|_{sq} < \gamma,$$

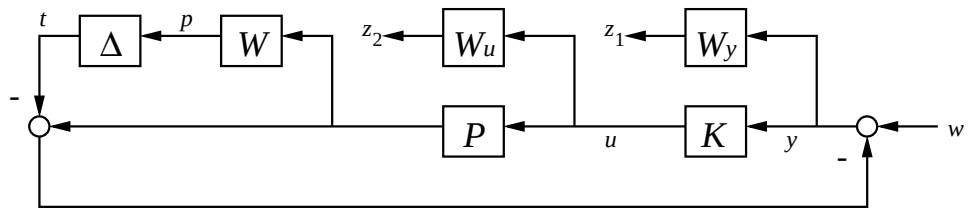


FIG. 4.2. *Generalized sensitivity minimization.*

where $T_{wz} = (I - \mathcal{F}_\ell(G,K)\,\Delta)^{-1}\,\mathcal{F}_\ell(G,K)$ *is the closed-loop map from* w *to* z *in Figure* 4.1.

In the case where $n_w = n_z = 1$, this problem reduces to

$$\sup_{\Delta \in \bar{\mathbf{\Delta}}} \|T_{wz}\|_{\ell_2 \to \ell_2} < \gamma.$$

Structured uncertainty may be transformed to block diagonal uncertainty (to which most existing theory applies) by appropriately redefining the plant G, although at some expense in the size of the problem. However, in the case of Problem 4.2 a direct analysis of the structured uncertainty is more tractable due an underlying connection with the square ℓ_2 problem, which we have shown to have a convex solution. Demonstrating this connection is the topic of this section. It should be noted that not all robust control problems may be formulated as in Figure 4.1 and Problem 4.2. To provide some concreteness the following example demonstrates how to formulate a problem under this framework.

Example. The generalized sensitivity minimization problem is depicted in Figure 4.2, where the aim is to minimize $\|z_1\| + \|z_2\|$ subject to the uncertainty $\|\Delta\|_{\ell_2 \to \ell_2} \le 1$ and $\|w\| \le 1$. Note that this is exactly the standard sensitivity minimization problem if the weight $W_u = 0$. This system may be reconfigured to fit the correct formulation by finding the operator which maps from $t + w$ and u to p, z, and y. From Figure 4.2,

$$\begin{aligned} p &= WPu, \\ z_1 &= W_y(w+t) - W_yPu, \\ z_2 &= W_uu, \\ y &= (w+t) - Pu. \end{aligned}$$

Thus this system is equivalent to the one in Figure 4.3, where

$$G = \begin{bmatrix} 0 & WP \\ \begin{bmatrix} W_y \\ 0 \end{bmatrix} & \begin{bmatrix} -W_yP \\ W_u \end{bmatrix} \\ I & -P \end{bmatrix}.$$

The problem may now be restated as follows: find a uniformly robustly stabilizing controller K such that

$$\sup_{\|\Delta\|_{\ell_2 \to \ell_2} \le 1} \sup_{\|w\| \le 1} \sum_{j=1}^{2} \|Q_j z\| < 1. \tag{4.1}$$

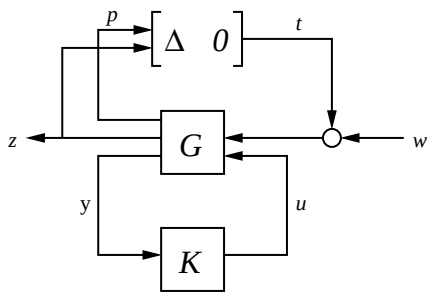


FIG. 4.3. *Reformulation of sensitivity minimization problem.*

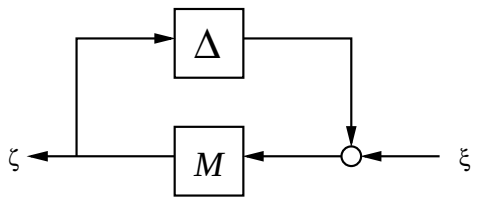


FIG. 4.4. *System for robust stability analysis.*

Thus we see that $n_z = 2$ and $n_w = 1$. To explicitly see that this example is in the framework of Figure 4.1, augment $\begin{bmatrix} \Delta & 0 \end{bmatrix}$ to get

$$\Delta' = \begin{bmatrix} \Delta^1 & \Delta^2 \\ \Delta & 0 \end{bmatrix},$$

where $\|\Delta^k\|_{\ell_2 \to \ell_2} \leq 1$, and augment G to obtain

$$G' = \begin{bmatrix} 0 & 0 & WP \\ \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \begin{bmatrix} W_y \\ 0 \end{bmatrix} & \begin{bmatrix} -W_y P \\ W_u \end{bmatrix} \\ 0 & I & -P \end{bmatrix}.$$

REMARK 4.3. *Note that, from Theorem* 3.5*, we can solve the above example exactly using the methods of this paper since here* $n_w \cdot n_z = 2$.

As an initial step in reformulating Problem 4.2 as a square ℓ_2 problem, we consider robust stability analysis of a related full structured uncertainty. Consider the system depicted in Figure 4.4, where ξ is partitioned into $n_\xi = n_q + n_w$ channels and ζ is partitioned into $n_\zeta = n_z + n_p$ channels.

THEOREM 4.4. *A system* M *is uniformly robustly stable with respect to the full-block structured uncertainty set* $\mathbf{\Delta} = \mathbf{\Delta}_{n_\xi, n_\zeta}$ *if and only if* $\|M\|_{sq} < 1$.

Proof. (If:) Start by choosing $\Delta \in \mathbf{\Delta}$. It is sufficient to show that the inverse $(I - M\Delta)^{-1}$ exists in $\mathcal{L}(\mathcal{Z}_1)$ and is bounded by a constant independent of Δ.

We begin by showing that $\|\Delta\|_{\mathcal{Z}_1 \to \mathcal{W}_\infty} \leq 1$. Take any $q \in \mathcal{Z}_1$. Then

$$\|P_i \Delta q\| = \left\| \sum_{j=1}^{n_\zeta} \Delta_{ij} Q_j q \right\| \leq \sum_{j=1}^{n_\zeta} \|\Delta_{ij}\|_{\ell_2 \to \ell_2} \|Q_j q\| \leq \sum_{j=1}^{n_\zeta} \|Q_j q\| = \|q\|_{\mathcal{Z}_1}.$$

Thus we conclude that $\|\Delta q\|_{\mathcal{W}_\infty} \leq \|q\|_{\mathcal{Z}_1}$, and so Δ is a contractive element of $\mathcal{L}(\mathcal{Z}_1, \mathcal{W}_\infty)$.

Using this and applying the submultiplicative inequality, we have

$$\|M\Delta\|_{\mathcal{Z}_1\to\mathcal{Z}_1} \le \|M\|_{\mathcal{W}_\infty\to\mathcal{Z}_1}\|\Delta\|_{\mathcal{Z}_1\to\mathcal{W}_\infty} \le \|M\|_{\mathcal{W}_\infty\to\mathcal{Z}_1} < 1.$$

Invoking Proposition 2.1, we see that $(I-M\Delta)^{-1}$ exists and that $\|(I-M\Delta)^{-1}\|_{\mathcal{Z}_1\to\mathcal{Z}_1} \le \frac{1}{1-\|M\|_{sq}}$.

(Only if:) This part of the proof is by contradiction. Suppose that M is uniformly robustly stable, and suppose, on the contrary, that $\|M\|_{sq} \ge 1$. Then it suffices to show that, given any $\epsilon > 0$, there exist $\xi \in \mathcal{W}_\infty$, with norm $\|\xi\|_{\mathcal{W}_\infty} = 1$, and $\Delta \in \mathbf{\Delta} = \mathbf{\Delta}_{n_\xi,n_\zeta}$ such that

$$\|(I - \Delta M)\xi\|_{\mathcal{W}_\infty} \le \epsilon.$$

Fix $\epsilon > 0$, and choose $\xi \in \mathcal{W}_\infty$, with norm $\|\xi\|_{\mathcal{W}_\infty} = 1$, such that $\|M\xi\|_{\mathcal{Z}_1} > 1 - \epsilon$.

Let $\alpha = \|M\xi\|_{\mathcal{Z}_1}$, and define the full structured perturbation Δ by

$$\Delta_{ij}(x) = \begin{cases} \frac{1-\epsilon}{\alpha\|Q_jM\xi\|}\langle Q_jM\xi, x\rangle P_i\xi & \text{if } Q_jM\xi \ne 0, \\ 0 & \text{if } Q_jM\xi = 0. \end{cases}$$

For notational simplicity, assume that $Q_jM\xi \ne 0$ for each j. Then, since $1 - \epsilon < \sum_{j=1}^{n_\zeta}\|Q_jM\xi\| = \alpha$ and using the Cauchy–Schwarz inequality,

$$\begin{aligned}\|\Delta_{ij}\|_{\ell_2\to\ell_2} &\le \frac{1-\epsilon}{\alpha\,\|Q_jM\xi\|}\,\|Q_jM\xi\|\,\|P_i\xi\| \\ &= \frac{1-\epsilon}{\alpha}\,\|P_i\xi\| < 1,\end{aligned}$$

so $\Delta \in \mathbf{\Delta} = \mathbf{\Delta}_{n_\xi,n_\zeta}$. Further,

$$\begin{aligned}P_i\Delta M\xi &= \sum_{j=1}^{n_\zeta}\Delta_{ij}\,(Q_jM\xi) \\ &= \sum_{j=1}^{n_\zeta}\frac{1-\epsilon}{\alpha\,\|Q_jM\xi\|}\,\langle Q_jM\xi, Q_jM\xi\rangle\,P_i\xi \\ &= (1-\epsilon)P_i\xi\left(\frac{1}{\alpha}\sum_{j=1}^{n_\zeta}\|Q_jM\xi\|\right) \\ &= (1-\epsilon)P_i\xi,\end{aligned}$$

and hence $\Delta M\xi = (1-\epsilon)\xi$. Thus $\|\,(I - \Delta M)\,\xi\|_{\mathcal{W}_\infty} = \epsilon$. □

Theorem 4.4 reformulates robust stability as a square ℓ_2 analysis problem, and so the problem of finding a robustly stabilizing controller immediately reduces to the square ℓ_2 synthesis problem: find a stabilizing K such that $\|\mathcal{F}_\ell\,(G,K)\|_{sq} < 1$. Problem 4.2 requires further reformulation before the square ℓ_2 results apply, and a technical result is first needed.

COROLLARY 4.5. *Let $\mathcal{A}$ be a set of operators on ℓ_2. Then $\sup_{A\in\mathcal{A}}\|A\|_{sq} < 1$ if and only if $\sup_{\Delta\in\mathbf{\Delta},A\in\mathcal{A}}\left\|(I - A\Delta)^{-1}\right\|_{\ell_2\to\ell_2} < \infty$.*

Proof. The proof of this result is nearly identical to the proof of Theorem 4.4, and the details are omitted. □

We now present an analogue to the so-called main loop theorem from μ theory; see [16]. In this result we show that uniform robust performance is equivalent to

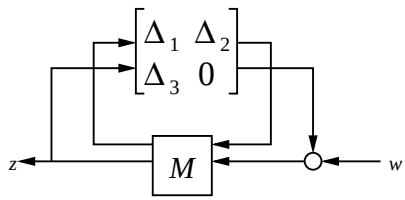


FIG. 4.5. *System for robust performance analysis.*

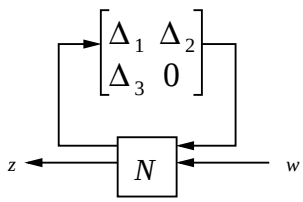


FIG. 4.6. *Equivalent system for robust performance analysis.*

uniform robust stability in an augmented uncertainty structure. This square version of the main loop theorem will provide the basis for the reformulation of Problem 4.2 as a square ℓ_2 problem.

PROPOSITION 4.6. *Define the augmented uncertainty set*

$$\mathbf{\Delta}_p = \left\{ \begin{bmatrix} \bar{\Delta} & 0 \\ 0 & \Delta_4 \end{bmatrix} : \bar{\Delta} \in \bar{\mathbf{\Delta}}, \Delta_4 \in \mathbf{\Delta}_{n_w,n_z} \right\}.$$

With N partitioned compatibly with the inputs and outputs,

$$\sup_{\bar{\Delta}\in\bar{\mathbf{\Delta}}} \left\| \left(I - N_{11}\bar{\Delta}\right)^{-1} \right\|_{\ell_2\to\ell_2} < \infty \text{ and } \sup_{\bar{\Delta}\in\bar{\mathbf{\Delta}}} \left\| \mathcal{F}_u\left(N, \bar{\Delta}\right) \right\|_{sq} < 1$$

if and only if

$$\sup_{\Delta_p\in\mathbf{\Delta}_p} \left\| (I - N\Delta_p)^{-1} \right\|_{\ell_2\to\ell_2} < \infty.$$

The proof of this proposition is a routine modification of the usual main loop theorem proof and is accordingly omitted.

The next theorem is the reformulation of Problem 4.2.

THEOREM 4.7. *Given a system $M \in \mathcal{L}(\ell_2)$, M is uniformly robustly stable to $\bar{\mathbf{\Delta}}$ and $\sup_{\bar{\Delta}\in\bar{\mathbf{\Delta}}} \|T_{wz}\|_{sq} < 1$, where $T_{wz} = \left(I - M\bar{\Delta}\right)^{-1} M$ is the closed-loop map from w to z in Figure 4.5 if and only if $\|M\|_{sq} < 1$.*

Proof. First notice that the systems in Figures 4.5 and 4.6 define the same mapping from w to z, where

$$N_{11} = M, \quad N_{12} = \begin{bmatrix} M_{12} \\ M_{22} \end{bmatrix}, \quad N_{21} = \begin{bmatrix} M_{21} & M_{22} \end{bmatrix}, \quad N_{22} = M_{22}.$$

Thus M is uniformly robustly stable to $\bar{\mathbf{\Delta}}$ and $\sup_{\bar{\Delta}\in\bar{\mathbf{\Delta}}} \|T_{wz}\|_{sq} < 1$ if and only if the inequalities $\sup_{\bar{\Delta}\in\bar{\mathbf{\Delta}}} \|(I - N_{11}\bar{\Delta})^{-1}\|_{l_2\to l_2} < \infty$ and $\sup_{\bar{\Delta}\in\bar{\mathbf{\Delta}}} \left\| \mathcal{F}_u\left(N, \bar{\Delta}\right) \right\|_{sq} < 1$ hold.

Applying Proposition 4.6, this holds if and only if $\sup_{\Delta_p \in \mathbf{\Delta}_p} \|(I - N\Delta_p)^{-1}\|_{l_2 \to l_2} < \infty$. Defining the operators

$$S = \begin{bmatrix} I & 0 \\ 0 & I \\ 0 & I \end{bmatrix}$$

and

$$T = MS^*\Delta_p = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & I & I \end{bmatrix} \begin{bmatrix} \Delta_1 & \Delta_2 & 0 \\ \Delta_3 & 0 & 0 \\ 0 & 0 & \Delta_4 \end{bmatrix},$$

then we see that $I - ST = I - N\Delta_p$ is invertible if and only if $I - TS = I - M\Delta$ is invertible, where

$$\Delta = \begin{bmatrix} \Delta_1 & \Delta_2 \\ \Delta_3 & \Delta_4 \end{bmatrix} \in \mathbf{\Delta} = \{S^*\Delta_p S : \Delta_p \in \mathbf{\Delta}_p\},$$

and further,

$$\begin{aligned} (I - M\Delta)^{-1} &= I + T(I - ST)^{-1} S \\ &= I + MS^*\Delta_p (I - N\Delta_p)^{-1} S. \end{aligned}$$

Thus $\sup_{\Delta_p \in \mathbf{\Delta}_p} \|(I - N\Delta_p)^{-1}\|_{l_2 \to l_2} < \infty$ holds if and only if $\sup_{\Delta \in \mathbf{\Delta}} \|(I - M\Delta)^{-1}\|_{l_2 \to l_2} < \infty$. Equivalently, applying Theorem 4.4, $\|M\|_{sq} < 1$. □

To summarize, Problem 4.2 requires a controller K which provides a robust performance level of γ, that is, such that $\sup_{\bar{\Delta} \in \bar{\mathbf{\Delta}}} \|T_{wz}\|_{sq} < \gamma$. Applying Theorem 4.7, it is sufficient to find a synthesis for the square ℓ_2 problem $\|\mathcal{F}_\ell(G, K)\|_{sq} < \gamma$, and through the development of section 3 we have given a sufficient controller synthesis for this problem as a set of operator inequalities.

5. Numerical example. Returning to the generalized sensitivity minimization, we now give an example to numerically demonstrate our results; it is without physical significance. Consider the standard sensitivity minimization, obtained by taking $W_u = 0$. Set $W_y = W = I$, and let the plant P be SISO with the following realization in which B, C, and D are not time-varying:

$$A_1 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & 2 \\ -2 & 0 \end{bmatrix}, \quad A_3 = \begin{bmatrix} -2 & 0 \\ 0 & -2 \end{bmatrix}, \quad A_4 = \begin{bmatrix} 0 & -2 \\ 2 & 0 \end{bmatrix},$$

$$B = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 2 \end{bmatrix}, \quad D = 0.$$

Then the system G has the realization

$$G = \left[\begin{array}{c|cc} A & 0 & B \\ \hline WC & 0 & 0 \\ -W_yC & W_y & 0 \\ -C & I & 0 \end{array}\right] = \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & 0 \end{array}\right]$$

with

$$B_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad B_2 = B, \quad C_1 = \begin{bmatrix} WC \\ -W_y C \end{bmatrix}, \quad C_2 = -C,$$
$$D_{11} = \begin{bmatrix} 0 \\ W_y \end{bmatrix}, \quad D_{12} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad D_{21} = I.$$

Recall that the uncertainty block has the form $[\Delta \quad 0]$ so in this problem, $n_w = 1$ and $n_z = 2$. Treating the system as a finite horizon problem with a horizon of 4 gives $\gamma_{opt} = 1.0010$, and treating it as a periodic system with a period of 4 gives $\gamma_{opt} = 17.4473$.

6. Conclusions. In this paper we have extended work on LTI robust synthesis to LTV systems. The tools required were generalizations of results for LTI systems with spatial constraints [7, 8] to operators in $\mathcal{L}(\ell_2)$, the framework for LTV state space systems from [9], and the powerful results in [10, 21] on the S-procedure. We have demonstrated how a typical control problem (robust sensitivity minimization) may be handled using this machinery.

This theory has been experimentally applied to robust trajectory tracking on a flexible beam as reported in [19]. Software which performs square ℓ_2 synthesis for periodic and finite horizon systems is available at http://epic.me.uiuc.edu/∼dullerud.

Future work includes weakening the constraints of uniform robust stability to robust stability and further developing a general necessary condition for the controller synthesis. Computational issues arise when trying to solve the LMIs for a large horizon problem due to the size of the problem and the number of decision variables it involves. Advances in LMI solution methods and investigation into exploiting the particular structure of these LMIs would be beneficial.

The framework and proof machinery presented here provide a streamlined way to treat generalized ℓ_2-synthesis and thus may find wider application in generalizing LTI results to both the LTV and distributed control settings.

Appendix. Proof of Theorem 3.5. Here we will prove Theorem 3.5 using the technique of quadratic forms and constraints. We remark that the periodic case of this result could be obtained readily from the proofs in [7, 8] by first applying a lifting technique to the periodic system. To obtain the general LTV results proved here a different approach is required. In fact, the proof given, when restricted to the LTI case, gives a more direct and transparent demonstration than specialized LTI proofs in [7, 8]. Thus it may be useful for extending the latter LTI work to more general classes of systems (e.g., infinite dimensional, distributed control architectures).

As with the proof of Theorem 3.4, γ will be taken to be 1 as this causes no loss in generality. For simplicity we will restrict our proof to the case where either M is periodic or $n_w n_z \leq 2$ holds. See Remark A.4 at the end of this appendix for the routine extension to the $n_w n_z \leq 3$ case.

Thus for the remainder of the section we have the standing assumptions that $\|M\|_{sq} < 1$ and either (a) M is periodic or (b) $n_w n_z \leq 2$ holds.

Our constructions will be based on the following duality-based characterization of the square norm:

$$\|M\|_{sq} = \sup_{w,v \neq 0} \frac{\mathrm{Re}\langle v, Mw \rangle}{\|v\|_{\mathcal{Z}_\infty} \|w\|_{\mathcal{W}_\infty}}.$$

Having made this observation, we define the quadratic forms

$$\psi_i(w,v) = \left\langle \begin{bmatrix} v \\ w \end{bmatrix}, \begin{bmatrix} 0 & \frac{1}{2}M \\ \frac{1}{2}M^* & -P_i^*P_i \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} \right\rangle = \mathrm{Re}\langle v, Mw\rangle - \|P_i w\|^2 \text{ for } i = 1, \dots, n_w,$$

$$\psi_i(w,v) = \left\langle \begin{bmatrix} v \\ w \end{bmatrix}, \begin{bmatrix} -Q_i^*Q_i & \frac{1}{2}M \\ \frac{1}{2}M^* & 0 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} \right\rangle = \mathrm{Re}\langle v, Mw\rangle - \|Q_i v\|^2 \text{ for } i = n_w+1, \dots, n_w+n_z$$

on the space $\ell_2 \times \ell_2$. For convenience set $\Psi(w,v) = (\psi_1(w,v), \dots, \psi_{n_w+n_z}(w,v))$, and define the set

$$\nabla = \{\Psi(w,v) : \ w, v \in \ell_2 \text{ and } \|w\|^2 + \|v\|^2 = 1\} \subset \mathbb{R}^{n_w+n_z},$$

and the nonnegative orthant $\Pi = \{(a_1, \dots, a_{n_w+n_z}) : \text{each } a_i \geq 0\} \subset \mathbb{R}^{n_w+n_z}$.

LEMMA A.1. *The closure* $\bar{\nabla}$ *is strictly separated from* Π.

Proof. Choose $(w,v) \in \ell_2 \times \ell_2$ satisfying $\|w\|^2 + \|v\|^2 = 1$. It is sufficient to show that for some $1 \leq k_0 \leq n_w + n_z$ the inequality $\psi_{k_0}(w,v) \leq (\|M\|_{sq} - 1)(n_w + n_z)^{-1}$ holds.

Set $\alpha = \|M\|_{sq}$, and thus we have

$$\mathrm{Re}\langle v, Mw\rangle \leq \alpha \|v\|_{\mathcal{Z}_\infty} \|w\|_{\mathcal{W}_\infty}.$$

Now let $\gamma = \max\{\|v\|^2_{\mathcal{Z}_\infty}, \|w\|^2_{\mathcal{W}_\infty}\}$, and then conclude

$$\mathrm{Re}\langle v, Mw\rangle - \gamma \leq (\alpha - 1)\gamma.$$

By equivalence of norms we have that $\gamma \geq (n_w + n_z)^{-1}(\|w\|^2 + \|v\|^2) = (n_w + n_z)^{-1}$, and so

$$\mathrm{Re}\langle v, Mw\rangle - \gamma \leq (\alpha - 1)(n_w + n_z)^{-1}.$$

By definition of the $\mathcal{W}_\infty$- and $\mathcal{Z}_\infty$-norms, there exists $1 \leq k_0 \leq n_w + n_z$ such that the left-hand side above is equal to $\psi_{k_0}(w,v)$. ∎

LEMMA A.2. *The closure of the convex hull* $\mathrm{co}(\bar{\nabla})$ *is strictly separated from* Π.

Proof. We consider the cases (a) M is periodic and (b) $n_w \cdot n_z \leq 2$ separately.

The first case is where M is q-periodic, in which case we prove that the closure $\bar{\nabla} = \mathrm{co}(\bar{\nabla})$, and thus by Lemma A.1 the conclusion follows. The fact that $\bar{\nabla}$ is convex follows directly from the main result in [14] on shift invariant quadratic forms by simply replacing the shift operator with Z^q and will not be reproduced here; the basic facts required in the proof are that $\psi(v,w) = \psi(Z^{qk}v, Z^{qk}w)$ and that Z^{qk} tends weakly to zero as k tends to infinity. See [17] for a proof in the style of this appendix; see also [18].

When M is not periodic but $n_w \cdot n_z \leq 2$ holds, we have at most *three* quadratic forms ψ_k or, equivalently, $\nabla \subset \mathbb{R}^3$. In this case we can appeal to [10], which states exactly the claim of the lemma for any three quadratic forms on a complex inner product space. Note that this result is essentially equivalent to the well-known μ-theory result, which says that the structured singular value is equal to its upper bound for three full-block uncertainties; see [16]. ∎

LEMMA A.3. *There exist* 1*-admissible scales* T *and* S *such that*

$$\begin{bmatrix} -T & M \\ M^* & -S \end{bmatrix} < 0. \tag{A.1}$$

Proof. By the separating hyperplane theorem and Lemma A.2, there exist vectors $\sigma \in \mathbb{R}^{n_w}$, $\tau \in \mathbb{R}^{n_z}$ and scalars ϵ, β such that

$$\begin{bmatrix}\sigma\\ \tau\end{bmatrix}^* x < -\epsilon < \beta < \begin{bmatrix}\sigma\\ \tau\end{bmatrix}^* a \qquad \text{for all } x \in \mathrm{co}(\bar{\nabla}) \text{ and } a \in \Pi.$$

From the properties of Π it is routine to verify that $\sigma_i, \tau_i \geq 0$ and that $\epsilon > 0$. Now $\mathrm{co}(\bar{\nabla})$ is compact, and so without loss of generality we may assume the strict inequalities $\sigma_i, \tau_i > 0$ so that

$$\begin{bmatrix}\sigma\\ \tau\end{bmatrix}^* x < -\epsilon \text{ still holds for each } x \in \mathrm{co}(\bar{\nabla}).$$

Using the definition of ∇, we have, in particular, that

$$\begin{bmatrix}\sigma\\ \tau\end{bmatrix}^* \Psi(w,v) = \sum_{i=1}^{n_w} \sigma_i(\mathrm{Re}\langle v, Mw\rangle - \|P_i w\|^2) + \sum_{i=1}^{n_z} \tau_i(\mathrm{Re}\langle v, Mw\rangle - \|Q_i v\|^2) < -\epsilon$$

for all $(w,v) \in \ell_2 \times \ell_2$ satisfying $\|w\|^2 + \|v\|^2 = 1$. Defining $\gamma = \sum_{i=1}^{n_w} \sigma_i + \sum_{i=1}^{n_z} \tau_i$ and multiplying the inequality through by $2\gamma^{-1}$, we get

$$2\mathrm{Re}\langle v, Mw\rangle - 2\gamma^{-1} \sum_{i=1}^{n_w} \sigma_i \|P_i w\|^2 - 2\gamma^{-1} \sum_{i=1}^{n_z} \tau_i \|Q_i v\|^2 < -2\gamma^{-1}\epsilon.$$

By setting $S = 2\gamma^{-1} \sum_{i=1}^{n_w} \sigma_i P_i^* P_i$ and $T = 2\gamma^{-1} \sum_{i=1}^{n_z} \tau_i Q_i^* Q_i$ we can rewrite this as

$$\left\langle \begin{bmatrix}v\\ w\end{bmatrix}, \begin{bmatrix}-T & M\\ M^* & -S\end{bmatrix} \begin{bmatrix}v\\ w\end{bmatrix} \right\rangle < -2\gamma^{-1}\epsilon.$$

This holds for any $(w,v) \in \ell_2 \times \ell_2$ with the unit norm constraint $\|w\|^2 + \|v\|^2 = 1$; therefore, inequality (A.1) holds.

To complete the proof notice that $\sum_{i=1}^{n_w} s_i + \sum_{i=1}^{n_z} t_i = 2$, and we can therefore perturb $s_1 > 0$ slightly to get $\sum_{i=1}^{n_w} s_i + \sum_{i=1}^{n_z} t_i < 2$ while still maintaining inequality (A.1). ∎

Proof of Theorem 3.5. Invoke Lemma A.3, and apply the Schur complement formula to (A.1) to get

$$M^* T^{-1} M - S < 0,$$

which implies $S^{-\frac{1}{2}} M^* T^{-1} M S^{-\frac{1}{2}} - I < 0$. Clearly, this means $\|T^{-\frac{1}{2}} M S^{-\frac{1}{2}}\| < 1$, and from Lemma A.3 we have directly that T and S are 1-admissible. ∎

REMARK A.4. *So far we have proved only the theorem for the case of $n_w \cdot n_z \leq 2$. For the remaining case of $n_w \cdot n_z \leq 3$, the above argument can again be employed, using instead the quadratic forms*

$$\psi(w) = \|Mw\|^2 - \|P_i w\|^2 \text{ for } i = 1, \ldots, n_w \text{ when } n_z = 1,$$

and $\psi(v) = \|M^* v\|^2 - \|Q_i v\|^2$ for $i = 1, \ldots, n_z$ when $n_w = 1$.

REMARK A.5. *If we replace the space of $\ell_2(\mathbb{C}^n)$ in the formulation of this paper with the real space $\ell_2(\mathbb{R}^n)$, condition* (b) *in the statement of Theorem* 3.5 *must be weakened to to $n_w \cdot n_z \leq 2$. This is because the result of* [10] *used in the proof of Lemma* A.2 *holds only for quadratic forms on complex spaces; however, the work in* [21] *says that* $\mathrm{co}(\bar{\nabla})$ *is strictly separated from* Π *if there are at most two real forms. This will be the case when $n_w \cdot n_z \leq 2$ and the forms from Remark* A.4 *are used.*

REFERENCES

[1] J. A. Ball, I. Gohberg, and M. A. Kaashoek, *Nevanlinna-Pick interpolation for time-varying input-output maps: The discrete case*, in Oper. Theory Adv. Appl. 56, Birkhäuser, Basel, 1992, pp. 1–51.

[2] B. Bamieh, J. B. Pearson, B. A. Francis, and A. Tannenbaum, *A lifting technique for linear periodic systems with applications to sampled-data control*, Systems Control Lett., 17 (1991), pp. 79–88.

[3] B. Bollobas, *Linear Analysis*, Cambridge University Press, Cambridge, UK, 1990.

[4] T. Chen and L. Qiu, *H_∞ design of general multirate sampled-data control systems*, Automatica J. IFAC, 30 (1994), pp. 1139–1152.

[5] J. B. Conway, *A Course in Functional Analysis*, Grad. Texts in Math., Springer-Verlag, New York, 1990.

[6] R. D'Andrea, *Generalizations of H_∞ Control / Control of Rotating Stall*, Ph.D. thesis, Department of Electrical Engineering, California Institute of Technology, 1996.

[7] R. D'Andrea, *$\mathcal{H}^\infty$ optimization with spatial constraints*, in Proceedings of the IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, 1995, pp. 4327–4332.

[8] R. D'Andrea, *Generalized ℓ_2 synthesis*, IEEE Trans. Automat. Control, 44 (1999), pp. 1145–1156.

[9] G. E. Dullerud and S. Lall, *A new approach for analysis and synthesis of time-varying systems*, IEEE Trans. Automat. Control, 44 (1999), pp. 1486–1497.

[10] A. Fradkov and V. A. Yakubovich, *The S-procedure and duality theorems for nonconvex problems of quadratic programming*, Vestnik Leningrad Univ., 31 (1973), pp. 81–87 (in Russian); English translation in Vestnik Leningrad Univ. Math., 6 (1979), pp. 73–93.

[11] P. Gahinet and P. Apkarian, *A linear matrix inequality approach to $\mathcal{H}_\infty$ control*, Internat. J. Robust Nonlinear Control, 4 (1994), pp. 421–448.

[12] A. Halanay and V. Ionescu, *Time-Varying Discrete Linear Systems*, Birkhäuser, Basel, 1994.

[13] P. A. Iglesias, *An entropy formula for time-varying discrete-time control systems*, SIAM J. Control Optim., 34 (1996), pp. 1691–1706.

[14] A. Megretsky and S. Treil, *Power distribution inequalities in optimization and robustness of uncertain systems*, J. Math. Systems Estim. Control, 3 (1993), pp. 301–319.

[15] A. Packard, *Gain scheduling via linear fractional transformations*, Systems Control Lett., 22 (1994), pp. 79–92.

[16] A. Packard and J. C. Doyle, *The complex structured singular value*, Automatica J. IFAC, 29 (1993), pp. 71–109.

[17] F. Paganini and J. C. Doyle, *Analysis of implicitly defined systems*, in Proceedings of the IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, 1994, pp. 3673–3678.

[18] C. Pirie, *Controller Synthesis for Uncertain Time-Varying Discrete-Time Systems*, M. Math. thesis, Applied Mathematics, University of Waterloo, Waterloo, Ontario, Canada, 1999.

[19] C. Pirie, S. Okubo, G. Dullerud, and D. Tortorelli, *Robust Nonlinear Trajectory Tracking on a Flexible Beam*, manuscript in preparation.

[20] P. M. Young, *Robustness analysis for full structured uncertainties*, in Proceedings of the IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, 1996, pp. 3464–3469.

[21] V. A. Yakubovich, *S-procedure in nonlinear control theory*, Vestnik Leningrad Univ., (1971), pp. 62–77 (in Russian); English translation in Vestnik Leningrad Univ. Math., 4 (1977), pp. 73-93.

[22] G. Zames and J. G. Owen, *Duality theory for MIMO robust disturbance rejection*, IEEE Trans. Automat. Control, 38 (1993), pp. 743–752.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 5, pp. 1333–1357

ON AUTOMATON RECOGNIZABILITY OF ABNORMAL EXTREMALS*

UGO BOSCAIN† AND BENEDETTO PICCOLI‡

Abstract. For a generic single-input planar control system $\dot{x} = F(x) + uG(x)$, $x \in \mathbb{R}^2$, $u \in [-1, 1]$, $F(0) = 0$, we analyze the properties of abnormal extremals for the minimum time stabilization to the origin. We prove that abnormal extremals are finite concatenations of bang arcs with switchings occurring on the set in which the vector fields F and G are collinear. Moreover, all the generic singularities of one parametric family of extremal trajectories near to abnormal extremals are studied. In particular, we prove that all possible sequences of these singularities, and hence all generic abnormal extremals, can be classified by a set of words recognizable by an automaton.

Key words. optimal control, abnormal extremals, synthesis theory, generic planar systems

AMS subject classifications. 49K15, 49J15, 49N35

PII. S0363012900381650

1. Introduction.

In this paper we deal with the minimum time stabilization problem to the origin for the planar single-input system $\dot{x} = F(x) + uG(x)$, $x \in \mathbb{R}^2$, $u \in [-1, 1]$, $F(0) = 0$. The Pontryagin maximum principle (PMP) [12, 19] provides a necessary condition for optimality, and trajectories satisfying it are called extremals. Abnormal extremals are extremals corresponding to the zero level of the Hamiltonian given by the PMP; see [1, 2].

The aim of this paper is to analyze completely the generic properties of abnormal extremals. We prove that these are finite concatenations of bang arcs (that is, corresponding to constant ± 1 control). Moreover, the switchings (discontinuity points of the control) happen exactly when the abnormal extremal crosses the set of zeros of the function $\Delta_A = F \wedge G$. In many cases, an abnormal extremal is formed by fold points; that is, there are two sheets of the cotangent bundle covered by extremal trajectories that project onto the same region of the plane, and the boundary of this region is the projection of the abnormal extremal.

The set of possible singularities along abnormal extremals is formed of 28 (equivalence classes of) singular points, but not all sequences of singularities can be realized. We aim to provide a classification of all possible sequences of singularities. A good classification is obtained if one can put the possible sequences in bijective correspondence with some algebraic or combinatorial object Ω with simple structure. If all possible sequences of singularities were admitted, then this classification could be done choosing Ω to be the set of all words formed with letters from the alphabet $\{1, \dots, 28\}$, with the meaning that each number corresponds to a singularity. This is not the case; however, we can still have some regular structure. More precisely, Ω can be chosen as a set of words recognizable by an automaton, and this is the most natural classification for this problem; see rules R1, R2, and R3 of section 5. We recall that a set of words Ω from a given finite alphabet is recognizable if there exists an automaton

*Received by the editors November 27, 2000; accepted for publication (in revised form) July 27, 2001; published electronically January 9, 2002.
http://www.siam.org/journals/sicon/40-5/38165.html
†Université de Bourgogne, Département de Mathématiques, Analyse Appliquée et Optimisation, 47870-21078 Dijon, France (uboscain@u-bourgogne.fr).
‡DIIMA, Universitá di Salerno, Via Ponte Don Melillo, 84084 Fisciano (SA), Italy and SISSA–ISAS, Via Beirut 2-4, 34014 Trieste, Italy (piccoli@sissa.it).

that generates exactly the words of Ω. An automaton is, roughly speaking, a graph with labelled edges, and one constructs words considering all paths starting from a given set of points and ending to another fixed set of points. The recognizable sets of words share a regular structure used in theoretical computer science; see [10, 11]. Since the sequence of singularities completely describes the abnormal extremal, we obtain that abnormal extremals are recognizable by an automaton.

In [15, 16, 22, 23], the authors studied the properties of extremal trajectories and via a finite dimensional reduction, obtained using the PMP, proved the existence of a regular optimal synthesis. Roughly speaking, an optimal synthesis is a function associating to each point an optimal trajectory. It happens that the synthesis is indeed generated by a feedback that is smooth on each strata of a Whitney stratification of the plane. For synthesis theory we refer also to [3, 9, 18]. In a series of papers [7, 8, 16, 17], the existence of a structurally stable optimal synthesis for generic smooth systems was proved, and complete classifications of the corresponding (nonsmooth) flows and relative singularities were given, in the same spirit of the work of Peixoto for two dimensional dynamical systems. For results in three dimensions, see [20, 21].

An alternative approach is used here and in [4, 5, 13, 14]. This amounts to constructing all extremals, that is, trajectory-covector pairs satisfying the PMP, and projecting the obtained set onto the plane. This approach is more involved but sheds more light on the links between synthesis singularities and the singularities of the minimum time function; see [6]. Moreover, the supports of extremals form a Whitney stratified set of dimension three in the cotangent bundle (see [5]). After normalization of the covector, one obtains a two dimensional stratified set in $\mathbb{R}^2 \times S^1$. The projection singularities can be classified in topological sense. Beside the classical folds and cusps (see [24]), new singularities appear. Some, called vertical, are due to the fact that the target (the origin) is of codimension two, while others are stable and independent of the target properties. These new singularities are called bifold and ribbon. In particular, the ribbon singularity can appear only along abnormal extremals.

All possible generic singularities of the synthesis on the plane occurring along (projections of) abnormal extremals are classified in section 4. However, one has to prove that all singularities indeed appear for some generic system, in particular, those corresponding to the new projection singularities, namely, bifold and ribbon. As a byproduct, we obtain the existence of systems presenting singular points corresponding to projection singularities of bifold and ribbon type (see Theorem 35).

The paper is organized in the following way. Section 2 introduces basic definitions and main results. The third section is dedicated to formulating and proving the main propositions about the switching strategy of abnormal extremals. In section 4 we describe the synthesis singularities (on the plane) occurring along abnormal extremals. Finally, in section 5 we describe the classification of abnormal extremals via a recognizable set of words and the corresponding automaton. Moreover, in section 5 we show that the ribbon singularity is realized, and we give an example of a synthesis containing a singularity of this kind.

2. Basic definitions and statement of the main result. Let Ξ be the set of all couples of $\mathcal{C}^\infty$ vector fields (F, G) such that the origin is an equilibrium point for F, that is, $F(0) = 0$. From now on we endow Ξ with the $\mathcal{C}^3$ topology induced by the norm

$$\|(F,G)\| = \sup \left\{ \left| \frac{\partial^{\alpha_1+\alpha_2} F_i(x)}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2}} \right|, \left| \frac{\partial^{\alpha_1+\alpha_2} G_i(x)}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2}} \right| :
$$

$$x \in \mathbb{R}^2;\ i = 1, 2;\ \alpha_1, \alpha_2 \in \mathbf{N} \cup \{0\}; \alpha_1 + \alpha_2 \leq 3 \Big\}.$$

We say that a subset of Ξ is generic if it contains an open and dense set. Analogously, a property P is said to be generic if the set satisfying P is generic.

For every $(F, G) \in \Xi$ we consider the minimum time stabilization problem to the origin for the control system

$$\dot{x} = F(x) + uG(x), \qquad x \in \mathbb{R}^2, \quad u \in [-1, 1]. \tag{1}$$

Reversing time, we deal with the equivalent problem of reaching every point of the plane in minimum time from the origin.

Given a measurable function $u : [a, b] \to [-1, 1]$, a trajectory of (1) corresponding to u is an absolutely continuous map $\gamma : [a, b] \to \mathbb{R}^2$ such that $\dot{\gamma}(t) = F(\gamma(t)) + u(t)G(\gamma(t))$ for almost every $t \in [a, b]$. Since the system is autonomous, we can always assume that $Dom(\gamma) = [0, a]$ for some $a \in \mathbb{R}$, $a > 0$, where Dom denotes the domain. Moreover, we denote by $\mathrm{Supp}(\gamma)$ the set $\gamma([a, b])$. A trajectory $\gamma : [0, a] \to \mathbb{R}^2$ is (time) optimal if for every trajectory $\gamma' : [0, b] \to \mathbb{R}^2$ with $\gamma(a) = \gamma'(b)$ we have $a \leq b$. A trajectory γ corresponding to a constant control ± 1 is called a *bang arc*. A *bang-bang* trajectory is a finite concatenation of bang arcs, and a time where the control changes sign is called *switching time*.

The well-known PMP in this special case states the following. Define for every $(x, p, u) \in \mathbb{R}^2 \times (\mathbb{R}^2)_* \times [-1, 1]$, where $(\mathbb{R}^2)_*$ is the set of row vectors,

$$\mathcal{H}(x, p, u) = p \cdot F(x) + u\, p \cdot G(x) \tag{2}$$

and

$$H(x, p) = \max\{p \cdot F(x) + u\, p \cdot G(x) : u \in [-1, 1]\}. \tag{3}$$

If $\gamma : [0, a] \to \mathbb{R}^2$ is a (time) optimal trajectory corresponding to a control $u : [0, a] \to [-1, 1]$, then there exist a nontrivial *field of covectors along* γ, that is, a function $\lambda : [0, a] \to (\mathbb{R}^2)_*$ never vanishing, and a constant $\lambda_0 \leq 0$ such that

(i) $\dot{\lambda}(t) = -\lambda(t) \cdot (\nabla F + u(t)\nabla G)(\gamma(t))$,
(ii) $\mathcal{H}(\gamma(t), \lambda(t), u(t)) + \lambda_0 = 0$ for almost every $t \in Dom(\gamma)$,
(iii) $\mathcal{H}(\gamma(t), \lambda(t), u(t)) = H(\gamma(t), \lambda(t))$ for almost every $t \in Dom(\gamma)$.

In this case we say that the pair (γ, λ) is extremal. If γ is optimal, we say that the pair (γ, λ) is optimal.

Given an extremal pair (γ, λ), one easily checks that for every $\alpha \in \mathbb{R}$, $\alpha > 0$, the pair $(\gamma, \alpha\lambda)$ is also extremal.

Definition 1. *Let (γ, λ) be an extremal pair. If the corresponding Hamiltonian satisfies $\mathcal{H}(\gamma(t), \lambda(t), u(t)) = 0$ for almost every $t \in Dom(\gamma)$ (i.e., if λ_0 in* (i) *above vanishes), we say that (γ, λ) is an abnormal extremal.*

The following theorem, proved in section 5, describes the set of generic abnormal extremals.

Theorem 2. *The set of abnormal extremals for control systems* (1)*, in a generic set of Ξ, can be classified through a set of words recognizable by an automaton.*

In the following we introduce some notation also used in [7, 16, 17]. Set

$$Y = F + G, \qquad X = F - G.$$

Let $\gamma : [t_1, t_2] \to \mathbb{R}^2$ be a trajectory of (1). If γ corresponds to constant control $+1$ (resp., -1) on $[t_1, t_2]$, we say that $\gamma|_{[t_1,t_2]}$ is a Y-trajectory (resp., X-trajectory), and we write $\gamma|_{[t_1,t_2]} \in \text{Traj}(Y)$ (resp., $\gamma|_{[t_1,t_2]} \in \text{Traj}(X)$). We say that $\gamma|_{[t_1,t_2]}$ is a Z-trajectory, and we write $\gamma|_{[t_1,t_2]} \in \text{Traj}(Z)$ if γ on $[t_1, t_2]$ corresponds to the control (called singular)

$$\varphi(x) = -\frac{\nabla \Delta_B(x) \cdot F(x)}{\nabla \Delta_B(x) \cdot G(x)}, \tag{4}$$

where

$$\Delta_B(x) = G(x) \wedge [F, G](x) = G_1(x)[F, G]_2(x) - G_2(x)[F, G]_1(x),$$

and $[F, G]$ is the Lie bracket of F and G. For later use we also define the function

$$\Delta_A(x) = F(x) \wedge G(x) = F_1(x)G_2(x) - F_2(x)G_1(x).$$

DEFINITION 3. *If $\gamma_1 : [t_1, t_2] \to \mathbb{R}^2$, $\gamma_2 : [t_2, t_3] \to \mathbb{R}^2$ $(t_1 < t_2 < t_3)$ are two trajectories of* (1) *with $\gamma_1(t_2) = \gamma_2(t_2)$, we set $(\gamma_2 * \gamma_1)(t) := \gamma_1(t)$ if $t \in [t_1, t_2[$ and $(\gamma_2 * \gamma_1)(t) := \gamma_2(t)$ if $t \in [t_2, t_3]$. Given an extremal trajectory γ, we denote by $n(\gamma)$ the smallest integer such that there exist $\gamma_i \in$ Traj(X)$\cup$Traj(Y) $\cup$Traj(Z), $i = 1, \dots, n(\gamma)$, satisfying $\gamma = \gamma_{n(\gamma)} * \cdots * \gamma_1$. The function $n(\gamma)$ is called the* number of arcs *of γ.*

In the following, we assume the following generic conditions:

(P1) The vectors $G(0)$ and $[F, G](0)$ are linearly independent.

(P2) Zero is a regular value for Δ_A and Δ_B.

(P3) The set $\Delta_A^{-1}(0) \cap \Delta_B^{-1}(0)$ is locally finite.

Let Tan_A be the set of points $x \in \Delta_A^{-1}(0)$ such that $X(x)$ or $Y(x)$ is tangent to $\Delta_A^{-1}(0)$. Define Tan_B in the same way using Δ_B rather than Δ_A.

(P4) Tan_A and Tan_B are locally finite sets.

Let $Bad := (\Delta_A^{-1}(0) \cap \Delta_B^{-1}(0)) \cup \text{Tan}_A \cup \text{Tan}_B$.

(P5) Bad is locally finite.

Notice that (P5) is a consequence of (P3) and (P4).

(P6) If $x \in Bad$ and $G(x) = 0$, then $F(x) \cdot \nabla(\Delta_A)(x) \neq 0$ and $F(x) \cdot \nabla(\Delta_B)(x) \neq 0$.

(P7) If $x \in Bad$, $G(x) \neq 0$, $x \in (\Delta_A^{-1}(0) \cap \Delta_B^{-1}(0)) \cap \text{Tan}_A$, and say, $X(x) \cdot \nabla \Delta_A(x) = 0$, then $\partial_y (X \cdot \nabla \Delta_A)|_{y=x} \neq 0$, $X(y) \neq 0$, and $Y(y) \neq 0$, where y takes values on $\Delta_A^{-1}(0)$ in a neighborhood of x.

(P8) If $x \in Bad$ and $x \in \text{Tan}_B$, then $\Delta_A(x) \neq 0$.

(P9) If $x \in Bad$, $G(x) \neq 0$, $X(x) = 0$, or $Y(x) = 0$, then $\Delta_B(x) \neq 0$.

The generic conditions (P6), (P7), and (P8) are clarified in Figure 2.1.

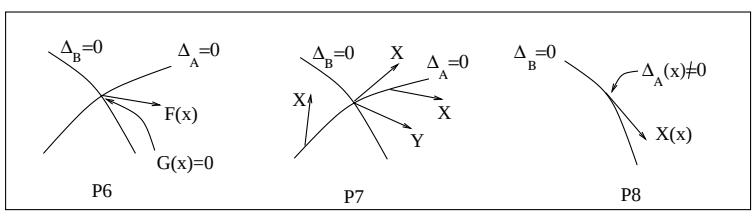


FIG. 2.1.

In [16] the following proposition was proved.

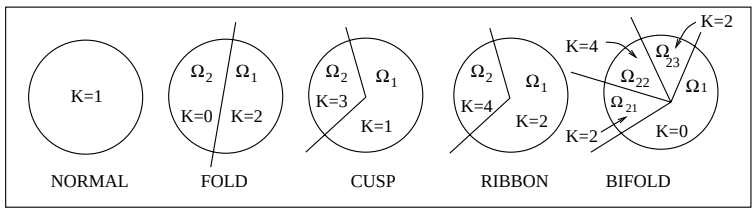


FIG. 2.2.

PROPOSITION 4. *If* (P1)–(P9) *hold, then each extremal trajectory* γ *can always be written in the form* $\gamma = \gamma_{n(\gamma)} * \cdots * \gamma_1$ *with* $n(\gamma) < \infty$*, and each arc is bang or singular (that is, a* Z*–trajectory).*

Let $\tau > 0$. We call the *reachable set* within time τ the set $\mathcal{R}(\tau) := \{x \in \mathbb{R}^2 :$ there exists $t \in [0,\tau]$ and a trajectory $\gamma : [0,\tau] \to \mathbb{R}^2$ of (1) such that $\gamma(0) = 0$, $\gamma(t) = x\}$. In [16] the following lemma was shown.

LEMMA 5. *Under the generic conditions* (P1)–(P9) $\gamma_i \in$ *Traj(Z) iff* $\mathrm{Supp}(\gamma_i) \subset \Delta_B^{-1}(0)$.

In the case in which $\gamma_i \in \mathrm{Traj}(Z)$ we say that $\mathrm{Supp}(\gamma_i)$ is a *turnpike*. We call $\gamma^{\pm} : [0, t_f^{\pm}] \to \mathbb{R}^2$ the extremal trajectories originating at 0 and corresponding to constant control ± 1 with $t_f^{\pm}$ the last times in which $\gamma^{\pm}$ are extremal (if they are less than τ) or τ (otherwise). Under generic assumptions, every extremal trajectory exits the origin with constant control $+1$ or -1.

DEFINITION 6. *Let* $(\gamma, \lambda) : [0,\tau] \to \mathbb{R}^2$ *be an extremal pair. The corresponding switching function is defined as* $\phi(t) := \lambda(t) \cdot G(\gamma(t))$.

We immediately have the following lemma.

LEMMA 7. *Let* $(\gamma, \lambda) : [0,\tau] \to \mathbb{R}^2$ *be an extremal pair, and* $[t_1, t_2] \subseteq [0,\tau]$ $(t_1 < t_2)$*. Then*

- *on* $[t_1, t_2]$, γ *corresponds to constant control* $+1$ *(resp.,* -1*) iff for each* $t \in [t_1, t_2]$ *we have* $\phi(t) \geq 0$ *(resp.,* $\phi(t) \leq 0$*) and meas*$(\{t \in [t_1, t_2] : \phi(t) = 0\}) = 0$*;*
- *on* $[t_1, t_2]$, γ *corresponds to the singular control* φ *(defined in* (4)*) iff for each* $t \in [t_1, t_2]$ *we have* $\phi(t) = 0$.

In [22] the following lemma was proved.

LEMMA 8. *Let* γ *be an extremal trajectory and* $\bar{t} \in Dom(\gamma)$ *a switching time. Let* $\bar{x} = \gamma(\bar{t})$*, and suppose that* $\bar{x} \notin \Delta_A^{-1}(0) \cup \Delta_B^{-1}(0)$*. Then* $\bar{t}$ *is a switching time from* X *to* Y *iff* $-\Delta_A(\bar{x})/\Delta_B(\bar{x}) > 0$*, and* $\bar{t}$ *is a switching time from* Y *to* X *iff* $-\Delta_A(\bar{x})/\Delta_B(\bar{x}) < 0$.

We now introduce some definitions to describe the projection singularities we encounter along abnormal extremals. Fix $\bar{x} \in \mathcal{R}(\tau)$ and an extremal trajectory $\bar{\gamma} : [0, \bar{a}] \to \mathcal{R}(\tau)$ such that $\bar{\gamma}(0) = 0$, $\bar{\gamma}(\bar{a}) = \bar{x}$, and define the function

$$K_\varepsilon^{\bar{x},\bar{\gamma}}(x) := \sharp\{\text{extremal trajectory } \gamma : \ \gamma(0) = 0,\ \gamma(a) = x,\ |\gamma(t) - \bar{\gamma}(t)| < \varepsilon \quad \forall\, t \in [0, \min(a,\bar{a})],\ |a - \bar{a}| < \varepsilon\}, \tag{5}$$

where $\sharp$ denotes the cardinality of a set. Referring to Figure 2.2, we introduce the following key definition.

DEFINITION 9. *Fix* $\bar{x} \in \mathcal{R}(\tau)$ *and an extremal trajectory* $\bar{\gamma} : [0, \bar{a}] \to \mathcal{R}(\tau)$ *such that* $\bar{\gamma}(0) = 0$ *and* $\bar{\gamma}(\bar{a}) = \bar{x}$.

- *We say that $\bar{x}$ is a* normal *point along $\bar{\gamma}$ if for ε sufficiently small there exists a neighborhood U of $\bar{x}$ such that $K_\varepsilon^{\bar{x},\bar{\gamma}}(U) = 1$.*
- *We say that $\bar{x}$ is a* fold *point along $\bar{\gamma}$ if there exists a one dimensional piecewise–$\mathcal{C}^1$ manifold l, with $\bar{x} \in l$, satisfying the following. For ε sufficiently small there exists a neighborhood U of $\bar{x}$ divided by l into two connected components Ω_1, Ω_2 such that $K_\varepsilon^{\bar{x},\bar{\gamma}}(\Omega_1) = 2$, $K_\varepsilon^{\bar{x},\bar{\gamma}}(\Omega_2) = 0$, $K_\varepsilon^{\bar{x},\bar{\gamma}}(l) = 1$.*
- *If l, $U(\varepsilon)$, Ω_1, Ω_2 are as above, then if $K_\varepsilon^{\bar{x},\bar{\gamma}}(\Omega_1) = 1$, $K_\varepsilon^{\bar{x},\bar{\gamma}}(\Omega_2) = 3$, and $K_\varepsilon^{\bar{x},\bar{\gamma}}(l) = 2$, we say that $\bar{x}$ is a* cusp point *along $\bar{\gamma}$.*
- *If l, $U(\varepsilon)$, Ω_1, Ω_2 are as above, then if $K_\varepsilon^{\bar{x},\bar{\gamma}}(\Omega_1) = 2$, $K_\varepsilon^{\bar{x},\bar{\gamma}}(\Omega_2) = 4$, and $K_\varepsilon^{\bar{x},\bar{\gamma}}(l) = 3$, we say that $\bar{x}$ is a* ribbon point *along $\bar{\gamma}$.*
- *We say that $\bar{x}$ is a* bifold point *along $\bar{\gamma}$ if there exists a one dimensional connected piecewise $\mathcal{C}^1$ embedded manifold l and two connected $\mathcal{C}^1$ embedded manifolds l_1 and l_2 satisfying the following.*
  1. *$l \cap \partial l_i = \{\bar{x}\}$ $(i = 1, 2)$; $\partial l_1 \cap \partial l_2 = \{\bar{x}\}$.*
  2. *For ε sufficiently small there exists a neighborhood U of x satisfying the following. The set $U \setminus l$ has two connected components Ω_1, Ω_2, and the set $\Omega_2 \setminus \{l_1 \cup l_2\}$ has three connected components Ω_{21}, Ω_{22}, and Ω_{23} (the names are chosen as in Figure 2.2) such that $K_\varepsilon^{\bar{x},\bar{\gamma}}(\Omega_1) = 0$, $K_\varepsilon^{\bar{x},\bar{\gamma}}(\Omega_{21}) = 2$, $K_\varepsilon^{\bar{x},\bar{\gamma}}(\Omega_{22}) = 4$, $K_\varepsilon^{\bar{x},\bar{\gamma}}(\Omega_{23}) = 2$, $K_\varepsilon^{\bar{x},\bar{\gamma}}(l) = 1$, and $K_\varepsilon^{\bar{x},\bar{\gamma}}(l_i) = 3$ $(i = 1, 2)$.*

Notice that this definition describes projection singularities. Indeed, the trajectories reaching the same point in the plane correspond to different lifts in the cotangent bundle.

Our aim is to prove that these are the only projection singularities that may happen along abnormal extremals and show that they are in fact realized for some control systems.

The behavior of abnormal extremals is individuated by the singularities it meets. To describe all possible singularities, we start describing those encountered in the optimal case. In [7, 16, 17], dealing with the optimal synthesis, it was proved that $\mathcal{R}(\tau)$ is a stratified subset of $\mathbb{R}^2$, and the one and zero dimensional strata that are singularities of the optimal flow are called, respectively, *frame curves* and *frame points* (FCs and FPs). Moreover, the authors use the letters X, Y, C, S, and K to indicate, respectively, the FCs corresponding to subsets of $\mathrm{Supp}(\gamma^+)$, subsets of $\mathrm{Supp}(\gamma^-)$, curves made of switching points (switching curves, see Figure 2.3 case 1), turnpikes and overlaps, that is, curves made of points reached optimally by two distinct trajectories. In [17], it was proved that these are generically all possible FCs. An FP x that is the intersection of two FCs F_1 and F_2 is called an (F_1, F_2) FP. Considering the extremal trajectories instead of the optimal ones, we do not have any K FC, but we have the following new FCs:

- an FC called $\bar{C}$ made of switching points on which X and Y point to opposite sides (see Figure 2.3, case 2);
- an FC called W that is an arc of an extremal trajectory characterized by the fact that all its points are fold points (see Figure 2.3, case 3).
- an FC called γ_0 that is an arc of an extremal trajectory that "transports" some special information, e.g., it switches every time it meets the locus $\Delta_A^{-1}(0)$, or it evolves into a W FC.

More details on the FCs W and γ^0 are given below.

To understand in detail the structure of abnormal extremals, we describe the set of all extremals, that is organized in one parameter families, called strips. Let Γ be

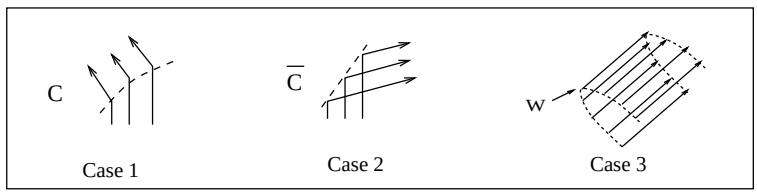


FIG. 2.3. *The FCs C, $\bar{C}$, W.*

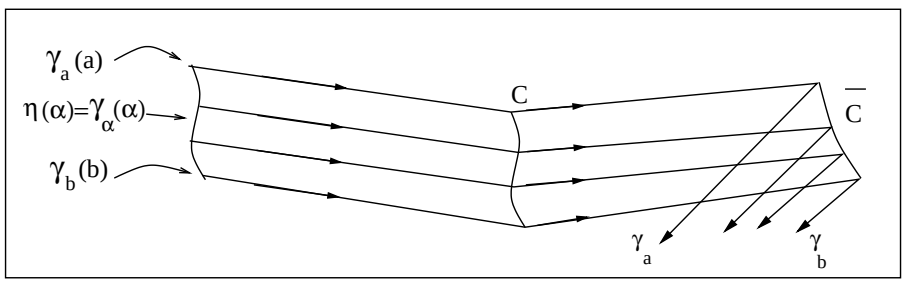


FIG. 2.4. *Strip.*

the set of all the extremal trajectories for the control problem (1) up to time $\tau > 0$. Again, under generic assumptions, extremals are finite concatenations of bang and singular arcs, so Γ can be obtained by a constructive algorithm that makes inductive steps on the number of arcs of extremal trajectories. We outline such an algorithm, described in detail in [5].

Outline of the algorithm. First, one constructs all the extremal trajectories in a neighborhood of $\gamma^{\pm}$. Then the set of extremal trajectories is subdivided into "strips." Each strip is a one dimensional continuous parametric family of extremal bang-bang trajectories having the same switching strategy. The evolution of each strip must be studied. In order to do that, the evolution of the boundary of strips must be analyzed separately. The more delicate case is that of abnormal extremals that here we treat in detail. The evolution of the interior of a strip may create new strips and cause the subdivision of each strip into smaller strips. This case is treated in [5], where the whole construction is completed.

Let us give the precise definition of strip.

DEFINITION 10. *Let a, b be two real numbers s.t. $0 \le a < b \le \tau$ and $x \in \mathcal{R}(\tau)$. A set of trajectories $\mathcal{S}^{a,b,x} = \{\gamma_\alpha : \ \alpha \in [a,b],\ \gamma_a(a) = x\}$ is called a strip if*

(i) $\forall\ \alpha \in [a,b],\ \ \gamma_\alpha : [0, \tau(\alpha)] \to \mathbb{R}^2$, *$\tau(\alpha) > \alpha$ is an extremal trajectory for the control problem* (1). *Moreover, there exists $\varepsilon > 0$ s.t. $\gamma_{[\alpha,\alpha+\varepsilon]}$ corresponds to a constant control ± 1;*

(ii) $\forall\ \alpha \in]a,b[,\ \ \gamma_\alpha$ *does not switch on $\Delta_A^{-1}(0) \cup \Delta_B^{-1}(0)$ after time α;*

(iii) *the set $\mathcal{B}^{a,b,x} = \{y \in \mathcal{R}(\tau) : \ \exists\ \alpha \in\ ]a,b[$ and $t \in]\alpha, \tau(\alpha)[$ s.t. $y = \gamma_\alpha(t)$, t is a switching time for $\gamma_\alpha\}$ is never tangent to X or Y.*

(iv) *the map $\eta : \ \alpha \in\ [a,b] \mapsto \gamma_\alpha(\alpha) \in \mathbb{R}^2$ is a bang or singular arc and $\gamma_\alpha|_{[0,\alpha']} = \gamma_{\alpha'}|_{[0,\alpha']}$ if $a \le \alpha' \le \alpha \le b$.*

The function $\eta : [a,b] \to \mathbb{R}^2$ is called the *base* of the strip, $\overset{\circ}{\mathcal{S}}{}^{a,b,x} := \{\gamma_\alpha : \ \alpha \in]a,b[\}$ is called an open strip, and $\partial\mathcal{S}^{a,b,x} := \{\gamma_a,\ \gamma_b\}$ is called the *strip border.* See Figure 2.4 for a graphical description of a strip.

Notice that in [7, 16] a similar algorithm was used to construct the optimal synthesis that can be organized as a finite collection of optimal strips. Also in this case the algorithm constructs Γ as a finite union of strips. The set of borders of strips is called $\partial\Gamma$. We clearly have the following lemma.

LEMMA 11. *Under generic assumptions, $\partial\Gamma$ is a finite set.*

Moreover, by construction, every border belongs to two different strips, and adjacent strips correspond locally to the same control. More precisely, the following lemma holds.

LEMMA 12. *Let $\gamma \in \partial\Gamma$, let $\mathcal{S}^1$ and $\mathcal{S}^2$ be the two strips such that $\{\gamma\} = \mathcal{S}^1 \cap \mathcal{S}^2$, and let $\bar{t}$ be a switching time for γ. Then the switching loci of $\mathcal{S}^1$ and $\mathcal{S}^2$, passing through $\gamma(\bar{t})$, correspond both to switchings from X to Y or both to switchings from Y to X.*

The algorithm produces the FCs in the following way:

- the FCs of kinds C and $\bar{C}$ lie in the interior of the strips;
- the borders of strips contain FCs of kinds W, γ^0, X, Y;
- the bases of strips contain FCs of kinds X, Y, S.

In fact, the FCs of kinds γ^0 and W are borders of strips, i.e., as follows.

DEFINITION 13. *Let $\gamma \in \partial\Gamma$, and suppose that it corresponds to a constant control in the interval $]b, c[$ $(0 < b < c \leq \tau)$.*

- *We say that in $]b, c[$ γ is a strip border of kind W if, for every $t \in\ ]b, c[$, $x := \gamma(t)$ is a fold point.*
- *Vice versa if, for every $t \in\ ]b, c[$ $\gamma(t)$ is a normal point, we say that in $]b, c[$ γ is a strip border of kind γ_0.*

By construction, we have the following lemma.

LEMMA 14. *If $\gamma \in \partial\Gamma$ and γ is of kind W in $]a, b[$ and of kind γ_0 in $]b, c[$ $(0 \leq a < b < c \leq \tau)$ or vice versa, then b is a switching time for γ.*

In the next section we see that in the case of abnormal extremals we need a more precise definition for FCs of kind W.

Given an extremal trajectory γ, let us define $v^\gamma(v_0, t_0; t)$ to be the solution to the Cauchy problem

$$\begin{cases} \dot{v^\gamma}(v_0, t_0; t) &= (\nabla F + u(t)\nabla G)(\gamma(t)) \cdot v^\gamma(v_0, t_0; t), \\ v^\gamma(v_0, t_0; t_0) &= v_0, \end{cases} \tag{6}$$

where $u(t)$ is the control corresponding to γ. In [16], denoting $\bar{v}^\gamma(t) := v^\gamma(G(\gamma(t)), t; 0)$, the following function was defined:

$$\theta^\gamma :\ Dom(\gamma) \to\ [-\pi, \pi], \quad \theta^\gamma(t) := \arg(\bar{v}^\gamma(0), \bar{v}^\gamma(t)), \tag{7}$$

where arg is the angle measured counterclockwise. We have the following.

LEMMA 15. *For fixed t_0, t, the map $f_{t_0,t} : v_0 \mapsto v^\gamma(v_0, t_0; t)$ is linear and injective. Moreover, let $\gamma : [a, b] \to \mathbb{R}^2$ be an extremal trajectory of (1) corresponding to a constant control $\bar{u}$. Then for every $t, t_0 \in [a, b]$ it holds that $v^\gamma((F+\bar{u}G)(\gamma(t_0)), t_0; t) = (F + \bar{u}G)(\gamma(t))$.*

In the following we use the notation $\theta^\pm(t) := \theta^{\gamma^\pm}(t)$ and $v^\pm := v^{\gamma^\pm}$. From Lemma 15 we have, for every $t \in [0, t_f^\pm]$, $v^\pm((F \pm G)(\gamma^\pm(t)), t; 0) = (F \pm G)(0) = \pm G(0)$. We consider the following generic conditions:

(GC1) For every $t \in [0, t_f^+]$, $G(\gamma^+(t)) \neq 0$.

(GC2) $\dot{\theta}^+(0) \neq 0$, $\dot{\theta}^+(t_f^+) \neq 0$.

(GC3) If $\dot{\theta}^+(t) = 0$, then $\theta^+(t) \neq 0$, $\ddot{\theta}^+(t) \neq 0$.

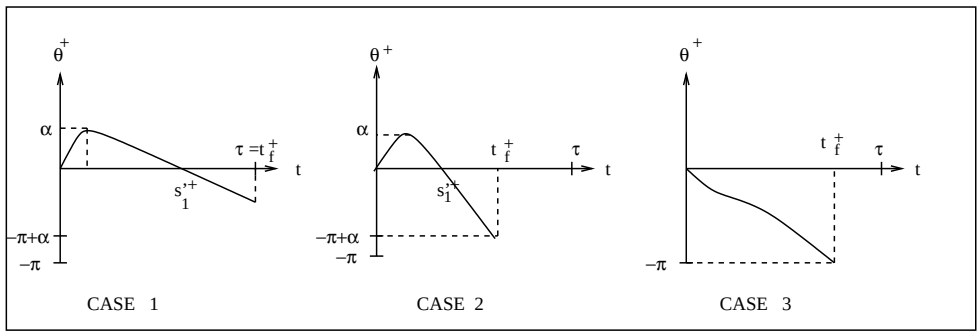


FIG. 2.5.

(GC4) If $t \neq s$ and $\dot\theta^+(s) = \dot\theta^+(t) = 0$, then $\theta^+(s) \neq \theta^+(t)$.
(GC5) If $t_f^+ = \tau$, then $\max\{|\theta^+(t) - \theta^+(\tau)|, \quad t \in [0,\tau]\} < \pi$.
In [7, 16] the following lemma was shown.

LEMMA 16. *We have the following:*

(A) $t_f^+ = \begin{cases} \tau \text{ if } |\theta^+(a) - \theta^+(b)| < \pi \quad \forall \quad a,\ b \in [0,\tau], \\ \min\{t \in [0,\tau] : \quad |\theta^+(a) - \theta^+(b)| = \pi \text{ for some } a,\ b \in [0,t]\} \text{ otherwise}; \end{cases}$

(B) *let γ be an extremal trajectory; then for almost every $t \in Dom(\gamma)$ we have $sgn(\dot\theta^\gamma(t)) = sgn(\Delta_B(\gamma(t))$;*
(C) *let γ be an extremal trajectory; then for each $t \in Dom(\gamma)$ the conditions $\phi(t) = \lambda(t) \cdot G(\gamma(t)) = 0, \quad \dot\theta^\gamma(t) \neq 0$ imply that t is a switching time for γ.*

We single out the following times:

$$s_1^+ := \min\{t \in]0, t_f^+] : \quad \theta^+(t) = 0, \dot\theta^+(0) < 0\},$$
$$s_1'^+ := \min\{t \in]0, t_f^+] : \quad \theta^+(t) = 0, \dot\theta^+(0) > 0\}.$$

The times s_1^-, $s_1'^-$ were defined similarly. From now on we assume the generic condition

(GAτ) (i) $\tau \notin \{s_i^\pm, s_i'^\pm\}$.
(ii) Let (γ, λ) be an extremal pairs. Then $\phi(\tau) = \lambda(\tau) \cdot G(\gamma(\tau)) = 0$ implies $\gamma(\tau) \notin \Delta_A^{-1}(0)$.

Observation 1. We can have three situations (cfr. Figure 2.5).
(1) $|\theta^\pm(a) - \theta^\pm(b)| < \pi$ for every $a, b \in [0,\tau]$. In this case $t_f^\pm = \tau$ and $|\theta^\pm(t_f^\pm)| < \pi$.
(2) $|\theta^\pm(a) - \theta^\pm(b)| = \pi$ for some $a, b \in]0,\tau]$. In this case, $|\theta^\pm(t_f^\pm)| < \pi$, $\theta^\pm$ has a maximum or a minimum in $]0, t_f^\pm[$, and either $s_1^\pm \neq 0$ or $s_1'^\pm \neq 0$.
(3) $|\theta^\pm(t_f^\pm)| = \pi$. In this case, $s_1^\pm$ and $s_1'^\pm$ are not defined, and generically we get $t_f^\pm < \tau$.

3. Generic properties of abnormal extremals. In this section we state and prove the main results about the switching strategies of abnormal extremals.

DEFINITION 17. *Let $\gamma : [0,\tau] \to \mathbb{R}^2$ be (the first component of) an abnormal extremal for the control problem* (1) *such that it switches at least one time, and let t_1 be its first switching time. We refer to the couples (γ, t_1) as nontrivial abnormal extremal (NTAE). By definition an NTAE is maximal if defined on $[0,\tau]$.*

DEFINITION 18. *Let $\gamma : [0,\tau] \to \mathbb{R}^2$ be an NTAE, and let $t_1 < t_2 < \cdots < t_{n(\gamma)-1} < t_{n(\gamma)} := \tau$ be the sequence of switching times. We set $AA(i) = \mathrm{Supp}\left(\gamma|_{[t_i,t_{i+1}]}\right)$ $(i = 1,\cdots,n(\gamma)-1)$, and we call it an* abnormal arc.

PROPOSITION 19. *Let $\gamma : [0,\tau] \to \mathbb{R}^2$ be an extremal trajectory for the control problem* (1) *such that it switches at least one time, let $\lambda : [0,\tau] \to \mathbb{R}_2$ be the corresponding covector, and let $t_1 < t_2 < \cdots < t_{n(\gamma)-1} < t_{n(\gamma)} := \tau$ be the sequence of switching times. Then the following hold:*

(A) *$\lambda(.)$ is unique (up to the multiplication by a positive constant);*

(B) *under generic assumptions the following conditions are equivalent:*

(a) *(γ, t_1) is an NTAE;*

(b) *$\gamma(t_i) \in \Delta_A^{-1}(0)$ for some $i \in \{1,\dots,n(\gamma)-1\}$;*

(c) *$\gamma(t_i) \in \Delta_A^{-1}(0)$ for each $i \in \{1,\dots,n(\gamma)-1\}$;*

(d) *$\gamma(\bar t) \in \Delta_A^{-1}(0)$ $(\bar t \in Dom(\gamma))$ iff $\bar t = t_i$ for some $i \in \{1,\dots n(\gamma)-1\}$;*

(C) *under generic assumptions,* (a) *(or, equivalently,* (b) *or* (c) *or* (d)*) implies that for every interval $[a,b] \subset [0,\tau]$ $(a < b)$, γ does not correspond to the singular control φ.*

Proof. (A) The covector associated to an extremal trajectory is completely determined after the first switching. From $n(\gamma) \geq 2$ it follows that $\lambda(.)$ is unique up to a positive constant.

Proof that (a) *implies* (c). Let (γ,λ) be an abnormal extremal such that $n(\gamma) \geq 2$. For each t_i $(i = 1,\dots,n(\gamma)-1)$ we have $\lambda(t_i)\cdot G(\gamma(t_i)) = 0$, and there exists a sequence $t'_m \nearrow t_i$ such that

$$\mathcal{H}(\gamma(t'_m),\lambda(t'_m),u(t'_m)) = \lambda(t'_m)\cdot(F + u(t'_m)G)(\gamma(t'_m)) = 0.$$

Hence $\lambda(t'_m)\cdot F(\gamma(t'_m)) \to 0$, and $\lambda(t_i)\cdot F(\gamma(t_i)) = 0$. We can conclude that $F(\gamma(t_i))$ and $G(\gamma(t_i))$ are parallel, $\lambda(t_i)$ being not equal to 0. (c) follows.

Proof of (C). From (P3), $\Delta_A^{-1}(0) \cap \Delta_B^{-1}(0) \cap \mathcal{R}(\tau)$ is a finite set. Generically γ does not switch on $\Delta_A^{-1}(0) \cap \Delta_B^{-1}(0)$ Using Lemma 5, we conclude.

Proof that (b) *implies* (a). Assume now that $\Delta_A(\gamma(t_i)) = 0$ for some $i \in \{1,\dots,n(\gamma)-1\}$. We have $\lambda(t_i)\cdot G(\gamma(t_i)) = 0$, and from $\Delta_A(\gamma(t_i)) = 0$ we get $\lambda(t_i)\cdot F(\gamma(t_i)) = 0$. There exists a sequence $t'_m \nearrow t_i$ such that

$$\mathcal{H}(\gamma(t'_m),\lambda(t'_m),u(t'_m)) = \lambda(t'_m)\cdot(F + u(t'_m)G)(\gamma(t'_m)) \to 0.$$

From the PMP we know that $\mathcal{H}$ is almost everywhere equal to a fixed constant in $Dom(\gamma)$; hence we can conclude $\mathcal{H}(\gamma(t),\lambda(t),u(t)) = 0$ almost everywhere.

Proof that (a) *implies* (d). Fix $\bar t$ such that $\gamma(\bar t) \in \Delta_A^{-1}(0)$. We have $F(\gamma(\bar t)) = \beta G(\gamma(\bar t))$ (by genericity we may assume $\beta \neq 0, \pm 1$), and there exists a sequence $t'_m \nearrow \bar t$ such that $|u(t'_m)| = 1$ and

$$\mathcal{H}(\gamma(t'_m),\lambda(t'_m),u(t'_m)) = \lambda(t'_m)\cdot(F + u(t'_m)G)(\gamma(t'_m)) = 0.$$

Hence $(1 + u(t'_m)\beta)\lambda(t'_m)\cdot G(t'_m) \to 0$ and, being that $\lim_{m\to\infty} u(t'_m) = \pm 1$, we have $\lambda(\bar t)\cdot G(\bar t) = 0$. From (C), $\Delta_B(\gamma(\bar t)) \neq 0$; thus $\dot\theta^\gamma(\bar t) \neq 0$, and $\bar t$ is a switching time (see Lemma 16). Vice versa, since (a) implies (c), we get that $\Delta_A(\gamma(t_i)) = 0$ for each i. Thus we are done.

The implications (c)⇒(b), (d)⇒(b) are obvious. This concludes the proof. ∎

We have the following lemma.

LEMMA 20. *Let (γ, t_1) be an NTAE; then $\gamma \in \partial\Gamma$.*

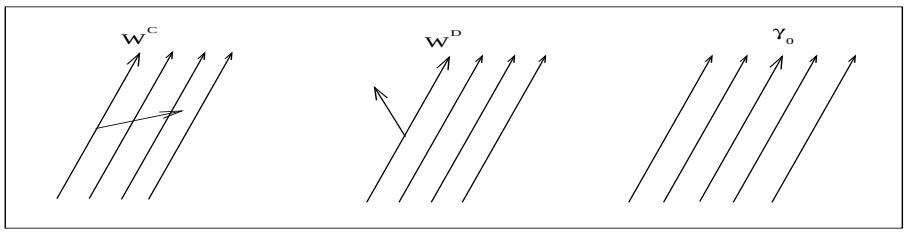


FIG. 3.1. *The abnormal extremals of type W^C, W^D, and γ_0.*

Proof. Let (γ, t_1) be an NTAE. From Proposition 19, $\gamma(t_1) \in \Delta_A^{-1}(0)$. Hence condition (ii) of Definition 10 is violated, so γ cannot be in the interior of a strip. ∎

Then it is natural to define $\partial\Gamma_A = \{\gamma \in \partial\Gamma,\ \gamma \text{ is an abnormal extremal}\}$.

DEFINITION 21. *Let $x \in \Delta_A^{-1}(0)$ be a switching point for an NTAE; then clearly $X(x) \neq 0$, $Y(x) \neq 0$, and $(F+G)(x) = \alpha(F-G)(x)$ for some $\alpha \neq 0$. If $\alpha > 0$ (resp., $\alpha < 0$), we say that* at *x, $\Delta_A^{-1}(0)$* is direct *(resp.,* inverse*).*

From Lemma 20 it follows that any $AA(i)$ is a strip border of kind γ_0 or W, but for abnormal extremals a more precise definition for the strip borders of kind W is necessary.

DEFINITION 22. *We refer to Figure* 3.1. *Let $\gamma \in \partial\Gamma_A$, and suppose that it corresponds to a constant control (say, $+1$) in the interval $]b, c[$ $(0 < b < c \leq \tau)$. Let $\mathcal{S}^1$ and $\mathcal{S}^2$ be the two strips such that $\{\gamma\} = \mathcal{S}^1 \cap \mathcal{S}^2$, and suppose that in the interval $]b, c[$ γ is a strip border of kind W.*

- *We say that in $]b, c[$ γ is a strip border of kind W^C if $\mathcal{S}^1$ and $\mathcal{S}^2$ both lie on the right (resp., on the left) of $\gamma|_{]b,c[}$ and X points to the right (resp., to the left) of $\gamma|_{]b,c[}$ at every point of* $\mathrm{Supp}(\gamma|_{]b,c[})$.
- *We say that in $]b, c[$ γ is a strip border of kind W^D if $\mathcal{S}^1$ and $\mathcal{S}^2$ both lie on the right (resp., on the left) of $\gamma|_{]b,c[}$ and X points to the left (resp., to the right) of $\gamma|_{]b,c[}$ at every points of* $\mathrm{Supp}(\gamma|_{]b,c[})$.

From Proposition 19 we have the following.

LEMMA 23. *Let $A_1, A_2 \in \{W^C, W^D, \gamma_0\}$ and $\gamma \in \partial\Gamma_A$. If γ is of kind A_1 in $]a, b[$ and of kind A_2 in $]b, c[$ $(A_1 \neq A_2,\ 0 < a < b < c \leq \tau)$, then b is a switching time for γ.*

Now the meaning of this more fine definition in the case of abnormal extremal strip borders is clear. An abnormal extremal can be of kind W^C (resp., W^D) on $[t-\varepsilon, t]$, $\varepsilon > 0$ and of kind W^D (resp., W^C) on $[t, t+\varepsilon]$, only if t is a switching time. On the contrary, in the case of strip borders that are not abnormal extremals the change from W^C to W^D or vice versa can occur without switching, and so the difference between W^C and W^D is not useful.

PROPOSITION 24. *Let γ be an NTAE, let $t_1 < t_2 < \cdots < t_{n(\gamma)-1} < t_{n(\gamma)} := \tau$ be the sequence of its switching times, and set $t_0 = 0$. From* (C) *of Proposition* 19, *generically $G(\gamma(t_i)) \neq 0$. If $F(\gamma(t_i)) = \beta_i G(\gamma(t_i))$, then clearly $\beta_i \neq \pm 1$. (Otherwise, $X(\gamma(t_i)) = 0$ or $Y(\gamma(t_i)) = 0$.) For all $i = 0, \ldots, n(\gamma) - 2$ it holds that*

$$v^\gamma(G(\gamma(t_{i+1})), t_{i+1}; 0) = (\beta_i + 1)/(\beta_{i+1} + 1) v^\gamma(G(\gamma(t_i)), t_i; 0)$$
if γ corresponds to control $+1$ on $[t_i, t_{i+1}]$,

$$v^\gamma(G(\gamma(t_{i+1})), t_{i+1}; 0) = (\beta_i - 1)/(\beta_{i+1} - 1) v^\gamma(G(\gamma(t_i)), t_i; 0)$$
if γ corresponds to control -1 on $[t_i, t_{i+1}]$.

Proof. Fix i, and suppose that γ corresponds to constant control $+1$ in $[t_i, t_{i+1}]$, the opposite case being similar. From $F(\gamma(t_{i+1})) = \beta_{i+1} G(\gamma(t_{i+1}))$, recalling Lemma 15, we have

$$\begin{aligned} F(\gamma(t_i)) + G(\gamma(t_i)) &= v^\gamma(F(\gamma(t_{i+1})) + G(\gamma(t_{i+1})), t_{i+1}; t_i) \\ &= (1+\beta_{i+1}) v^\gamma(G(\gamma(t_{i+1})), t_{i+1}; t_i). \end{aligned}$$

Now from $F(\gamma(t_i)) = \beta_i G(\gamma(t_i))$ (notice that in the case when $i = 0$ we have $F(0) = 0$, and hence $\beta_0 = 0$) and using again Lemma 15, we have

$$(1+\beta_i) v^\gamma(G(\gamma(t_i)), t_i; 0) = (1+\beta_{i+1}) v^\gamma(G(\gamma(t_{i+1})), t_{i+1}; 0),$$

which concludes the proof. □

PROPOSITION 25. *Let $\gamma : [0, \tau] \to \mathbb{R}^2$ be an extremal trajectory for the control problem* (1) *that switches at least one time, let θ^γ be the corresponding function defined in* (7)*, and let $t_1 < t_2 < \cdots < t_{n(\gamma)-1} < t_n(\gamma) := \tau$ be the sequence of switching times. Then under generic assumptions the following conditions are equivalent:*

(a) (γ, t_1) *is an NTAE;*
(b) $\theta^\gamma(t_i) \in \{0, \pm\pi\}$ *for some* $i \in \{1, \dots, n(\gamma) - 1\}$*;*
(c) $\theta^\gamma(t_i) \in \{0, \pm\pi\}$ *for each* $i \in \{1, \dots, n(\gamma) - 1\}$*;*
(d) $\theta^\gamma(\bar t) \in \{0, \pm\pi\}$, $(\bar t \in Dom(\gamma))$ *iff* $\bar t = t_i$ *for some* $i \in \{1, \dots n(\gamma) - 1\}$.

Proof. Proof that (a) *implies* (c). By definition $\theta^\gamma(0) = 0$. Proposition 24 implies that the vectors $v^\gamma(G(\gamma(t_{i+1}), t_{i+1}; 0)$ and $v^\gamma(G(\gamma(t_i), t_i; 0)$ are parallel. Since $\theta^\gamma(t_{i+1})$ and $\theta^\gamma(t_i)$ measure precisely the angle between those vectors and $G(0)$, we have $\theta^\gamma(t_{i+1}) = \theta^\gamma(t_i) \pm \pi$.

Proof that (c) *implies* (a). From (c) we have $\theta^\gamma(t_1) \in \{0, \pm\pi\}$; then for some $b \in \mathbb{R}$ (which by genericity we may assume different from 0 and 1) it holds that $v^\gamma(G(\gamma(t_1)), t_1; 0) = bG(\gamma(0))$. Now if we suppose that γ corresponds to constant control $+1$ in the interval $[0, t_1]$ (the opposite case being similar), we have $bG(\gamma(0)) = b(F+G)(\gamma(0)) = v^\gamma(b(F+G)(\gamma(t_1)), t_1; 0)$. From the injectivity of the map $v_0 \to v^\gamma(v_0, t_0; t_1)$ we obtain $\gamma(t_1) \in \Delta_A^{-1}(0)$. Using Proposition 19 (a) follows.

Proof that (b) *implies* (c) *and vice versa.* Clearly (b) follows from (c); let us prove the opposite. Let λ be the covector associated to γ. From (b) we have that $\lambda(0)$ is orthogonal to $G(0)$, and hence γ switches iff $\theta^\gamma \in \{0, \pm\pi\}$. Thus (c) follows.

Proof that (a) *implies* (d). It is a consequence of the generic condition $\theta^\gamma(\bar t) = 0 \Rightarrow \dot\theta^\gamma(\bar t) \neq 0$.

The implication (d)⇒(c) is obvious. This concludes the proof. □

PROPOSITION 26. *Let (γ, t_1) be an NTAE, and let $0 =: t_0 < t_1 < t_2 < \cdots < t_{n(\gamma)-1} < t_{n(\gamma)} := \tau$ be the sequence of switching times. Suppose that for some $i \in \{0, 1, \dots, n(\gamma)-2\}$, $\Delta_A^{-1}(0)$ is inverse at the points $\gamma(t_i)$, $\gamma(t_{i+1})$. Then $\theta^\gamma(t_{i+1}) = \theta^\gamma(t_i)$.*

Proof. Set $(F+G)(\gamma(t_i)) = \alpha_i (F-G)(\gamma(t_i))$ and $F(\gamma(t_i)) = \beta_i G(\gamma(t_i))$. Under generic assumptions, α_i and β_i are well defined for each $i = 1, \dots, n(\gamma) - 1$ and it holds that

$$\alpha_0 = -1, \quad \beta_0 = 0, \text{ and } \alpha_i, \beta_i \notin \{0, \pm 1\}, \quad \beta_i = \frac{1+\alpha_i}{1-\alpha_i}, \quad i \in \{1, ..., n(\gamma) - 1\}.$$

Now if $\Delta_A^{-1}(0)$ is inverse at both points $\gamma(t_i), \gamma(t_{i+1})$ ($i \in \{0, \dots, n(\gamma) - 2\}$, then $\alpha_i, \alpha_{i+1} < 0$ and we have $\beta_i, \beta_{i+1} \in]-1, 1[$ (see Figure 3.2). Recalling the definition of θ^γ, from Proposition 24 the conclusion follows. □

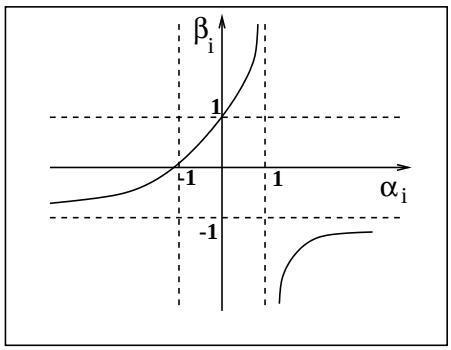


FIG. 3.2.

From Propositions 19 and 25, using the definitions of the times s_1^+, $s_1'^+$, t_f^+, s_1^-, $s_1'^-$, t_f^-, we have the following corollary.

COROLLARY 27. *Let γ be an extremal trajectory exiting the origin with control $+1$; then its first switching can occur on $\Delta_A^{-1}(0)$ only if $s_1 \neq 0$ (condition* (A)*) or $s_1' \neq 0$ (condition* (B)*) or $|\theta^+(t_f^+)| = \pi$ (condition* (C)*). Moreover, at most one of the conditions* (A), (B), (C) *holds, and the corresponding time is the first switching time of γ and the first time at which γ^+ intersect $\Delta_A^{-1}(0)$. A similar result holds for γ^- and for the times s_1^-, $s_1'^-$, t_f^-.*

Referring to conditions 1, 2, 3 of Observation 1, conditions A and B correspond either to case 2 or case 1 (with $s_1^+ \neq 0$ or $s_1'^+ \neq 0$), and condition C corresponds to case 3. Moreover, it is clear that for an NTAE (γ, t_1), t_1 is the first time at which γ reaches $\Delta_A^{-1}(0)$. In particular, if the trajectory exits the origin with control $+1$, we have $t_1 = s_1^+$ or $t_1 = s_1'^+$ or $t_1 = t_f^+$. In the case when $s_1^+ = 0$ and $s_1'^+$ are not defined or vice versa and $|\theta^+(t_f^+)| < \pi$ (which implies $t_f^+ = \tau$), there are no NTAEs exiting the origin with control $+1$. (An abnormal extremal exists but it never switches.) This case corresponds to case 1 of Observation 1 with $s_1^+ = 0$ and s'^+ not defined or vice versa. These observations are collected in the following.

COROLLARY 28. *There are at most two maximal NTAE. Moreover,*

(♠) *one exits the origin with control $+1$, and its first switching is at*

- s_1^+ *iff* $s_1^+ \neq 0$;
- $s_1'^+$ *iff* $s_1'^+ \neq 0$;
- t_f^+ *iff* $|\theta^+(t_f^+)| = \pi$;

(♣) *the other exits the origin with control -1, and its first switching is at*

- s_1^- *iff* $s_1^- \neq 0$;
- $s_1'^-$ *iff* $s_1'^- \neq 0$;
- t_f^- *iff* $|\theta^-(t_f^-)| = \pi$.

Finally, if $|\theta^\pm(t_f^\pm)| = \pi$, then $\Delta_A^{-1}(0)$ is direct at $\gamma(t_f^\pm)$.

The following two propositions describe the position of the switching curves of the strips whose borders are abnormal extremals.

PROPOSITION 29. *Let (γ, t_1) be an NTAE, let t_i and t_{i+1} be two consecutive switching times, let $\mathcal{S}$ be a strip such that $\gamma \in \partial S$, and let U^i, U^{i+1} be two sufficiently small neighborhoods of $\gamma(t_i)$ and $(\gamma(t_{i+1})$. Moreover, let U_{in}^i and U_{out}^i (resp., U_{in}^{i+1}, U_{out}^{i+1}) be the two connected components of $U^i \setminus \Delta_A^{-1}(0)$ (resp., $U^{i+1} \setminus \Delta_A^{-1}(0)$) chosen in such a way that γ enters U_{in}^i (resp., U_{in}^{i+1}). Under generic conditions we have the following cases:*

(1) $\theta^\gamma(t_i) = \theta^\gamma(t_{i+1})$ *and* $\Delta_A^{-1}(0)$ *direct at* $\gamma(t_i)$. *In this case if the switching locus of* $\mathcal{S}$ *passing through* $\gamma(t_i)$ *lies in* U^i_{in} *(resp.,* U^i_{out}*), then the switching locus of* $\mathcal{S}$ *passing through* $\gamma(t_{i+1})$ *lies in* U^{i+1}_{out} *(resp.,* U^{i+1}_{in}*).*

(2) $\theta^\gamma(t_i) = \theta^\gamma(t_{i+1})$ *and* $\Delta_A^{-1}(0)$ *inverse at* $\gamma(t_i)$. *In this case if the switching locus of* $\mathcal{S}$ *passing through* $\gamma(t_i)$ *lies in* U^i_{in} *(resp.,* U^i_{out}*), then the switching locus of* $\mathcal{S}$ *passing through* $\gamma(t_{i+1})$ *lies in* U^{i+1}_{in} *(resp.,* U^{i+1}_{out}*).*

(3) $\theta^\gamma(t_i) = \theta^\gamma(t_{i+1}) \pm \pi$ *and* $\Delta_A^{-1}(0)$ *direct at* $\gamma(t_i)$. *In this case we have the same conclusion as in case* (2).

(4) $\theta^\gamma(t_i) = \theta^\gamma(t_{i+1}) \pm \pi$ *and* $\Delta_A^{-1}(0)$ *inverse at* $\gamma(t_i)$. *In this case we have the same conclusion as in case* (1).

Proof. Let f_i (resp., A_i) be the sign of $-\Delta_B/\Delta_A$ (resp., Δ_A) on U^i_{in}, and let B_i be the sign of Δ_B on U^i. By hypothesis, taking U^i sufficiently small, these quantities are well defined and we have $f_i = -A_iB_i$. Moreover, set $\theta_i = +1$ if $\theta^\gamma(t_i) = \theta^\gamma(t_{i+1}) \pm \pi$ and $\theta_i = -1$ if $\theta^\gamma(t_i) = \theta^\gamma(t_{i+1})$.

Claim. $B_{i+1} = \theta_i B_i$.

Proof of the Claim. From $sgn(\dot\theta^\gamma(t)) = sgn(\Delta_B(\gamma(t)))$ we have that $\bar v^\gamma(t)$ (see formula (7)) is a vector rotating counterclockwise in U^i (resp., U^{i+1}) iff $B_i > 0$ (resp., $B_{i+1} > 0$). Recalling that $\theta^\gamma(t_i), \theta^\gamma(t_{i+1}) \in \{0, \pm\pi\}$ it is clear that B_i and B_{i+1} have the same sign iff $\theta^\gamma(t_{i+1}) = \theta^\gamma(t_i) \pm \pi$. The claim is proved.

Case 1. First suppose Δ_A is direct at $\gamma(t_i)$. In this case from Proposition 19 we clearly have $A_{i+1} = -A_i$. Now if the switching loci of $\mathcal{S}$ lie one in U^i_{in} and one in U^{i+1}_{in} (resp., U^i_{out} and U^{i+1}_{out}), then, from Lemma 8 we have $f_i = -f_{i+1}$. This occurs iff $-A_iB_i = +A_{i+1}B_{i+1} = -A_iB_i\theta_i$, from which it follows that $\theta_i = +1$.

On the other hand, if the switching loci of $\mathcal{S}$ lie one in U^i_{in} and one in U^{i+1}_{out} (resp., U^i_{out} and U^{i+1}_{in}), then $f_i = +f_{i+1}$, which occurs iff $\theta_i = -1$.

Case 2. If Δ_A is inverse at $\gamma(t_i)$, we have $A_{i+1} = +A_i$. Now if the switching loci of $\mathcal{S}$ lie one in U^i_{in} and one in U^{i+1}_{in} (resp., U^i_{out} and U^{i+1}_{out}), we have $f_i = -f_{i+1}$. This occurs iff $\theta_i = -1$.

On the other hand, if the switching loci of $\mathcal{S}$ lie one in U^i_{in} and U^{i+1}_{out} (resp., U^i_{out} and U^{i+1}_{in}), then $f_i = +f_{i+1}$, which occurs iff $\theta_i = +1$. □

PROPOSITION 30. *Let* (γ, t_1) *be an NTAE, and let* $\mathcal{S}^1$ *and* $\mathcal{S}^2$ *be two strips such that* $\{\gamma\} = \mathcal{S}^1 \cap \mathcal{S}^2$. *Let* $\bar t$ *be a switching time for* γ, *and let* U *be a small neighborhood of* $\gamma(\bar t)$ *such that* $U \setminus \Delta_A^{-1}(0)$ *has two connected components* U_{in} *and* U_{out}, *chosen in such a way that* γ *enters* U *from* U_{in}. *Then, under generic conditions, the switching loci of* $\mathcal{S}^1$ *and* $\mathcal{S}^2$ *passing through* $\gamma(\bar t)$ *satisfy the following:*

(a) *they both lie in* U_{in} *or both lie in* U_{out};

(b) *they are tangent to* $\mathrm{supp}(\gamma)$ *in* $\gamma(\bar t)$.

Proof of (a). By the analysis of the singularities at the first switching time (see Figure 4.1) we know that (a) is true in the special case $\bar t = t_1$. Using Proposition 29 and by induction the thesis follows.

Proof of (b). We prove (b) by induction on the switching time. We start considering the first switching time and treat only the case in which γ exits the origin with constant control $+1$, the opposite case being similar. First suppose that $\bar t$ is the first switching time for γ, and assume $s_1^+ > 0$ (i.e., $\bar t = s_1^+$), the case $s_1'^+ > 0$ and $|\theta^+(t_f^+)| = \pi$ being similar. Set $x = (x_1, x_2)$, and choose a local system of coordinates in such a way that

$$Y \equiv \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad X(x) = \begin{pmatrix} -1 + a_1x_1 + a_2x_2 + O(|x|^2) \\ b_1x_1 + b_2x_2 + O(|x|^2) \end{pmatrix},$$

$$X(x) = \begin{pmatrix} c_0 + c_1(x_1 - s_1^+) + c_2 x_2 + O(|x - (s_1^+, 0)|^2) \\ d_1(x_1 - s_1^+) + d_2 x_2 + O(|x - (s_1^+, 0)|^2) \end{pmatrix}.$$

Generically $\Delta_B(\gamma^+(s_1^+)) \neq 0$, and thus a switching curve generates at $\gamma^+(s_1^+)$.

Let $(\bar\gamma, \bar\lambda)$ be an extremal trajectory of (1) corresponding to constant control -1 in the interval $[0, \tau_1[$ and to constant control $+1$ in the interval $[\tau_1, \tau_2]$. We consider the trajectories $\bar\gamma$ that are near to γ, that is, those corresponding to (τ_1, τ_2) in a neighborhood of $(0, s_1^+)$. We have

$$\bar\gamma(\tau_1) = \begin{pmatrix} -\tau_1 + O(\tau_1^2) \\ -\frac{1}{2} b_1 \tau_1^2 + O(\tau_1^3) \end{pmatrix}, \quad \bar\gamma(\tau_2) = \begin{pmatrix} -\tau_1 + \tau_2 + O(\tau_1^2) \\ -\frac{1}{2} b_1 \tau_1^2 + O(\tau_1^3) \end{pmatrix}.$$

Now if τ_1, τ_2 are switching times for $\bar\gamma$, we must have

$$\bar\lambda(\tau_1) \cdot G(\bar\gamma(\tau_1)) = 0, \tag{8}$$

$$\bar\lambda(\tau_2) \cdot G(\bar\gamma(\tau_2)) = 0. \tag{9}$$

Moreover, from equation (i) of the PMP we have $\bar\lambda(\tau_1) = \bar\lambda(\tau_2) =: \bar\lambda$. Finally, set $\bar\lambda = (\bar\lambda_1, \bar\lambda_2)$, and normalize $\bar\lambda$ in such a way $\bar\lambda_2 = \sqrt{1 - \bar\lambda_1}$. Equations (8) and (9) become

$$f_1(\lambda_1, \tau_1, \tau_2) := \bar\lambda_1(1 + a_1\tau_1) + \sqrt{1 - \bar\lambda_1}\left(\frac{1}{2} b_1 \tau_1\right) + O(\tau_1^2) = 0,$$

$$f_2(\lambda_1, \tau_1, \tau_2) := \bar\lambda_1\left(\frac{1}{2}(1 - c_0) - c_1(-\tau_1 + (\tau_2 - s_1^+))\right) + \sqrt{1 - \bar\lambda_1}\left(-\frac{1}{2} d_1(-\tau_1 + (\tau_2 - s_1^+))\right) + O(\tau_1^2) + O((\tau_2 - s_1^+)^2) = 0.$$

These are two equations for the variable $(\bar\lambda_1, \tau_1, \tau_2)$, and $(0, 0, s_1^+)$ is a solution. We compute the 2×2 Jacobian matrix of partial derivatives of f_1 and f_2 with respect to (λ_1, λ_2) and check that its determinant at the point $(0, 0, s_1^+)$ is equal to $-d_1/2$. Under the generic assumption $d_1 \neq 0$, we can solve the system in a neighborhood of $(0, 0, s_1^+)$ expressing $(\bar\lambda_1, \tau_2)$ as a function of τ_1. This yields

$$\left.\frac{\partial \tau_2}{\partial \tau_1}\right|_{(0,0,s_1^+)} = -\frac{(b_1(1 - c_0) - 2d_1)}{2d_1} =: m; \tag{10}$$

hence $\tau_2 = s_1^+ + m\tau_1 + O(\tau_1^2)$, and under the generic condition $m \neq 1$ the parametric expression for the switching curve starting at $(s_1^+, 0)$ is

$$x_1(\tau_1) = s_1^+ + (m - 1)\tau_1 + O(\tau_1^2),$$
$$x_2(\tau_1) = -\frac{1}{2} b_1 \tau_1^2 + O(\tau_1^3).$$

If $m \neq 0$, we can use τ_2 as a parameter, so

$$x_1(\tau_2) = s_1^+ + \frac{m - 1}{m}(\tau_2 - s_1^+) + O((\tau_2 - s_1^+)^2),$$
$$x_2(\tau_2) = -\frac{b_1}{2m^2}(\tau_2 - s_1^+)^2 + O((\tau_2 - s_1^+)^3).$$

Assuming the generic condition $c_0 \neq 1$ and $b_1 \neq 0$, we may express x_2 as a function of x_1:

$$x_2 = -\frac{2d_1^2}{b_1(1-c_0)^2}(x_1 - s_1^+)^2 + O((x_1 - s_1^+)^3). \tag{11}$$

Notice that this curve can be of kind C or $\bar{C}$. This proves that the switching curve starting at the first switching time is tangent to $\mathrm{Supp}(\gamma)$.

Now we have to prove the induction step; that is, if t_n, t_{n+1} are two consequent switching times for γ and the switching curve passing through t_n is tangent to γ, then the switching curve passing through t_{n+1} is tangent to γ as well. Let us consider only the case in which γ corresponds to the constant control $+1$ on $[t_n, t_{n+1}]$, the opposite case being similar. Choose a local system of coordinates, and rescale the time in such a way that

$$\begin{aligned}
&t_n = 0, \quad \gamma(t_n) = 0,\\
&Y \equiv \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad X(x) = \begin{pmatrix} e_0 + e_1 x_1 + e_2 x_2 + O(|x|^2) \\ f_1 x_1 + f_2 x_2 + O(|x|^2) \end{pmatrix},\\
&X(x) = \begin{pmatrix} g_0 + g_1(x_1 - t_{n+1}) + g_2 x_2 + O(|x - (t_{n+1}, 0)|^2) \\ h_1(x_1 - t_{n+1}) + h_2 x_2 + O(|x - (t_{n+1}, 0)|^2) \end{pmatrix}.
\end{aligned}$$

We can parameterize the switching curve C_n through $\gamma(t_n)$ in the following way. Given $x \in C_n$, there exists (γ_x, λ_x) extremal that switches at x at time $\tau_n(x)$. By induction we may assume that τ_n is invertible for x near $\gamma(t_n)$ and parameterize C_n by τ_n:

$$\begin{aligned}
x_1(\tau_n) &= \alpha_n \tau_n + O(\tau_n^2),\\
x_2(\tau_n) &= \beta_n \tau_n^2 + O(\tau_n^3)
\end{aligned}$$

for some α_n, β_n that by genericity we may assume different from zero.

Let $(\tilde{\gamma}, \tilde{\lambda}) := (\tilde{\gamma}, \tilde{\lambda})_{\tau_n}$ be the extremal trajectory of (1) switching at τ_n on C_n, and let τ_{n+1} be the next switching time. We have

$$\tilde{\gamma}(\tau_n) = \begin{pmatrix} \alpha_n \tau_n + O(\tau_n^2) \\ \beta_n \tau_n^2 + O(\tau_n^3) \end{pmatrix}, \quad \tilde{\gamma}(\tau_{n+1}) = \begin{pmatrix} \alpha_n \tau_n + \tau_{n+1} + O(\tau_n^2) \\ \beta_n \tau_n^2 + O(\tau_n^3) \end{pmatrix}$$

and

$$\tilde{\lambda}(\tau_n) \cdot G(\tilde{\gamma}(\tau_n)) = 0, \tag{12}$$

$$\tilde{\lambda}(\tau_{n+1}) \cdot G(\tilde{\gamma}(\tau_{n+1})) = 0. \tag{13}$$

With computations entirely similar to the previous ones, using similar generic conditions we conclude that the switching curve passing through $\gamma(t_{n+1})$ has the expression

$$\begin{aligned}
x_1(\tau_{n+1}) &= t_{n+1} + \alpha_{n+1}(\tau_{n+1} - t_{n+1}) + O((\tau_{n+1} - t_{n+1})^2),\\
x_2(\tau_{n+1}) &= \beta_{n+1}(\tau_{n+1} - t_{n+1})^2 + O((\tau_{n+1} - t_{n+1})^3)
\end{aligned}$$

for some $\alpha_{n+1}, \beta_{n+1} \neq 0$. This conclude the proof. □

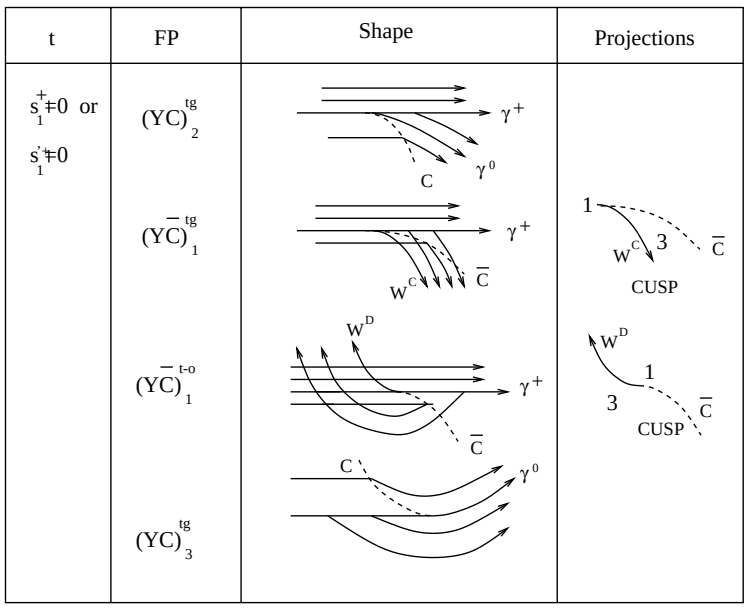


FIG. 4.1.

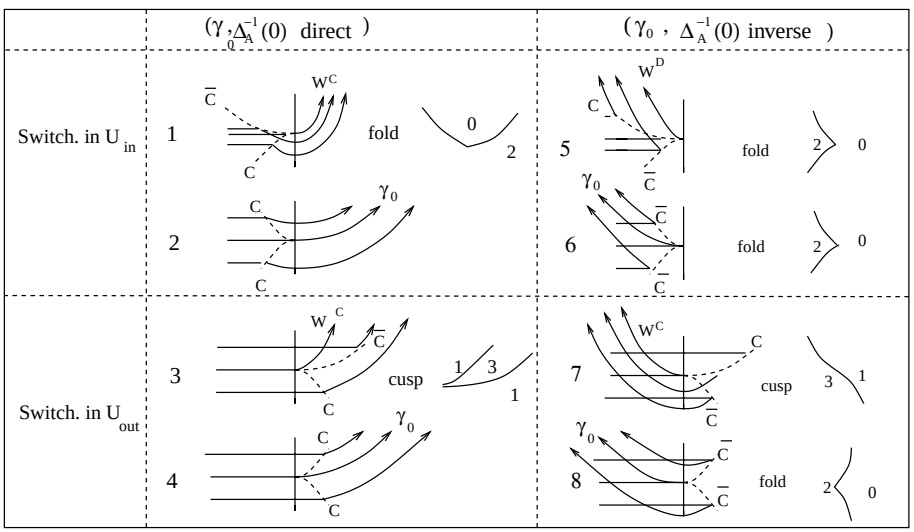


FIG. 4.2.

4. Singularities. In this section we describe all possible FPs occurring along an NTAE.

We start to describe the first singularity for the NTAE exiting the origin with control $+1$, the opposite case being similar. We refer to Figure 4.1, where all extremal trajectories are depicted in a neighborhood of the FPs. Following Corollary 28, an NTAE generates at time s_1^+ (iff $s_1^+ \neq 0$), or at time $s_1'^+$ (iff $s_1'^+ \neq 0$), or at time t_f^+ (iff $|\theta^+(t_f^+)| = \pi$). Assume $s_1^+ \neq 0$; then a switching curve tangent to $\mathrm{Supp}(\gamma^+)$ bifurcates from $\gamma^+(s_1^+)$. Recall formula (11) of the proof of (b) of Proposition 30. Using the

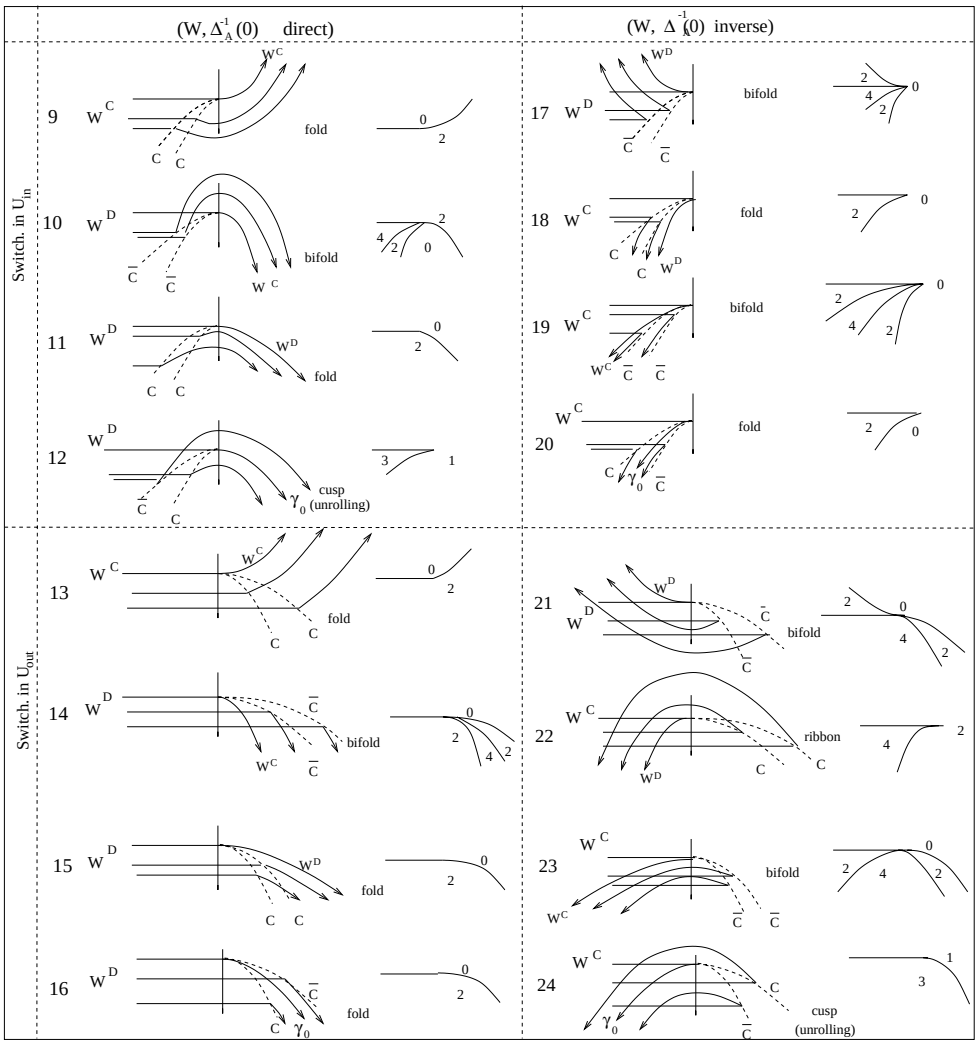


FIG. 4.3.

definition of θ^+ and s_1^+, as in Proposition 3.1 of [7], one gets that $b_1 > 0,\ d_1 < 0$ (we would have $b_1 < 0,\ d_1 > 0$, if $s_1^+ \neq 0$), and $c_0 < 1$. Hence the switching curve is bifurcating to the right of $\mathrm{Supp}(\gamma^+)$. Moreover, it follows that m, defined in formula (10), is bigger than 1, and being that $t_1 > 0$, we have (in formula (11)) $x_1 > s_1^+$. If the switching curve is of kind C, we call the singularity $(Y, C)_2^{tg}$. If the switching curve is of kind $\bar{C}$ and $\Delta_A^{-1}(0)$ is direct at $\gamma^+(s_1^+)$, we call the singularity $(Y, \bar{C})_1^{tg}$. Finally, if the switching curve is of kind $\bar{C}$ and $\Delta_A^{-1}(0)$ is inverse at $\gamma^+(s_1^+)$, we call the singularity $(Y, \bar{C})_1^{t-o}$. These names are chosen in accordance with [5, 17]. The case in which $s_1'^+ \neq 0$ is entirely similar.

The case in which the NTAE starts at $\gamma(t_f^+)$ (that happens iff $|\theta^+(t_f^+)| = \pi$) is again described by formula (11) with s_1^+ replaced by t_f^+. In this case, reasoning as in [7], we get $c_0 > 1$ and $b_1 d_1 < 0$. It follows that the switching curve bifurcates to the right, $m < 1$, and in formula (11) we have $x_1 < t_f^+$. Moreover, at t_f^+, γ^+ stops to be

extremal and $\Delta_A^{-1}(0)$ is direct at $\gamma(t_f^+)$ (see Proposition 26). We call this singularity $(Y, C)_3^{tg}$.

To classify the other generic singularities involving an NTAE, we consider at the FPs

- $\Delta_A^{-1}(0)$ direct or inverse;
- switching in U_{in} or U_{out} according to Proposition 29;
- all the *essentially different* directions of the exiting abnormal trajectory.

We obtain 24 types of singularities. The singularities with entering abnormal extremals of kind γ_0 are shown in Figure 4.2, while in Figure 4.3 all the possible singularities for an NTAE of the kinds W^C and W^D are listed. In these figures we also indicate the labels *fold, cusp, bifold,* or *ribbon* in accordance with Definition 9.

5. Classification of abnormal extremals. In this section we prove Theorem 2.

From now on we fix a positive time $\tau > 0$. Using Corollary 28, we have that for each system $(F, G) \in \Xi$ there exists exactly two maximal abnormal extremals $\gamma_A^{\pm} = \gamma_A^{\pm}((F, G), \tau)$ in time τ exiting the origin, respectively, with control ± 1.

DEFINITION 31. *We say that a set A provides a generic classification of abnormal extremals (in time τ) if there exist a generic subset Π of Ξ and a map $\Phi : \Pi \to A$ such that $\Phi((F, G)) = \Phi((F', G'))$, $(F, G), (F', G') \in \Pi$ iff the corresponding maximal abnormal extremals present the same finite sequence of generic singularities.*

Remark. From section 4, we know the structure of the set of extremal trajectories near each generic singularity. Hence, if two abnormal extremals present the same sequence of singularities, then the synthesis near them is exactly the same.

We provide a generic classification through a set of words A recognizable by an *automaton*. Therefore, proving Theorem 2 amounts precisely to constructing an automaton describing all possible sequences of generic singularities along an NTAE.

First, we build an automaton, naturally associated to a system, with the simplest possible set of edges. By this automaton we can prove that the ribbon and the bifold singularities are realized, but more than one sequence of singularities may correspond to a recognizable word. Then we build a more complicated automaton that has the required property; i.e., to every recognizable word it corresponds one and only one sequence of generic singularities.

Let us first recall some definitions from automata theory. For a more extensive and detailed treatment of the subject we refer to [10, 11].

DEFINITION 32. *Let Σ be a finite set, and consider the set Σ^* of ordered n-tuples $s = (\sigma_1, \dots, \sigma_k)$, $\sigma_i \in \Sigma$ $(i = 1, \dots, k)$, $k \geq 0$. We call Σ the* alphabet, *$\sigma \in \Sigma$ a* letter, *$s = (\sigma_1, \dots, \sigma_k) \in \Sigma^*$ a* word of length *k, and Σ^* the set of words generated by Σ.*

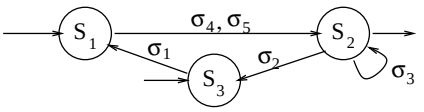
The set of words generated by an alphabet is a set with a simple structure, and a classification based on such a set is quite satisfying. Such a kind of classification was given, for example, for the sequence of generic singularities along $\gamma^{\pm}$ in [15]. For abnormal extremals we have to use a set with a more complicated structure.

DEFINITION 33. *Let Σ be a finite* alphabet*; an* automaton *$\mathcal{A}$ over Σ consists of the following:*

- *a finite set $\mathbb{S}$ whose elements are called* states*;*
- *a set of* initial states *$\mathbb{I} \subseteq \mathbb{S}$;*
- *a set of* terminal states *$\mathbb{T} \subseteq \mathbb{S}$;*
- *a set of* edges*, that is, a subset $\mathbb{E} \subseteq \mathbb{S} \times \Sigma \times \mathbb{S}$. An* edge *is indicated as (S_1, σ, S_2), and we say that it begins at S_1, it ends at S_2, and it carries the*

label σ.

Usually an *automaton* is represented by a set of circles (states) and a set of arrows that connect the circles (the edges). The initial (resp., final) states are labelled by arrows pointing toward (resp., away from) the circle. If there are several edges beginning and ending at the same states, they are replaced by a single arrow carrying several labels.



A *path* in $\mathcal{A}$ is a finite sequence of edges of the type $(S_1, \sigma_1, S_2)(S_2, \sigma_2, S_3), \dots, (S_k, \sigma_k, S_{k+1})$. If $S_1 \in \mathbb{I}$ and $S_{k+1} \in \mathbb{T}$, we say that the *path* is successful.

DEFINITION 34. *A set of words* $\Omega \subset \Sigma^*$ *is said to be recognizable by* $\mathcal{A}$ *if for every word* $(\sigma_1, \sigma_2, \dots, \sigma_m) \in \Omega$ *of length* m *there exists* $S_1, \dots, S_{m+1} \in \mathbb{S}$ *such that*

- $(S_i, \sigma_i, S_{i+1}) \in \mathbb{E}$ *for every* $i = 1, \dots, m$*;*
- $(S_1, \sigma_1, S_2)(S_2, \sigma_2, S_3), \dots, (S_m, \sigma_m, S_{m+1})$ *is a successful path.*

The set of words recognizable by an automaton share some regularity properties; in particular, they are studied in automata theory (see [10, 11]). Let us resume all the information on the switching strategy of an abnormal extremal via three rules. Let (γ, t_1) be an NTAE, and let $t_1 < t_2, \dots, < t_{n(\gamma)-1} < t_{n(\gamma)} := \tau$ be the sequence of its switching times.

R1. Let $\mathcal{S}^1$ and $\mathcal{S}^2$ be the two strips such that $\gamma \in \mathcal{S}^1 \cap \mathcal{S}^2$, and let $\bar{\mathcal{S}} := \mathcal{S}^1 \cup \mathcal{S}^2$. We have the following cases:

(1) $\theta^\gamma(t_i) = \theta^\gamma(t_{i+1})$ and $\Delta_A^{-1}(0)$ direct at $\gamma(t_i)$. In this case if the switching locus of $\bar{\mathcal{S}}$ passing through $\gamma(t_i)$ lies in U^i_{in} (resp., U^i_{out}), then the switching locus of $\bar{\mathcal{S}}$ passing through $\gamma(t_{i+1})$ lies in U^{i+1}_{out} (resp., U^{i+1}_{in}).

(2) $\theta^\gamma(t_i) = \theta^\gamma(t_{i+1})$ and $\Delta_A^{-1}(0)$ inverse at $\gamma(t_i)$. In this case if the switching locus of $\bar{\mathcal{S}}$ passing through $\gamma(t_i)$ lies in U^i_{in} (resp., U^i_{out}), then the switching locus of $\bar{\mathcal{S}}$ passing through $\gamma(t_{i+1})$ lies in U^{i+1}_{in} (resp., U^{i+1}_{out}).

(3) $\theta^\gamma(t_i) = \theta^\gamma(t_{i+1}) \pm \pi$ and $\Delta_A^{-1}(0)$ direct at $\gamma(t_i)$. In this case we have the same conclusion as in case (2).

(4) $\theta^\gamma(t_i) = \theta^\gamma(t_{i+1}) \pm \pi$ and $\Delta_A^{-1}(0)$ inverse at $\gamma(t_i)$. In this case we have the same conclusion as in case (1).

The rule R1 is a direct consequence of Propositions 29 and 30(a).

R2. If $\Delta_A^{-1}(0)$ is inverse at t_i and t_{i+1}, then $\theta^\gamma(t_{i+1}) = \theta^\gamma(t_i)$. The rule R2 follows from Proposition 26.

R3. Only the following consecutive singularities are possible:

Singularity at t_i	Possible singularity at t_{i+1}	AA(i)
$(YC)^{tg}_2$,$(YC)^{tg}_3$,2,4,6,8,12,16,20,24	1,2,3,4,5,6,7,8	γ^0
$(Y\bar{C})^{tg}_1$,1,3,7,9,10,13,14,19,23	9,13,18,19,20,22,23,24	W^C
$(Y\bar{C})^{t-o}_1$,5,11,15,17,18,21,22	10,11,12,14,15,16,17,21	W^D

Clearly, if an abnormal arc of kind γ^0 (resp., W^C, W^D) exits the singularity $\gamma(t_i)$ $(i = 1, \dots, n(\gamma) - 1)$, then an abnormal arc of kind γ^0 (resp., W^C, W^D) enters the singularity $\gamma(t_{i+1})$. Thus R3 can be directly checked using Figures 4.1, 4.2, and 4.3.

We now are ready to build an automaton $\mathcal{A}$. For us the set of *states* is the set of the 28 singularities

$$\mathbb{S} := \{(YC)^{tg}_2, (Y\bar{C})^{tg}_1, (Y\bar{C})^{t-o}_1, (YC)^{tg}_3,$$

TABLE A.

Letters→ states ↓	0	π
$(YC)_2^{tg}$	1,2,5,6	3,4,7,8
$(Y\bar{C})_1^{tg}$	9,18,19,20	13,22,23,24
$(Y\bar{C})_1^{t-o}$	14,15,16,21	10,11,12
$(YC)_3^{tg}$	1,2,5,6	3,4,7,8
1	13,22,23,24	9,18,19,20
2	3,4,7,8	1,2,5,6
3	9,18,19,20	13,22,23,24
4	1,2,5,6	3,4,7,8
5	10,11,12,17	14,15,16
6	1,2,5,6	3,4
7	13,22,23,24	9
8	3,4,7,8	1,2
9	13,22,23,24	9,18,19,20
10	13,22,23,24	9,18,19,20
11	14,15,16,21	10,11,12,17
12	3,4,7,8	1,2,5,6
13	9,18,19,20	13,22,23,24
14	9,18,19,20	13,22,23,24
15	10,11,12,17	14,15,16,21
16	1,2,5,6	3,4,7,8
17	10,11,12,17	14,15,16
18	10,11,12,17	14,15,16
19	9,18,19,20	13
20	1,2,5,6	3,4
21	14,15,16,21	10,11,12
22	14,15,16,21	10,11,12
23	13,22,23,24	9
24	3,4,7,8	1,2

$$1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24\},$$

and the alphabet is

$$\Sigma := \{0, \pi\},$$

that is (if we are considering two singularities at t_i and t_{i+1}), the set of values assumed by the function $\Delta\theta_i^\gamma := |\theta^\gamma(t_{i+1}) - \theta^\gamma(t_i)|$. The set of *initial states* is constituted by the singularities $(YC)_2^{tg}, (Y\bar{C})_1^{tg}, (Y\bar{C})_1^{t-o}, (YC)_3^{tg}$, and the set of *terminal states* coincides with $\mathbb{S}$. Using rules R1–R3, we obtain Table A, which shows how the *edges* connect the *states*; that is, it describes the set of edges $\mathbb{E}$. For example, from the *state* (singularity) 18, using the *letter* π, we may reach the states $14, 15, 16$. This means that the edges of $\mathbb{E}$ with label π that start at the state 18 are $(18, \pi, 14), (18, \pi, 15), (18, \pi, 16)$. It is clear that for this automaton every word of Σ^* is recognizable, but $\mathcal{A}$ does not provide a generic classification because a word corresponds to more than one sequence of singularities. However, it describes in a simple way the set of abnormal extremals, and, in particular, from Table A we have the following theorem.

THEOREM 35. *All states* 1–24 *can be reached with at most two edges. More precisely, only the singularity* 17 *needs, in fact, two edges. Moreover, the singularity number* 22 *(the ribbon) can be realized with the edge* $((Y\bar{C})_1^{tg}, \pi, 22)$ *and the singularity number* 10 *(which is a bifold) with the edge* $((Y\bar{C})_1^{t-o}, \pi, 10)$.

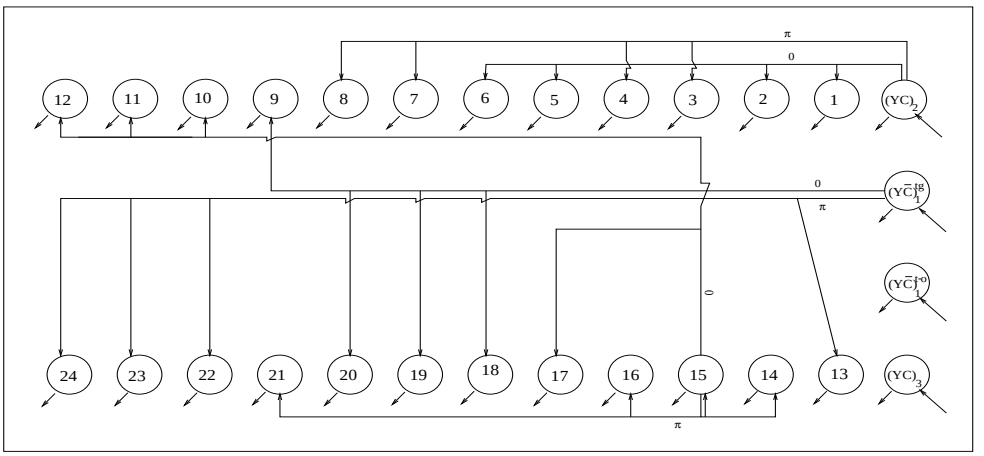


FIG. 5.1. *The automaton. (Some edges are omitted.)*

TABLE B *(first part).*

Letters→ states ↓	$(0, D, \gamma^0)$	$(0, D, W^C)$	$(0, D, W^D)$	$(0, I, \gamma^0)$	$(0, I, W^C)$	$(0, I, W^D)$
$(YC)_2^{tg}$	2	1	–	6	–	5
$(Y\bar{C})_1^{tg}$	–	9	–	20	19	18
$(Y\bar{C})_1^{t-o}$	16	14	15	–	–	21
$(YC)_3^{tg}$	2	1	–	6	–	5
1	–	13	–	24	23	22
2	4	3	–	8	7	–
3	–	9	–	20	19	18
4	2	1	–	6	–	5
5	12	10	11	–	–	17
6	2	1	–	6	–	5
7	–	13	–	24	23	22
8	4	3	–	8	7	–
9	–	13	–	24	23	22
10	–	13	–	24	23	22
11	16	14	15	–	–	21
12	4	3	–	8	7	–
13	–	9	–	20	19	18
14	–	9	–	20	19	18
15	12	10	11	–	–	17
16	2	1	–	6	–	5
17	12	10	11	–	–	17
18	12	10	11	–	–	17
19	–	9	–	20	19	18
20	2	1	–	6	–	5
21	16	14	15	–	–	21
22	16	14	15	–	–	21
23	–	13	–	24	23	22
24	4	3	–	8	7	–

Figure 5.1 shows the automaton with all states but only the edges of the kind $((YC)_2^{tg}, ., .)$, $((Y\bar{C})_1^{tg}, ., .)$, $(15, ., .)$.

To build a new automaton $\mathcal{A}'$ that provides a generic classification, we need to include more information in the alphabet. First, we assign a label to the entry arrows I_1, I_2, I_3, I_4, corresponding, respectively, to the singularities $(YC)_2^{tg}$, $(Y\bar{C})_1^{tg}$, $(Y\bar{C})_1^{t-o}$, $(YC)_3^{tg}$. Then we introduce more information in the letters; i.e., we need a bigger alphabet. To do this, given a generic singularity on an NTAE, we want to

TABLE B *(second part).*

Letters→ states ↓	(π, D, γ^0)	(π, D, W^C)	(π, D, W^D)	(π, I, γ^0)	(π, I, W^C)	(π, I, W^D)
$(YC)_2^{tg}$	4	3	–	8	–	7
$(Y\bar{C})_1^{tg}$	–	13	–	24	23	22
$(Y\bar{C})_1^{t-o}$	12	10	11	–	–	–
$(YC)_3^{tg}$	4	3	–	8	7	–
1	–	9	–	20	19	18
2	2	1	–	6	–	5
3	–	13	–	24	23	22
4	4	3	–	8	7	–
5	16	14	15	–	–	–
6	4	3	–	–	–	–
7	–	9	–	–	–	–
8	2	1	–	–	–	–
9	–	9	–	20	19	18
10	–	9	–	20	19	18
11	12	10	11	–	–	17
12	2	1	–	6	–	5
13	–	13	–	24	23	22
14	–	13	–	24	23	22
15	16	14	15	–	–	21
16	4	3	–	8	7	–
17	16	14	15	–	–	–
18	16	14	15	–	–	–
19	–	–	–	–	13	–
20	4	3	–	–	–	–
21	12	10	11	–	–	–
22	12	10	11	–	–	–
23	–	9	–	–	–	–
24	2	1	–	–	–	–

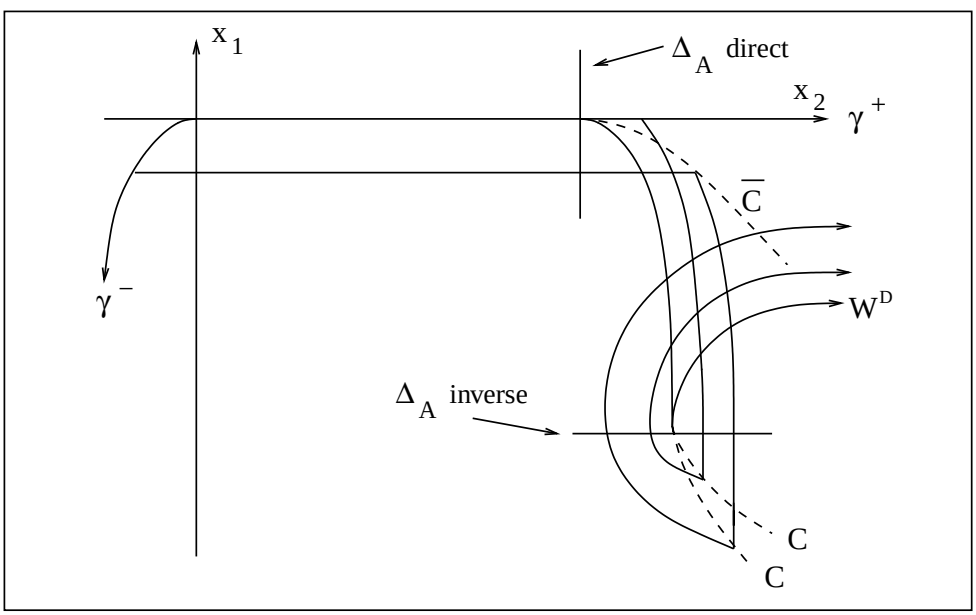


FIG. 5.2. *An example of a synthesis involving a* ribbon singularity.

include the following data relative to the subsequent singularity:

- $\Delta_A^{-1}(0)$ direct or inverse (indicated by D and I, resp.)

- the kind of exiting abnormal arc (i.e., γ^0, W^C or W^D).

In this way, the automaton $\mathcal{A}'$ is formed by $\mathbb{S}' = \mathbb{S}$, $\Sigma' = \{0, \pi\} \times \{D, I\} \times \{\gamma^0, W^C, W^D\} \cup \Sigma'_1$, where $\Sigma'_1 = \{I_1, I_2, I_3, I_4\}$ and set of edges $\mathbb{E}'$. Every element of Σ' (that is not in Σ'_1) is indicated by a triplet (. , . , .). In Table B the set of edges $\mathbb{E}'$ (with labels not in Σ'_1) is completely described. Notice that not all words of $(\Sigma')^*$ are recognizable by $\mathcal{A}'$. For example, the word $I_2(\pi, I, W^D)(0, D, W^C)$ is recognizable, and it corresponds to the sequence of singularities $(Y\bar{C})_1^{tg} \to 22 \to 14$, while the word $I_1(0, I, W^C)$ is not recognizable.

With these definitions, to every word recognizable by $\mathcal{A}'$ corresponds one and only one possible sequence of generic singularities along a NTAE. Theorem 2 is therefore proved.

We refer to Figure 5.2 for a graphic example of synthesis involving a *ribbon* singularity.

REFERENCES

[1] A. A. Agrachev and A. V. Sarychev, *Abnormal sub-Riemannian geodesics: Morse index and rigidity*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 13 (1996), pp. 635–690.

[2] A. A. Agrachev and A. V. Sarychev, *On abnormal extremals for Lagrange variational problems*, J. Math. Systems Estim. Control, 8 (1998), pp. 87–118.

[3] V. G. Boltyanskii, *Sufficient conditions for optimality and the justification of the dynamic programming method*, SIAM J. Control, 4 (1966), pp. 326–361.

[4] U. Boscain and B. Piccoli, *Projection singularities of extremals for planar systems*, Proceeding of the 38th IEEE Conference on Decision and Control, Phoenix, AZ, 1999, pp. 2936–2941.

[5] U. Boscain and B. Piccoli, *Extremal syntheses for generic planar systems*, J. Dynam. Control Systems, 7 (2001), pp. 209–258.

[6] U. Boscain and B. Piccoli, *Morse properties for the minimum time function on 2-d manifolds*, J. Dynam. Control Systems, 7 (2001), pp. 385–423.

[7] A. Bressan and B. Piccoli, *Structural stability for time–optimal planar syntheses*, Dynam. Contin. Discrete Impuls. Systems, 3 (1997), pp. 335–371.

[8] A. Bressan and B. Piccoli, *A generic classification of time-optimal planar stabilizing feedbacks*, SIAM J. Control Optim., 36 (1998), pp. 12–32.

[9] P. Brunovsky, *Existence of regular syntheses for general problems*, J. Differential Equations, 38 (1980), pp. 317–343.

[10] S. Eilenberg, *Automata, Languages, and Machines*, Vol. A, Academic Press, New York, 1974.

[11] S. Eilenberg, *Automata, Languages, and Machines*, Vol. B, Academic Press, New York, 1976.

[12] V. Jurdjevic, *Geometric Control Theory*, Cambridge University Press, Cambridge, UK, 1997.

[13] M. Kiefer and H. Schättler, *Cut-loci and cusp singularities in parametrized families of extremals*, in Optimal Control, Appl. Optim. 15, Kluwer, Dordrecht, 1998, pp. 250–277.

[14] M. Kiefer and H. Schättler, *Parametrized families of extremals and singularities in solution to the Hamilton–Jacobi–Bellman equation*, SIAM J. Control Optim., 37 (1999), pp. 1346–1371.

[15] B. Piccoli, *A Generic Classification of Time Optimal Planar Stabilizing Feedback*, Ph.D. thesis, SISSA, Trieste, Italy, 1994.

[16] B. Piccoli, *Regular time–optimal syntheses for smooth planar systems*, Rend. Sem Mat. Univ. Padova, 95 (1996), pp. 59–79.

[17] B. Piccoli, *Classification of generic singularities for the planar time-optimal synthesis*, SIAM J. Control Optim., 34 (1996), pp. 1914–1946.

[18] B. Piccoli and H. J. Sussmann, *Regular synthesis and sufficiency conditions for optimality*, SIAM J. Control Optim., 39 (2000), pp. 359–410.

[19] L. S. Pontryagin, V. Boltianski, R. Gamkrelidze, and E. Mitchtchenko, *The Mathematical Theory of Optimal Processes*, John Wiley and Sons, New York, 1961.

[20] H. Schättler, *On the local structure of time-optimal bang-bang trajectories in* $\mathbb{R}^3$, SIAM J. Control Optim., 26 (1988), pp. 186–204.

[21] H. Schättler, *The local structure of time optimal trajectories in dimension three under generic conditions*, SIAM J. Control Optim., 26 (1988), pp. 899–918.

[22] H. J. Sussmann, *The structure of time-optimal trajectories for single-input systems in the plane: The C^∞ nonsingular case*, SIAM J. Control Optim., 25 (1987), pp. 433–465.
[23] H. J. Sussmann, *Regular synthesis for time-optimal control of single-input real analytic systems in the plane*, SIAM J. Control Optim., 25 (1987), pp. 1145–1162.
[24] H. Whitney, *On singularities of mappings of Euclidean spaces.* I. *Mappings of the plane into the plane*, Ann. of Math. (2), 62 (1955), pp. 374–410.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 5, pp. 1358–1383

VISCOSITY SOLUTIONS OF THE BELLMAN EQUATION FOR EXIT TIME OPTIMAL CONTROL PROBLEMS WITH VANISHING LAGRANGIANS*

MICHAEL MALISOFF†

Abstract. We study the Hamilton–Jacobi–Bellman equation for undiscounted exit time optimal control problems for fully nonlinear systems and fully nonlinear singular Lagrangians using the dynamic programming approach. We prove a local uniqueness theorem characterizing the value functions for these problems as the unique viscosity solutions of the corresponding Hamilton–Jacobi–Bellman equations that satisfy appropriate boundary conditions. The novelty of this theorem is in the relaxed hypotheses on the lower bound on the Lagrangian and the very general assumptions on the target set. As a corollary, we show that the value function for the Fuller problem is the unique viscosity solution of the corresponding Hamilton–Jacobi–Bellman equation that vanishes at the origin and satisfies certain growth conditions. This implies as special cases first that the value function of this problem is the unique proper viscosity solution of the corresponding Hamilton–Jacobi–Bellman equation, in the class of all functions which are continuous in the plane and null at the origin, and second that this value function is the unique viscosity solution of that equation in a class which includes functions which are not bounded below. We also apply our results to the degenerate eikonal equation of geometric optics and to the shape-from-shading equations in image processing. Our theorem also applies to problems with noncompact targets and unbounded control sets whose Lagrangians take negative values.

Key words. viscosity solutions, dynamical systems, optimal control

AMS subject classifications. 49L25, 93C

PII. S0363012900368594

1. Introduction.

This paper studies Hamilton–Jacobi–Bellman equations (HJBEs) for a large class of unbounded optimal control problems for fully nonlinear systems having the form

$$\begin{cases} y'(t) \;=\; f(y(t),\alpha(t)), \qquad t \geq 0, \quad \alpha(t) \in A, \\ y(0) \;=\; x. \end{cases} \tag{1.1}$$

Our hypotheses will be such that (1.1) has a unique solution trajectory which is defined on $[0,\infty)$ for each input α and $x \in \mathbb{R}^N$. The optimal control problems are of the form

$$\text{Minimize} \;\; J(x,\alpha) \;\; \text{over} \;\; \alpha \in \mathcal{A}^f(x), \tag{1.2}$$

where $y_x(\cdot,\alpha)$ is the solution of (1.1) for each input $\alpha \in \mathcal{A} := \{$ measurable functions $[0,\infty) \to A\}$, $t_x(\alpha)$ is the infimum of those times t at which $y_x(t,\alpha)$ lies in a given

*Received by the editors February 28, 2000; accepted for publication (in revised form) May 26, 2000; published electronically January 9, 2002. This research was supported in part by NSF grant DMS95-00798. This work was done while the author was a Louis Bevier Graduate Fellow in the Department of Mathematics at Rutgers University. It includes part of the author's dissertation under the direction of Professor Héctor J. Sussmann. Some of this work was presented during the session "Optimal Control I" of the 38th IEEE Conference on Decision and Control in Phoenix, AZ on December 7, 1999, and published in preliminary form in the conference proceedings.
http://www.siam.org/journals/sicon/40-5/36859.html
†Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803-4918 (malisoff@math.lsu.edu).

closed set $\mathcal{T} \subseteq \mathbb{R}^N$, $\mathcal{A}^f(x)$ is the set of inputs α for which $t_x(\alpha) < \infty$, and

$$J(x,\alpha) := \int_0^{t_x(\alpha)} \ell(y_x(s,\alpha),\alpha(s))\, ds \;+\; g\left(y_x(t_x(\alpha),\alpha)\right) \quad \forall\, x \in \mathbb{R}^N \text{ and } \alpha \in \mathcal{A}^f(x).$$

We refer to the function ℓ as the Lagrangian (or instantaneous cost) of (1.2). We refer to $\mathcal{T}$ as a target, and thus (1.2) is a problem of minimizing the cost of reaching a target.

The value function of (1.2) will be denoted by v, and $\mathcal{R}$ denotes the set of all points x that can be brought to $\mathcal{T}$ in finite time using the evolution (1.1) and some input α in $\mathcal{A}$. Thus,

$$v(x) \;:=\; \begin{cases} \inf\left\{J(x,\alpha) :\; \alpha \in \mathcal{A}^f(x)\right\} & \text{if } x \in \mathcal{R}, \\ +\infty, & \text{otherwise.} \end{cases}$$

The class of problems we consider includes the Fuller problem (FP) and other well-known physical applications (cf. [14], [23], [26], and section 4), as well as problems for which the control set A and $\partial\mathcal{T}$ are both unbounded. The HJBE for (1.2) is

$$\sup_{a \in A} \left\{ -f(x,a) \cdot Du(x) \;-\; \ell(x,a) \right\} \;=\; 0, \tag{1.3}$$

which we wish to solve on $\Omega \setminus \mathcal{T}$, where Ω is a suitable open subset of $\mathbb{R}^N$. We will study (1.3) in the framework of viscosity solutions and relaxed controls (cf. [1] and [2]).

We will prove a local uniqueness theorem which characterizes the value functions for (1.2) as the unique viscosity solutions of the associated HJBEs on open sets of the form $\Omega \setminus \mathcal{T}$ that satisfy appropriate boundary and growth conditions. As a consequence, we show that the FP value function is the unique viscosity solution of the corresponding HJBE in the class of functions which are zero at the origin, are continuous in the plane, and satisfy a certain growth regularity condition. This regularity condition is a generalization of properness, i.e., of the condition $w(x) \to +\infty$ as $||x|| \to \infty$, which can also be satisfied by functions which are not bounded below (cf. Remark 2.8).

Uniqueness characterizations of this kind have been studied and applied by many authors for a large number of stochastic and deterministic optimal control problems and for differential games. Recent work in these areas may be found in [2], [9], and in the hundreds of references therein. For a detailed account of uniqueness questions for the HJBE for exit time problems with Lagrangians which are bounded below by positive constants on $\mathbb{R}^N \setminus \mathcal{T}$, see [5] and [16]. Also, uniqueness characterizations for HJBEs have been used to study the convergence of numerical schemes for approximating value functions and differential game values (cf. [4] and [19]) and also in the study of H^∞ control, singular perturbations, and much more.

However, Fuller's exit time problem is not covered by these results, since its Lagrangian ℓ vanishes at some points outside $\mathcal{T}$. For example, see [2], where the main comparison results for exit time problems require $\ell \geq m > 0$, and [16], where this requirement is relaxed to requiring[1] that for each $\varepsilon > 0$ there be a constant C_ε such that

$$\ell(x,a) \geq C_\varepsilon > 0 \qquad \forall a \in A \quad \text{and} \quad x \in \mathbb{R}^n \setminus B(\mathcal{T},\varepsilon). \tag{1.4}$$

[1] We set $\operatorname{dist}(p,S) := \inf\{||p-s|| : s \in S\}$ and $B(S,\varepsilon) := \{x \in \mathbb{R}^N :\ \operatorname{dist}(x,S) < \varepsilon\}$ for any $S \subseteq \mathbb{R}^N$, $p \in \mathbb{R}^N$, and $\varepsilon > 0$. For cases where $S = \{x\}$, we sometimes write $B_\varepsilon(x)$ instead of $B(\{x\},\varepsilon)$.

Also, the earlier results do not cover important Bellman equations for shape-from-shading problems and many eikonal equations from physics (cf. section 4). In fact, one easily finds HJBEs for optimal control problems with exit times that have several proper viscosity solutions when the lower bound requirement (1.4) is violated. For example, use the system

$$\dot{x}(t) \ = \ u(t) \ \in \ [-1,1],$$

choose

$$\ell(x,a) \ := \ (x+2)^2\,(x-2)^2\,x^2\,(x+1)^2\,(x-1)^2,$$

and set $g \equiv 0$. Let v_1 and v_2 denote the value functions for the associated problem (1.2) with the targets $\mathcal{T}_1 = \{0\}$ and $\mathcal{T}_2 = \{0, 2, -2\}$, respectively. One can easily check that v_1 and v_2 are both proper and that both are viscosity solutions of the associated HJBE (1.3) on $\mathbb{R} \setminus \mathcal{T}$ with $\mathcal{T} := \mathcal{T}_1$. Also, with the target $\mathcal{T} = \{0\}$, the problem satisfies all the hypotheses of the well-known theorems which characterize value functions of exit time control problems as the unique proper viscosity solutions of (1.3) in the class of all functions which are zero on $\mathcal{T}$, save for the fact that the positive lower bound requirement (1.4) on ℓ is not satisfied.

One may notice that the ingredient missing from this example is an optimal control analogue of the La Salle invariance condition, which we will specify below. This condition will guarantee that there are no spurious solutions for the HJBE for the FP. Roughly stated, the condition says that each trajectory of (1.1) "immediately leaves" any subset of the state space outside the target in which the running cost is null along some trajectory (cf. section 2). This condition would not be satisfied for the problem in the previous paragraph, since the constant trajectories at ± 2 accumulate zero costs. Our approach is therefore based on the properties of the trajectories $t \mapsto y_x(t,\alpha)$ of (1.1) at times close to zero. For a very different treatment of (1.3) based on the asymptotic behavior of trajectories as $t \to \infty$ and which characterizes value functions for exit time problems as the unique nonnegative viscosity solutions of the corresponding HJBEs that satisfy appropriate boundary conditions, see [17] and [18].

Our work is part of a larger research program which generalizes uniqueness results for viscosity solutions to versions that cover well-known optimal control problems whose dynamics do not necessarily admit unique trajectories for some choices of open loop controls and initial positions, or whose Lagrangians violate the usual boundedness requirements. For uniqueness characterizations for the Bellman equation for linear-quadratic problems, see [3] (which covers finite horizon cases) and [7] (which covers the infinite horizon case). A uniqueness characterization for (1.3) for exit time problems whose dynamics are continuous but not Lipschitz continuous, including Sussmann's reflected brachystochrone problem (cf. [21] and [22]), appears in [12]. For results for exit problems where f and ℓ are bounded under other special conditions that are generally not satisfied in the problems we consider below, see [25].

This paper is organized as follows. In section 2, we list assumptions on the data which will be in force in most of what follows, and we give definitions needed to specify a class of functions in which the value function v will be the unique viscosity solution. We also review the definitions of viscosity solutions and relaxed controls. In section 3, we state our main results, and section 4 shows how to apply them to a variety of physical problems, including the FP, geometric optics, and image processing. This is followed in section 5 with statements of certain converse dynamic

programming principles used to prove our uniqueness characterizations, a lemma on weak-$\star$ convergence of sequences of relaxed controls, and proofs of our main results. In section 6, we show how to extend these results so that they cover cases where the assumptions of section 2 are not satisfied, including cases where the control set A is unbounded and the Lagrangian ℓ takes negative values.

2. Assumptions and definitions. Let us make the following assumptions:

- (A_0) The control set A is a compact metric space.
- (A_1) The dynamics $f : \mathbb{R}^N \times A \to \mathbb{R}^N$ is continuous, and there exists a constant $L > 0$ such that $||f(x,a) - f(y,a)|| \leq L\,||x-y||$ for all $x, y \in \mathbb{R}^N$ and all $a \in A$.
- (A_2) The target $\mathcal{T} \subseteq \mathbb{R}^N$ is closed, $\mathcal{T} \neq \emptyset$, $g \in C(\mathcal{T})$, and g is bounded below.[2]
- (A_3) The Lagrangian $\ell : \mathbb{R}^N \times A \to \mathbb{R}$ is continuous.

The compactness requirement (A_0) is included to simplify the exposition. We will consider the case where A is noncompact separately in section 6.3 below. Roughly speaking, all of our results still hold when A is unbounded if we replace the inputs $\alpha \in \mathcal{A}$ with vector field valued inputs and add mild assumptions on the dynamics f and the Lagrangian ℓ. Alternatively, we can extend our results to cover problems with noncompact control sets by adding assumptions on f and ℓ from [3] that penalize "large" control values (cf. (A_4)–(A_5) below). We emphasize that, unlike in the usual treatments of exit time problems, $\partial\mathcal{T}$ is not required to be compact or smooth, f and ℓ need not be bounded, and ℓ need not be bounded below.

By an elementary application of the Schauder fixed point theorem (cf. [2]), conditions (A_0)–(A_1) will guarantee that for each input $\alpha \in \mathcal{A}$ and each point $x \in \mathbb{R}^N$ there is a unique solution $y_x(\cdot,\alpha)$ of the dynamics (1.1) which is defined on $[0,\infty)$. Moreover, the following estimates easily follow from (A_0)–(A_1):

- ($\mathcal{E}_1$) $\quad ||y_x(t,\alpha) - x|| \leq M_x\, t \ \text{ for all } \ t \in [0, 1/M_x]$,
- ($\mathcal{E}_2$) $\quad ||y_x(t,\alpha)|| \leq (||x|| + \sqrt{2Kt})e^{Kt} \ \text{ for all } \ t \geq 0 \ \text{ and } \ \alpha \in \mathcal{A}$,

where $M_x := \sup\{||f(z,a)|| : z \in B_1(x), a \in A\}$ when this supremum is nonzero, $M_x = 1$ otherwise, and $K := \sup\{||f(0,a)|| : a \in A\}$. (Recall from the footnote in section 1 that $B_1(x)$ is the open unit ball centered at x.) The preceding estimates are shown in [2].

Since A is a compact metric space, we can view our controls $\alpha \in \mathcal{A}$ as members of the larger class $\mathcal{A}^r$ of relaxed controls (cf. [2] and [24]). Thus,

$$\mathcal{A}^r := \{ \text{ measurable functions } [0,\infty) \to A^r \},$$

where A^r is the set of Radon probability measures on A (i.e., the probability measures on the smallest σ-algebra on A containing all the open sets of A). We topologize A^r as a subset of the dual of $C(A)$ with the topology of weak-$\star$ convergence. By identifying $\mathcal{A}$ with the set of Dirac probability measures on A (i.e., probability measures that put weight one on a single point in the control set at each time), we will view $\mathcal{A}$ as a subset of $\mathcal{A}^r$. The topology on $\mathcal{A}^r$ will be such that each sequence in $\mathcal{A}^r$ has a weak-$\star$ convergent subsequence in $\mathcal{A}^r$ (cf. Lemma 5.3). By $\mathcal{A}^r \ni \alpha_n \to \bar{\alpha} \in \mathcal{A}^r$ weak-$\star$ we mean

[2] We let $C(U)$ denote the set of continuous real-valued functions on any space U. For cases in which U is an open subset of a Euclidean space, $C^1(U)$ denotes the set of continuous real-valued functions on U with one continuous derivative.

$$\int_0^\tau \int_A (h(s))(a)\, d(\alpha_n(s))(a)\, ds \quad \to \quad \int_0^\tau \int_A (h(s))(a)\, d(\bar{\alpha}(s))(a)\, ds \quad \text{as} \quad n \to \infty \tag{2.1}$$

for each Lebesgue integrable function $h : [0, \tau] \to C(A)$ and each $\tau > 0$.

Define $\ell^r : \mathbb{R}^N \times A^r \to \mathbb{R}$ and $f^r : \mathbb{R}^N \times A^r \to \mathbb{R}^N$ by

$$\ell^r(x, m) := \int_A \ell(x, a)\, dm(a) \quad \text{and} \quad f^r(x, m) := \int_A f(x, a)\, dm(a). \tag{2.2}$$

One checks that f^r and ℓ^r satisfy (A_1) and (A_3), respectively, and that A^r is compact. We leave the details to the reader (cf. [2]). We can therefore define $y_x^r(\cdot, \alpha) : [0, \infty) \to \mathbb{R}^N$ to be the unique trajectory of $y'(s) = f^r(y(s), \alpha(s))$ starting at x for each $\alpha \in \mathcal{A}^r$. We will refer to $y_x^r(\cdot, \alpha)$ as a relaxed trajectory. Define the *Hamiltonian* $H : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ by

$$H(x, p) = \sup_{a \in A} \{ -f(x, a) \cdot p - \ell(x, a) \} \tag{2.3}$$

and set

$$P := \left\{ x \in \mathbb{R}^N : \int_0^t \ell^r(y_x^r(s, \alpha), \alpha(s)) ds > 0 \quad \forall\ t \in (0, \infty],\ \alpha \in \mathcal{A}^r \right\},$$

where we allow the integral to diverge to $+\infty$ in the definition of P. We sometimes write H_A instead of H to emphasize the control set. Notice that assumptions (A_0)–(A_3) guarantee that $H(x, p)$ is always a finite supremum. Notice also that we could have $P \supseteq \mathcal{R} \setminus \mathcal{T}$ even if $\ell(x, a) = 0$ for some $x \in \mathcal{R} \setminus \mathcal{T}$ and $a \in A$, since the dynamics (1.1) could be such that points in the state space that give zero values for the Lagrangians are "immediately" brought out of the null sets of the vector fields $\ell(\cdot, a)$, producing positive running costs. This will be the case for the FP (cf. section 4.1). For this problem, as well as for the other examples we give below, condition (1.4) is violated, but $P \supseteq \mathcal{R} \setminus \mathcal{T}$ (cf. section 4).

We will prove that for suitable open sets S, there is at most one viscosity solution $w \in C(S)$ of HJBE (1.3) on $S \setminus \mathcal{T}$ that satisfies a certain growth property and a certain compatibility property (cf. Definition 2.1 and Definition 2.6). In particular, we give uniqueness characterizations for viscosity solutions $w \in C(\mathcal{R})$ of (1.3) on $\mathcal{R} \setminus \mathcal{T}$. For the case of the FP, these side conditions reduce to properness or some generalized notion of properness (cf. Remark 2.7 and section 6). In particular, we show that the value function for the FP is the unique viscosity solution of the corresponding HJBE in a class which includes functions which are not bounded below. We also apply our results to degenerate eikonal equations and to the equations of shape-from-shading problems (cf. sections 4.2–4.3).

We use the following definition of viscosity solutions.

DEFINITION 2.1. *Let $\mathcal{G} \subseteq \mathbb{R}^N$ be open, let $S \supseteq \mathcal{G}$, let $F : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ be continuous, and let $w : S \to \mathbb{R}$.*

1. *We call w a* (semicontinuous) viscosity supersolution *of $F(x, Dw(x)) = 0$ on $\mathcal{G}$ if* (i) *w is lower semicontinuous on $\mathcal{G}$ and* (ii) *for all $\gamma \in C^1(\mathcal{G})$ and all local minima $x_o \in \mathcal{G}$ of $w - \gamma$ we have $F(x_o, D\gamma(x_o)) \geq 0$.*
2. *We call w a* (semicontinuous) viscosity subsolution *of $F(x, Dw(x)) = 0$ on $\mathcal{G}$ if* (i) *w is upper semicontinuous on $\mathcal{G}$ and* (ii) *for all $\lambda \in C^1(\mathcal{G})$ and all local maxima $x_1 \in \mathcal{G}$ of $w - \lambda$, we have $F(x_1, D\lambda(x_1)) \leq 0$.*

3. *We call w a* viscosity solution *of $F(x, Dw(x)) = 0$ on $\mathcal{G}$ if it is simultaneously a (semicontinuous) viscosity supersolution and a (semicontinuous) viscosity subsolution of this equation on $\mathcal{G}$.*

This definition of viscosity solutions is elementarily equivalent to the definition of viscosity solutions based on super- and subdifferentials used in [7] (cf. [2]). Also, by a denseness argument, the above definition is equivalent to the one obtained by replacing $C^1(\mathcal{G})$ with $C^\infty(\mathcal{G})$ (i.e., infinitely differentiable functions on $\mathcal{G}$). Most of what follows is based on the viscosity solution definition given above, so we omit the definition based on semidifferentials. We also use the following notions of quick growth and boundary levelness.

DEFINITION 2.2. *Let $S \subset \mathbb{R}^N$ be open, let $w : S \to \mathbb{R}$ be continuous, and let $\omega_o \in \mathbb{R} \cup \{+\infty\}$.*

1. *We say w is* quickly growing *with respect to S and write $w \in QG(S)$, provided the following:*
   $QG(S)$ *For each pair (x, y) with $x \in S$ and $y \in \partial S$, there exists an $\varepsilon > 0$ such that if $p \in S$ and if $|| p - y || < \varepsilon$, then $w(x) < w(p)$.*
2. *We say w is* boundary level *for S and ω_o and write $w \in BC_{\omega_o}(S)$, provided the following:*
   $BC_{\omega_o}(S)$ *For each $x \in S$, we have $w(x) < \omega_o$. In addition, for each $x_o \in \partial S$, we have $w(x) \to \omega_o$ as $S \ni x \to x_o$.*

We will show that if $w \in QG(\Omega)$ is a viscosity subsolution of (1.3) on $\Omega \setminus \mathcal{T}$, where Ω is an open set containing $\mathcal{T}$, and if $\Omega \setminus \mathcal{T} \subseteq P$, then $w \le v$ on $\Omega \setminus \mathcal{T}$ (but see section 6.4 for cases where $\Omega \setminus \mathcal{T} \not\subseteq P$). This is done in Proposition 5.4. This inequality is an easy consequence of the dynamic programming principle for cases where $\Omega = \mathcal{R}$ but nontrivial for other cases, because we must then consider trajectories which reach $\mathcal{T}$ in finite time but exit Ω before reaching $\mathcal{T}$.

Notice that condition $QG(S)$ is weaker than the standard requirement $BC_{\omega_o}(S)$ in the known uniqueness characterizations for cases where S is bounded (cf. [2], [5], and [16]). Notice also that the condition $QG(S)$ does not require the function w to be defined on ∂S, while for $\omega_o \neq +\infty$, $BC_{\omega_o}(S)$ holds if S is the ω_o-sublevel set of w and w is continuous on $\bar{S}$. In some of what follows, we will take $S = \mathcal{R} = \mathbb{R}^N$ and $\omega_o = +\infty$, in which case $QG(S)$ and $BC_{\omega_o}(S)$ are satisfied vacuously. We will do this when we consider Fuller's example.

We will be especially interested in viscosity solutions on sets which can be decomposed into smaller sets with "controllable boundaries." This controllability condition will guarantee the nondegeneracy of the Hamiltonian H, which is needed for our uniqueness characterizations. The controllability condition we need is as follows.

DEFINITION 2.3. *Let $\mathcal{O} \subseteq \mathbb{R}^N$ and assume* (A_0)–(A_3) *are satisfied. We say that $\mathcal{O}$ satisfies the* strong small time control condition *with respect to $\mathcal{T}$ and write SSTC($\mathcal{O}$, $\mathcal{T}$) if there is a sequence of bounded open sets $\{\Omega_j\}$ such that the following conditions hold:*

($SSTC_1$) The Ω_j's increase to $\mathcal{O}$, i.e., $\Omega_j \subseteq \Omega_{j+1}$ for all j and $\mathcal{O} = \bigcup_{j=1}^{\infty} \Omega_j$.

($SSTC_2$) If $T_j : \Omega_j \setminus \mathcal{T} \to \mathbb{R} \cup \{+\infty\}$ is defined by
$$T_j(x) := \inf\{t : y_x(t, \alpha) \in \partial(\Omega_j \setminus \mathcal{T}),\ \alpha \in \mathcal{A}\} \quad \text{for} \quad j = 1, 2, \ldots,$$
and if $\Omega_j \setminus \mathcal{T} \ni x \to x_0 \in \partial(\Omega_j \setminus \mathcal{T})$, then $T_j(x) \to 0$.

When these conditions hold, $\{\Omega_j\}$ is called the associated controllability sequence.

The following example illustrates the SSTC requirements.

Example 2.4. Take $N = 2$, $A = [-1, +1]$, $f(x, y, a) = (y, a)$, and $\mathcal{T} = \{\vec{0}\} \subseteq \mathbb{R}^2$. As we will show below, we can then satisfy SSTC($\mathbb{R}^2 \setminus \{0\}, \{0\}$) by taking Ω_j to be the open set bounded by the trajectory from $(0, j)$ to $(0, -j)$ using the constant control $a \equiv -1/2$, followed by the trajectory back to $(0, j)$ using the constant control $a \equiv +1/2$ with the origin removed. If we now choose $\ell(x, y, a) = x^2$, then we get the usual FP. We discuss the FP more fully in section 4.1, where we prove a uniqueness characterization for the FP Bellman equation.

Remark 2.5. Uniqueness of viscosity solutions requires nondegeneracy of the Hamiltonian, i.e., extra assumptions on the dynamics, which is implied by SSTC. Using proximal normals, we can characterize this nondegeneracy in terms of $f(x, a)$ and the boundary of the domain, as follows. Since the boundaries $\partial(\Omega_j \setminus \mathcal{T})$ in Definition 2.3 may not be smooth, we use generalized exterior normals (cf. [2]). Recall that for any closed nonempty set $S \subset \mathbb{R}^N$, the *(generalized) exterior normal* to S at $z \in \partial S$, which is denoted by $N(z)$, is defined to be the set of all unit vectors ν for which there exists $x \notin S$ satisfying

$$x = z + \operatorname{dist}(x, S)\nu \qquad \text{and} \qquad \{z\} = \mathcal{P}(x),$$

where $\mathcal{P}(x) := \{z \in \partial S : \operatorname{dist}(x, S) = ||x - z||\}$. Condition $(SSTC_2)$ is then satisfied if the following holds:

$(SSTC_3)$ For each $j \in \mathbb{N}$ and $\bar{x} \in \partial(\Omega_j \setminus \mathcal{T})$ there exist $s > 0$ and $\delta > 0$ such that

$$\inf_{a \in A} f(\bar{x}, a) \cdot \nu \ \leq \ -\delta$$

for any $\nu \in N(x)$ with $x \in \partial(\Omega_j \setminus \mathcal{T}) \cap B_s(\bar{x})$

(cf. [2, Chapter IV]). This of course reduces to a standard "outer field" condition in terms of outer normals when the sets $\Omega_j \setminus \mathcal{T}$ have piecewise smooth boundaries. Also, for the special case in which $\Omega_j \equiv \mathcal{O}$, we have nondegeneracy in terms of $f(x, a)$ and the boundary $\partial(\mathcal{O} \setminus \mathcal{T})$ of the domain.

One can also express the nondegeneracy in terms of the standard small-time controllability condition from [15]. Recall that for any $\mathcal{U} \subseteq \mathbb{R}^N$, $STC\mathcal{U}$ is defined to be the requirement that $\mathcal{U} \subset \text{Interior}\, \mathcal{R}(\varepsilon)$ for all $\varepsilon > 0$, where

$$\mathcal{R}(\varepsilon) \ := \ \left\{x \ \in \ \mathbb{R}^N : \ \exists\, t \ \in \ [0, \varepsilon) \ \text{and} \ \alpha \ \in \ \mathcal{A} \ \ \text{s.t.} \ \ y_x(t, \alpha) \ \in \ \mathcal{U}\right\}.$$

The condition $STC\mathcal{U}$ is satisfied when suitable assumptions are made on the directions of the vector field f and of its Lie brackets at $\partial\mathcal{U}$ (cf. [20]). Condition $(SSTC_2)$ is then satisfied if the following holds:

$(SSTC_4)$ The condition $\text{STC}(\Omega_j^c) \wedge \text{STC}\mathcal{T}$ holds for $j = 1, 2, \ldots$.

Condition $(SSTC_4)$ says that the dynamics (1.1) can be controlled to $\partial\mathcal{T}$ or $\partial(\Omega_j)$ for any $j \in \mathbb{N}$. This condition is stronger than $(SSTC_2)$ for cases in which $\mathcal{T} \cap \partial\Omega_j \neq \emptyset$ for some j, since $(SSTC_2)$ does not require controllability to the portions of $\partial(\Omega_j)$ in $\mathcal{T}$.

Finally, we use the following generalized notion of properness.

Definition 2.6. *Assume* (A_0)–(A_3) *and SSTC($\mathcal{O}$, $\mathcal{T}$). Let* $\{\Omega_j\}$ *denote the associated controllability sequence, and let* $w : \mathcal{O} \to \mathbb{R}$ *and* $\omega_o \in \mathbb{R} \cup \{+\infty\}$ *be given. We say that* w *is* $(\mathcal{O}, \mathcal{T}, \omega_o)$-compatible *if the following conditions hold:*

1. $\partial(\Omega_j) \setminus \mathcal{T} \subseteq \mathcal{O}$ *for all* j*;*
2. $\lim_{j \to \infty} \min\{w(p) : \ p \in \partial(\Omega_j) \setminus \mathcal{T}\} = \omega_o$.

Remark 2.7. The preceding compatibility conditions can be easily verified when $\mathcal{T} = \{0\}$. For example, any proper function is $(\mathbb{R}^N \setminus \{0\}, \{0\}, +\infty)$-compatible if $STC\{0\}$ and $STC(\Omega_j^c)$ hold for an increasing sequence Ω_j of bounded open sets that satisfy $\min\{||x|| : x \in \partial(\Omega_j), x \neq 0\} \to \infty$ and whose union is $\mathbb{R}^N \setminus \{0\}$. However, $(\mathcal{O}, \mathcal{T}, \omega_o)$-compatibility is a much more general properness condition, since it allows functions which are not proper and cases where $\mathcal{O}$ is bounded.

Remark 2.8. Notice for future reference that $(\mathcal{O}, \mathcal{T}, \omega_o)$-compatibility does not put any restrictions on the behavior of compatible functions at points which are not on the boundaries of any of the sets Ω_j in the controllability sequences. In particular, we allow $(\mathcal{O}, \mathcal{T}, +\infty)$-compatible functions which are not bounded below. We will generally be concerned with functions that are compatible in the sense of the preceding definition. (But see section 6.2 for analogues of our main results, which characterize the value function of (1.2) as a unique viscosity solution of (1.3) in a class of functions satisfying an even more general notion of properness than the one given by Definition 2.6.)

3. Statements of main results. We will prove the following theorem.

THEOREM 3.1. *Assume that* (1.2) *satisfies* (A_0)–(A_3), *that* $\Omega \subseteq \mathbb{R}^N$ *is an open set containing* $\mathcal{T}$, *that* $\omega_o \in \mathbb{R} \cup \{+\infty\}$, *and that* $w \in BC_{\omega_o}(\Omega)$ *is a viscosity solution of* [3]

$$\begin{cases} H(x, Dw(x)) \;=\; 0, & x \in \Omega \setminus \mathcal{T}, \\ w(x) \;=\; g(x), & x \in \mathcal{T}. \end{cases} \tag{3.1}$$

Assume $\Omega \setminus \mathcal{T} \subseteq P$ *and that* w *is* $(\Omega \cap P,\ \mathcal{T}, \omega_o)$*-compatible. Then* $w \equiv v$ *on* Ω.

Remark 3.2. The hypothesis that $w \in BC_{\omega_o}(\Omega)$, which is unnecessarily strong, is used to simplify the statement of the theorem. As we will show in section 5.2, Theorem 3.1 remains true if we weaken the assumption that $w \in BC_{\omega_o}(\Omega)$ to the assumptions that $w < \omega_o$ on Ω and $w \in QG(\Omega)$ (cf. Definition 2.2).

Remark 3.3. Let us compare the assumptions of Theorem 3.1 to those of the previously known uniqueness characterizations for exit time problems, and let us show how the assumptions of the theorem can easily be verified. As we explained in the introduction, the earlier results assume (1.4), i.e., positive lower bounds on the Lagrangian ℓ outside neighborhoods of the target. This is much more restrictive than requiring $\Omega \setminus \mathcal{T} \subseteq P$. For example, it could be that ℓ depends only on the state, and $\int_0^t \ell(y_x(s,\beta))\,ds > 0$ even though $\ell(y_x(s,\beta)) = 0$ for some values $s \in [0,t]$. This will be the case for the FP (cf. Example 2.4 and section 4.1), since the Lagrangian vanishes along the y-axis but all the trajectories starting outside $\vec{0}$ and running for positive time give positive running costs. The condition $\Omega \setminus \mathcal{T} \subseteq P$ is a control analogue of the La Salle invariance principle, and as we show in Remark 4.3, this condition cannot be omitted from the theorem. The condition $\Omega \setminus \mathcal{T} \subseteq P$ also allows cases where ℓ is always positive, or null only for state space points in the target, but still violates (1.4) because $\liminf_{||x|| \to \infty} \ell(x,a) = 0$ for certain control values a while $\mathcal{T}$ is bounded. This will be the case for the shape-from-shading and degenerate eikonal equations (cf. sections 4.2–4.3). As discussed in Remark 2.7, condition $BC_{\omega_o}(\Omega)$ is standard (cf. [2], [5], and [16]), and $(\mathbb{R}^N \setminus \mathcal{T}, \mathcal{T}, +\infty)$-compatibility holds for any proper continuous function as long as $STC\mathcal{T}$ and $STC(\Omega_j)$ hold along a suitable increasing sequence Ω_j

[3] By a viscosity solution of (3.1) we will mean a viscosity solution of $H(x, Dw(x)) = 0$ on $\Omega \setminus \mathcal{T}$ which equals g on $\mathcal{T}$. Notice that since we assume $\mathcal{T} \subset \Omega$, the target condition $w(x) = g(x)$ in (3.1) holds on all of $\mathcal{T}$.

of bounded open sets. This observation is the basis for our uniqueness result for the FP.

In the context of Theorem 3.1, we of course have $\Omega \subseteq \{x \in \mathbb{R}^N : v(x) < \omega_o\}$, since $w \equiv v$ on Ω. Also, while the condition $\Omega \setminus \mathcal{T} \subseteq P$ of the theorem implies that

$$\int_0^t \ell(y_p(s,\alpha),\alpha(s))\,ds \geq 0 \quad \forall \alpha \in \mathcal{A},\ t \geq 0, \ \text{ and } \ p \in \Omega \setminus \mathcal{T}, \tag{3.2}$$

this nonnegativity can fail for trajectories starting outside Ω, since (1.4) is not assumed. This suggests that a slight strengthening of the condition $\Omega \setminus \mathcal{T} \subseteq P$ to require nonnegativity of costs for trajectories running *outside* Ω might guarantee that $\Omega = \{x \in \mathbb{R}^N : v(x) < \omega_o\}$. This motivates the following corollary.

COROLLARY 3.4. *Under the assumptions of Theorem* 3.1*, with ℓ nonnegative,* $\Omega = \{x \in \mathbb{R}^N : v(x) < \omega_o\}$.

Remark 3.5. As we will show in the proof of Corollary 3.4, the assumption that ℓ is nonnegative can be relaxed to the following requirement:

($\sharp$) If $p \in \mathbb{R}^N$, and if $\alpha \in \mathcal{A}$ and $t > 0$ are such that $y_p(t,\alpha) \in \partial\Omega$, then $\int_0^t \ell(y_p(s,\alpha),\alpha(s))\,ds \geq 0$.

This is a strengthening of (3.2), since ($\sharp$) requires the nonnegativity of costs for trajectories starting outside of $\Omega \setminus \mathcal{T}$, while (3.2) requires only the nonnegativity of these costs for initial values on $\Omega \setminus \mathcal{T}$. Of course, ($\sharp$) may hold even if (1.4) fails (cf. the examples in section 4 below). Corollary 3.4 extends the local uniqueness characterizations based on (1.4) (cf. [2, Chapter 2] and [16]), which prove the uniqueness of viscosity solutions of HJBE (1.3) on neighborhoods Ω of $\mathcal{T}$ and which then characterize the neighborhoods containing proper viscosity solutions as sublevel sets of v. We will prove Corollary 3.4 in section 5.2.

The following corollary will be an immediate consequence of the fact that the value function v of (1.2) is a viscosity solution of the HJBE (1.3) on $\mathcal{R} \setminus \mathcal{T}$.

COROLLARY 3.6. *Assume* (A_0)–(A_3)*, $\mathcal{R}$ open, and $v \in QG(\mathcal{R})$. If $\mathcal{R} \setminus \mathcal{T} \subseteq P$ and v is $(\mathcal{R} \cap P, \mathcal{T}, +\infty)$-compatible, then v is the unique viscosity solution of*

$$\begin{cases} H(x, Dw(x)) \ = \ 0, & x \in \mathcal{R} \setminus \mathcal{T}, \\ w(x) \ = \ g(x), & x \in \mathcal{T}, \end{cases} \tag{3.3}$$

in the class of all $(\mathcal{R} \cap P, \mathcal{T}, +\infty)$-compatible functions $w \in QG(\mathcal{R})$.

The hypotheses of this corollary are especially easy to check when $\mathcal{R} = \mathbb{R}^N$, $\mathcal{T} = \{0\}$, and v is proper, since then $QG(\mathcal{R})$ holds vacuously and it suffices to check $SSTC(P)$ for a sequence Ω_j satisfying $\min\{||x|| : x \in \partial(\Omega_j), x \neq 0\} \to \infty$ and $\partial\Omega_j \setminus \mathcal{T} \subseteq P$ to get the uniqueness characterization.

Remark 3.7. The novelty of these results is that we do not need to assume that the Lagrangian ℓ satisfies condition (1.4) or that ℓ is even bounded below, that $\mathcal{T}$ can be very general, and that the uniqueness characterization is within a class which contains functions which are not bounded below. For the special cases in which $\omega_o = +\infty$, Corollary 3.4 characterizes the reachability set $\mathcal{R}$ as the unique set Ω for which there exists a function w satisfying the hypotheses of the theorem. By replacing the inputs $t \mapsto \alpha(t)$ with vector field valued inputs $t \mapsto (f(\cdot,\alpha(t)), \ell(\cdot,\alpha(t)))$ and adding standard assumptions on the input choice map $a \mapsto f(\cdot,a) \times \ell(\cdot,a)$, we can generalize Theorem 3.1 so that it also covers exit time problems with closed but not necessarily compact control sets (cf. section 6.3). One can also relax the assumption $\Omega \setminus \mathcal{T} \subseteq P$ and allow problems where ℓ takes negative values (cf. section 6.4).

4. Physical applications. Before proving our main results, we show how they apply to physical problems. Our uniqueness results apply to a variety of first-order equations which can be expressed in the HJBE form (1.3) but which are not tractable using the uniqueness results requiring (1.4). These include the Bellman equation of the FP (cf. [26]), degenerate eikonal equations from geometric optics (cf. [17]), and shape-from-shading problems in image processing (cf. [11]). This section discusses these applications.

4.1. The Fuller problem. We explain how the results of section 3 give uniqueness characterizations for the HJBE of the FP. Recall that the FP (with exponent q) is

$$\text{Infimize} \quad \int_0^{t_p(\alpha)} |y_{1,p}(t,\alpha)|^q \, dt \quad \text{over all} \quad \alpha \in \mathcal{A}^f(p) \tag{4.1}$$

for each $p \in \mathbb{R}^2$, where $t \mapsto y_p(t,u) := (y_{1,p}(t,u), y_{2,p}(t,u))'$ is defined to be the trajectory of

$$\begin{cases} \dot{x}(t) \;=\; y(t), \;\; \dot{y}(t) \;=\; u(t) \;\in\; A := [-1,+1], \\ (x(0),y(0))' \;=\; p, \end{cases}$$

and $t_p(u)$ is the first time this trajectory reaches the target $\mathcal{T} := \{0\} \subseteq \mathbb{R}^2$. (We use 0 to denote $0 \in \mathbb{R}$ or $\vec{0} \in \mathbb{R}^2$.) We allow any $q > 1$, but we fix q in what follows. From [26], we know that (4.1) has an optimal control for each initial state. In particular, $\mathcal{R} = \mathbb{R}^2$. We will verify the hypotheses of Corollary 3.6.

We first verify condition $SSTC(\mathbb{R}^2 \setminus \{0\}, \{0\})$. Let $\eta > 0$ be given. The FP trajectory ϕ_1 from $(0,\eta)'$ using the constant control $u \equiv -1/2$ reaches the point $(0,-\eta)'$ at time 4η. The FP trajectory ϕ_2 from $(0,-\eta)'$ using $u \equiv 1/2$ reaches $(0,\eta)'$ at time 4η. Let ζ_η denote the concatenation of the corresponding restrictions $\phi_1 : [0,4\eta] \to \mathbb{R}^2$ followed by $\phi_2 : [0,4\eta] \to \mathbb{R}^2$, and let G_η denote the open set bounded by this concatenation with the origin removed. By elementary calculations of trajectories, one sees that $STC(G_\eta^c)$ holds for each $\eta > 0$. The calculation is based on the fact that ϕ_1 solves $x = \eta^2 - y^2$, and ϕ_2 solves $x = y^2 - \eta^2$. From [15], we know that $STC(\{0\})$ holds. Since the G_η's engulf $\mathbb{R}^2 \setminus \{0\}$, it follows from Remark 2.5 that condition SSTC($\mathbb{R}^2 \setminus \{0\}, \{0\}$) holds.

For the case of the FP, we have $\int_0^t |y_{1,p}^r(s,\alpha)|^q \, ds > 0$ for all $\alpha \in \mathcal{A}^r$, $p \neq 0$, and $t > 0$, since $\dot{x} \equiv y$ is continuous along any FP trajectory (which implies that $|x(t)| > 0$ for a positive measure of small times whenever $y(0) \neq 0$). This establishes the condition $\mathcal{R} \setminus \mathcal{T} \subseteq P$. Next we verify that the FP value function v is $(\mathbb{R}^2 \setminus \{0\}, \{0\}, +\infty)$-compatible, which will be a consequence of the fact that v is proper (i.e., that it satisfies $\lim_{||x|| \to \infty} v(x) = +\infty$). The fact that v is proper is an elementary consequence of the dilation symmetries of the problem. Indeed, suppose

$$\lim_{m \to \infty} ||(x_m, y_m)'|| = \infty \quad \text{and} \quad v((x_n,y_n)') \leq M < \infty \quad \forall\, n.$$

For large n, pick a $\lambda_n > 0$ so that

$$\Lambda_n \;:=\; (\lambda_n^2 x_n, \lambda_n y_n)' \;\in\; \partial\,[B_1(0)]\,, \tag{4.2}$$

and thus $\lambda_n \to 0$. Assume $t \mapsto (x_n(t), y_n(t))'$ is an optimal FP trajectory for the initial point $(x_n, y_n)'$. Since $t \mapsto (\lambda_n^2 x_n(t/\lambda_n), \lambda_n y_n(t/\lambda_n))'$ is optimal for the initial

point Λ_n for each n (cf. [26]), it follows that

$$M \geq \int_0^\infty |x_n(t)|^q \, dt = \int_0^\infty \left|\lambda_n^2 \, x_n(u/\lambda_n)\right|^q \frac{1}{\lambda_n^{1+2q}} du = \frac{v(\Lambda_n)}{\lambda_n^{1+2q}} \quad \forall n. \tag{4.3}$$

Since $\partial[B_1(0)]$ is compact and v is continuous and positive definite, it follows that v is bounded away from 0 on $\partial[B_1(0)]$. To show that v is continuous, notice that v is strictly convex on $\mathbb{R}^2$ (cf. [26] and p. 81 of [13]), which follows from the fact that the FP admits optimal controls for each initial position and the convexity of $x \mapsto |x|^q$. Since (4.3) would mean that $v(\Lambda_n) \to 0$ as $n \to \infty$, we arrive at a contradiction with (4.2). Therefore, the FP value function is proper. Since the minimum norm of any point on $\partial(G_\eta) \setminus \{0\}$ increases without bound as $\eta \to +\infty$, it follows that the FP value function is $(\mathbb{R}^2 \setminus \{0\}, \{0\}, +\infty)$-compatible (with the controllability sequence defined by $\Omega_j = G_j$). Corollary 3.6 therefore gives the following.

COROLLARY 4.1. *For each $q > 1$, the value function v_q for the FP with exponent q is the unique viscosity solution of the Bellman equation*

$$-y(Dw((x,y)'))_1 + |(Dw((x,y)'))_2| - |x|^q = 0 \tag{4.4}$$

on $\mathbb{R}^2 \setminus \{0\}$ in the class of proper functions $w \in C(\mathbb{R}^2)$ that satisfy $w(0) = 0$. Moreover, v_q is the unique $(\mathbb{R}^2 \setminus \{0\}, \{0\}, +\infty)$-compatible viscosity solution $w \in C(\mathbb{R}^2)$ of (4.4) on $\mathbb{R}^2 \setminus \{0\}$ that vanishes at 0.

Remark 4.2. Since the FP Lagrangians ℓ vanish outside $\mathcal{T}$, this result does not follow from the earlier uniqueness results for viscosity solutions of HJBEs. Notice also that Corollary 3.6 actually gives a uniqueness characterization for (4.4) in a class containing functions which are not bounded below, as explained in Remark 2.8. Moreover, the FP value function is the unique viscosity solution of (4.4) on $\mathbb{R}^2 \setminus \{0\}$ in the class of functions $w \in C(\mathbb{R}^2)$ which are null at 0 and satisfy the even more general regularity condition (REG) of section 6.2 with $\Omega = \mathbb{R}^2$.

4.2. Degenerate eikonal equations. Our results also apply to equations from physics which do not arise as exit time problem HJBEs but which can be expressed in that form. For example, in geometric optics, the propagation of light is described by equations of the form

$$\sum_{i,j=1}^N a_{i,j}(x) \, u_{x_i}(x) \, u_{x_j}(x) + \sum_{i=1}^N 2 \, b_i(x) \, u_{x_i}(x) - h^2(x) = 0, \tag{4.5}$$

where $a_{i,j} = \sum_{k=1}^N \sigma_{i,k}\sigma_{j,k}$ and $\sigma = [\sigma_{i,j}]$ is a symmetric matrix. Such equations are called *degenerate eikonal equations*, and the viscosity solutions of (4.5) have been studied and applied extensively in physics (cf. [6]). The function $h : \mathbb{R}^N \to (0,1)$ represents the refraction index of the medium. Under the standard assumption that $b(x) = 2\sigma(x)\bar{b}(x)$ for some vector field $\bar{b}$, (4.5) is equivalent to

$$\left|\left|\sigma(x) \, Du(x) + \bar{b}(x)\right|\right| = \left[h^2(x) + \left|\left|\bar{b}\right|\right|^2(x)\right]^{1/2}.$$

This includes as a special case

$$\sup_{||a||=1} \{ -a \cdot Du(x) - |h(x)| \} = 0, \tag{4.6}$$

which is an HJBE for an exit problem for any closed target $\mathcal{T}$ with the dynamics $f(x,a) = a \in \partial B_1(0)$ and the Lagrangian $\ell(x,a) = |h(x)|$. (Recall from section 1 that $B_\varepsilon(p) := \{x \in \mathbb{R}^M : ||x-p|| < \varepsilon\}$ for all $\varepsilon > 0$ and $p \in \mathbb{R}^M$.) Equation (4.6) is satisfied by the travel time in geometric optics if $1/|h(x)|$ is the speed of the medium. Now take $\mathcal{T}$ to be bounded and choose any continuous refraction index h satisfying (i) $h(x) = \tilde{h}(||x||)$ for some decreasing function $\tilde{h} : \mathbb{R}_+ \to (0,1)$, (ii) $h(x) \to 0$ as $||x|| \to \infty$, and (iii) $||x||h(x) \to +\infty$ as $||x|| \to \infty$, e.g., take $\ell(x,a) \equiv h(x) = (1+||x||)^{-\beta}$ for any $\beta \in (0,1)$, so that (1.4) is violated. (One can also allow cases where h takes the value zero at some points in $\mathcal{T}$.) This defines an exit time problem with the control set $\partial B_1(0) \subseteq \mathbb{R}^N$ and the final cost $g \equiv 0$, and one can check that the structural assumptions of Theorem 3.1 apply to this problem. Using the fact that $t_p(\alpha) \geq ||p||/2$ for $||p||$ large and any input α that drives p to $\mathcal{T}$, the calculation

$$\int_0^{t_p(\alpha)} \tilde{h}(||y_p(s,\alpha)||)ds \;\geq\; \int_0^{||p||/2} \tilde{h}(3||p||/2)ds \;=\; \frac{1}{3}\left[\frac{3||p||}{2}\tilde{h}(3||p||/2)\right] \;\to\; +\infty$$

shows that the value functions for the exit time problems defined in this way are proper, and they are also continuous. Therefore, we can take the sets Ω_j in the SSTC($\mathbb{R}^N, \mathcal{T}$) definition to be the sublevel sets $\{x : v(x) < j\}$ of such value functions. This gives a characterization of the value function of the exit time problem, i.e., the travel time function, as the unique proper viscosity solution of (4.6) on $\mathbb{R}^N \setminus \mathcal{T}$ that vanishes on $\mathcal{T}$. In this case, ℓ is bounded below by positive constants on compact sets, so the result follows from the uniqueness results in [16], applied on the increasing sequence of bounded open sublevel sets of any continuous proper function. Theorem 3.1 also gives a uniqueness characterization for (4.6) on sublevel sets of the corresponding value function v which contain $\mathcal{T}$.

Remark 4.3. Since the set of relaxed trajectories for $f(x,a) = a \in \partial B_1(0) \subseteq \mathbb{R}^N$ includes all constant trajectories, we cannot allow eikonal equations in which h vanishes at points outside the target, although we can still allow degenerate cases in which h has a very general zero set in $\mathcal{T}$. Indeed, if h vanished at a point $\bar{p} \notin \mathcal{T}$, then the constant relaxed trajectory at $\bar{p}$ would generate zero running costs, which would violate the requirement $\mathbb{R}^N \setminus \mathcal{T} \subseteq P$. In fact, as shown in [17], the eikonal equation

$$||Du(x)||^2 \;=\; x^2(1-x^2)^2 \tag{4.7}$$

has *two* proper viscosity solutions on $\mathbb{R} \setminus \{0\}$ that vanish at 0. This shows that if we set

$$\tilde{P} := \left\{ x \in \mathbb{R}^N : \int_0^t \ell(y_x(s,\alpha),\alpha(s))\,ds > 0 \quad \forall t \in (0,\infty),\ \alpha \in \mathcal{A} \right\},$$

then the statement of Theorem 3.1 becomes false if $\Omega \setminus \mathcal{T} \subseteq P$ is replaced with $\Omega \setminus \mathcal{T} \subseteq \tilde{P}$. However, our results show that there is only one $(\mathbb{R}\setminus\{-1,0,+1\}, \{-1,0,+1\}, +\infty)$-compatible viscosity solution $w \in C(\mathbb{R})$ of (4.7) on $\mathbb{R} \setminus \{-1,0,+1\}$ which is null on $\{-1,0,1\}$. This extends a result from section 2 of [17].

4.3. Shape-from-shading equations. Our results also apply to equations of the form

$$I(x)\,\Psi(Du(x)) \;-\; b(x)\cdot Du(x) \;-\; h^2(x) \;=\; 0$$

for I nonnegative and Ψ a convex function with $\Psi(0) = 0$. This equation is studied in [17]. Taking the Legendre transform $\Psi^\star$ of Ψ, which is nonnegative, we can rewrite

this equation as

$$\max_{a \in \text{domain}(\Psi^\star)} \left\{ -\left(b(x) - I(x)a\right) \cdot Du(x) \; - \; \left[h^2(x) \; + \; I(x)\Psi^\star(a)\right] \right\} \; = \; 0.$$

A particular case of this equation is

$$I(x) \left[1 + ||Du(x)||^2\right]^{1/2} \; - \; 1 \; = \; 0, \quad x \in \Omega \subseteq \mathbb{R}^2, \tag{4.8}$$

for open sets Ω, which can in fact (cf. [17]) be written as

$$\max_{||a|| \leq 1} \left\{ I(x)\, a \cdot Du(x) \; - \; \left[1 \; - \; I(x) \left(1 - ||a||^2\right)^{1/2}\right] \right\} \; = \; 0. \tag{4.9}$$

Equation (4.9) arises in shape-from-shading models in image processing, where $I(x) \in [0, 1)$ is the intensity of light reflected by an object (cf. [11]). The objective in image processing is to reconstruct the unknown function u, representing the height of the surface over some subset Ω of the plane, from the brightness of a single two-dimensional image of the surface. As shown in [11], for the case of a Lambertian surface which is not self-shadowing and which is illuminated by a single distant vertical light source, the height u satisfies (4.9). Now take any positive Lipschitz continuous intensity function $I(x)$ for which $\lim_{||x|| \to \infty} I(x) = 1$, e.g.,

$$I(x) \; = \; 1 \; - \; \alpha\, e^{-||x||^2}, \qquad 0 < \alpha < 1. \tag{4.10}$$

Just as before, (4.9) is a Bellman equation for an exit time problem with the dynamics $f(x, a) = -I(x)a$ for any compact target $\mathcal{T}$ and the Lagrangian $\ell(x, a) = 1 - I(x)(1 - ||a||^2)^{1/2}$, which violates (1.4). However, with the choice of (4.10), we can use our theorem to get a global uniqueness characterization for viscosity solutions of (4.9) on $\mathbb{R}^N \setminus \mathcal{T}$ if the height u is $(\mathbb{R}^2, \mathcal{T}, \omega_o)$-compatible for some $\omega_o \in \mathbb{R} \cup \{+\infty\}$, continuous, and null on $\mathcal{T}$. This compatibility will hold if, for example, there is a positive constant ω_o so that (i) $u(x) < \omega_o$ for all x and (ii) $u(x) \to \omega_o$ as $||x|| \to \infty$. More generally, the compatibility requirement can be relaxed to the requirement (REG) in section 6.2 with $\Omega = \mathbb{R}^2$, $w = u$, and $P = \mathbb{R}^N$. These results also do not follow from the earlier uniqueness characterizations from [17], where the uniqueness results for (4.9) require $I(x) \leq C < 1$ for C constant. We also get local uniqueness characterizations on sublevel sets of the height function (cf. Theorem 3.1).

Remark 4.4. As discussed above, uniqueness characterizations for HJBEs form the basis for numerical schemes for approximating viscosity solutions. For example, in [11], uniqueness characterizations for (4.8) are used to give stable, robust numerical algorithms for finding the surface. Also, there is a large literature on approximating the minimal time function for exit problems with $\ell \equiv 1$ (cf. [2], [4], and the many references therein). One question which should be considered but which will not be discussed here is how the new uniqueness results for HJBEs can be used to study the convergence of schemes for approximating value functions for cases such as the FP, where $\ell(x, a)$ vanishes for certain $x \notin \mathcal{T}$ and $a \in A$. This type of result could have physical applications. This question will be addressed by the author in a separate paper.

5. Main lemmas and proof of main results. In this section, we recall well-known results from dynamic programming and relaxed controls. These are used to prove Theorem 3.1 and Corollary 3.4.

5.1. Main lemmas. Under (A_0)–(A_3) and $STC\mathcal{T}$, one easily proves that the value function v is a viscosity solution of the HJBE (1.3) on $\mathcal{R}\setminus\mathcal{T}$. The proof follows, since v satisfies the dynamic programming equality

$$v(x) = \inf_{\alpha\in\mathcal{A}}\left\{\int_0^t \ell(y_x(s,\alpha),\alpha(s))\,ds \;+\; v(y_x(t,\alpha))\right\} \quad \forall x\in\mathcal{R},\ t\in[0,\inf_\alpha t_x(\alpha)). \tag{5.1}$$

Our main results are based on the following representation lemmas, which say that viscosity solutions of (1.3) on $\mathcal{R}\setminus\mathcal{T}$ satisfy analogues of (5.1). The proofs of these lemmas are based on uniqueness characterizations for finite horizon problems (cf. [2, Chapter 3] or [10], and for generalizations, see [17] and [18]).

LEMMA 5.1. *Assume that conditions* (A_0)–(A_3) *are satisfied and that* $u\in C(\bar E)$ *is a viscosity subsolution of* $H(x,Du(x))=0$ *on* E, *where* $E\subset\mathbb{R}^N$ *is bounded and open. If we set*

$$\tau_q(\beta) \;=\; \inf\{\,t\ge 0:\ y_q(t,\beta)\ \in\ \partial E\}$$

for each $\beta\in\mathcal{A}$ *and* $q\in E$, *then, for all* $\beta\in\mathcal{A}$ *and* $q\in E$, *we have*

$$u(q) \;\le\; \int_0^r \ell(y_q(s,\beta),\beta(s))\,ds \;+\; u(y_q(r,\beta)) \tag{5.2}$$

for $0\le r<\tau_q(\beta)$.

LEMMA 5.2. *Assume* (A_0)–(A_3) *hold and assume* $w\in C(\bar B)$ *is a viscosity supersolution of* (1.3) *on* B, *where* B *is open and bounded. Set*

$$T_\delta(p) \;:=\; \inf\{\,t:\operatorname{dist}(y_p(t,\alpha),\ \partial B)\le\delta,\ \alpha\in\mathcal{A}\}$$

for each $p\in B$ *and* $\delta>0$. *Then for any* $p\in B$ *and any* $\delta\in(0,\operatorname{dist}(p,\partial B)/2]$, *we have*

$$w(p) \;\ge\; \inf_{\alpha\in\mathcal{A}}\left\{\int_0^t \ell(y_p(s,\alpha),\alpha(s))\,ds \;+\; w(y_p(t,\alpha))\right\} \tag{5.3}$$

for all $t\in(0,T_\delta(p))$.

Notice that we can put $r=\tau_q(\beta)$ in (5.2) when $\tau_q(\beta)<\infty$. We also need the following elementary consequence of the Gronwall inequality and the sequential compactness of $\mathcal{A}^r$ (cf. [24]).

LEMMA 5.3. *Let* A *be a compact metric space, let* $\{\alpha_n\}$ *be a sequence in* $\mathcal{A}^r$, *and let* $c>0$. *Assume that* $f:\mathbb{R}^N\times A\to\mathbb{R}^N$ *satisfies the condition* (A_1). *Then there exists a subsequence of* $\{\alpha_n\}$ *(which we do not relabel) and an* $\alpha\in\mathcal{A}^r$ *such that the following conditions hold:*

1. $\alpha_n\to\alpha$ *weak-*$\star$ *on* $[0,c]$.
2. *If* $x_n\to x$ *in* $\mathbb{R}^N$, *then* $y^r_{x_n}(\cdot,\alpha_n)\to y^r_x(\cdot,\alpha)$ *uniformly on* $[0,c]$.

We sometimes apply Lemma 5.3 to sequences in $\mathcal{A}$ identified with Dirac measure valued relaxed controls.

5.2. Proof of main results. Corollary 3.6 follows from Theorem 3.1 because v is a viscosity solution of the associated HJBE, which is shown by first establishing that v satisfies the dynamic programming equality (5.1) (cf. [2]). We leave the details to the reader and prove only Theorem 3.1 and Corollary 3.4.

5.2.1. Proof of Theorem 3.1. The inequality "$w \leq v$" follows from the following more general prolongation result.

PROPOSITION 5.4. *Assume that* (1.2) *satisfies* (A_0)–(A_3)*, that* $\Omega \subseteq \mathbb{R}^N$ *is an open set containing* $\mathcal{T}$ *which satisfies* $\Omega \setminus \mathcal{T} \subseteq P$*, and that* $w \in QG(\Omega)$ *is a viscosity subsolution of* (3.1)*. Then* $w \leq v$ *on* Ω*.*

Proof. Let $x \in \Omega \setminus \mathcal{T}$ be given, and let B be any bounded open subset of Ω that contains x and is such that $\bar{B} \subseteq \Omega \setminus \mathcal{T}$. Then w is also a viscosity subsolution of the HJBE (1.3) on B. Since $w \in C(\overline{B})$, it follows from Lemma 5.1 that

$$w(x) \;\leq\; \int_0^t \ell(y_x(s,\alpha),\alpha(s))\,ds \;+\; w(y_x(t,\alpha)) \tag{5.4}$$

for all $\alpha \in \mathcal{A}$ and $t \in [0, \tau_x(\alpha))$, where the $\tau_x(\alpha) := \inf\{t \geq 0 : y_x(t,\alpha) \notin B\} \in [0, +\infty]$. Moreover, we can put $t = \tau_x(\alpha)$ in (5.4) when $\tau_x(\alpha) < \infty$.

Suppose that $w(x) > v(x)$. By the definition of the infimum, there is an $\tilde{\alpha} \in \mathcal{A}^f(x)$ such that

$$\int_0^{t_x(\tilde{\alpha})} \ell(y_x(s,\tilde{\alpha}),\tilde{\alpha}(s))\,ds \;+\; g(y_x(t_x(\tilde{\alpha}),\tilde{\alpha})) < w(x). \tag{5.5}$$

If $y_x(s,\tilde{\alpha}) \in \Omega$ for all $s \in [0, t_x(\tilde{\alpha})]$, then $t_x(\tilde{\alpha})$ is a limit of exit times from sets $B = B_k$, as above, as $k \to \infty$. For example, take B_k to be an open tube around the restriction of the trace of $y_x(\cdot,\tilde{\alpha})$ to $[0,\, t_x(\tilde{\alpha}) - 1/k]$ for k large enough. In that case we arrive at a contradiction once we put $\alpha = \tilde{\alpha}$ and $t = t_x^k(\tilde{\alpha})$ in (5.4), where $t_x^k(\alpha) := \inf\{t \geq 0 : y_x(t,\alpha) \notin B_k\}$, and pass to the limit as $k \to \infty$. Otherwise, let $\hat{\tau}$ be the last time in $(0, t_x(\tilde{\alpha}))$ that $y_x(\cdot,\tilde{\alpha})$ is in $\partial\Omega$, and apply Lemma 5.1 with the choices

$$q = z_n := y_x(\hat{\tau} + 1/n, \tilde{\alpha}), \quad \beta(\cdot) = \alpha_n(\cdot) := \tilde{\alpha}(\cdot + \hat{\tau} + 1/n), \quad r = t_{z_n}(\tilde{\alpha}(\cdot + \hat{\tau} + 1/n)),$$

and with E chosen to be a tube in Ω which contains a portion of the trajectory $y_x(\cdot,\tilde{\alpha})$ that runs from z_n to $\mathcal{T}$. We then get $w(z_n) < w(x)$ for all n. Indeed, (5.5) would give

$$\begin{aligned} w(x) &> \int_0^{\hat{\tau}+1/n} \ell(y_x(s,\tilde{\alpha}),\tilde{\alpha}(s))\,ds \;+\; \int_0^{t_x(\tilde{\alpha})-\hat{\tau}-1/n} \ell(y_{z_n}(s,\alpha_n),\alpha_n(s))\,ds \\ &\quad + w(y_x(t_x(\tilde{\alpha}),\tilde{\alpha})) \\ &\geq \int_0^{\hat{\tau}+1/n} \ell(y_x(s,\tilde{\alpha}),\tilde{\alpha}(s))\,ds + w(z_n). \end{aligned} \tag{5.6}$$

Since $x \in \Omega \setminus \mathcal{T} \subseteq P$, the last integral is nonnegative for all n, so we get $w(x) > w(z_n)$ for large n. But $\Omega \ni z_n \to y_x(\hat{\tau},\tilde{\alpha}) \in \partial\Omega$, and this contradicts the assumption $w \in QG(\Omega)$. Since $w \equiv v$ on $\mathcal{T}$, the result follows. □

We turn next to the proof of the inequality "$w \geq v$." Fix $x \in \Omega \setminus \mathcal{T}$. Since $\Omega \setminus \mathcal{T} \subseteq P$, we know that $x \in \Omega_j \setminus \mathcal{T}$ for large enough j, and for such a j we set $\mathcal{S} = \Omega_j$. We later put some restrictions on the value of j. We now use (5.3) of Lemma 5.2 (with $B = \mathcal{S} \setminus \mathcal{T}$) to prove the existence of a trajectory which starts at x and reaches $\mathcal{T}$ in finite time. The lemma applies since $\overline{\mathcal{S} \setminus \mathcal{T}} \subseteq \mathcal{S} \setminus \mathcal{T} \cup \partial\mathcal{S} \cup \partial\mathcal{T} \subseteq \Omega$ (by the compatibility assumption) and w is continuous on Ω.

We will first assume that $\omega_0 < \infty$. The proof is similar in spirit to the proof of Theorem IV.3.15 in [2] and arguments in [8], but we use a weak-$\star$ convergence

and strong controllability argument to replace the assumptions that w is bounded below and ℓ is bounded away from zero. We will apply inequality (5.3) of Lemma 5.2 repeatedly to a sequence of points $p = x_1, x_2, \dots$ in $\mathcal{S} \setminus \mathcal{T}$. We will assume that all of the corresponding δ's in the lemma can be chosen to be 1. The general case then follows by replacing the corresponding $1/k$'s with δ_k/k's for a suitable sequence $\delta_k \searrow 0$ in the argument below. In what follows, we set

$$I(x,t,\alpha) \;=\; \int_0^t \ell(y_x(s,\alpha),\alpha(s))ds \;+\; w(y_x(t,\alpha)).$$

Given $\varepsilon > 0$, let us begin by constructing an $\hat{\alpha} \in \mathcal{A}$ such that $\tau_x(\hat{\alpha}) < +\infty$ and such that

$$w(x) \;\geq\; \int_0^{\tau_x(\hat{\alpha})} \ell(y_x(s,\hat{\alpha}),\hat{\alpha}(s))\,ds \;+\; \lambda_x(\hat{\alpha}) \;-\; \varepsilon, \tag{5.7}$$

where

$$\tau_x(\alpha) \;:=\; \inf\{t \geq 0 : y_x(t,\alpha) \notin \mathcal{S} \setminus \mathcal{T}\}$$

and

$$\lambda_x(\alpha) := \begin{cases} \dfrac{w(x)+\omega_0}{2}, & t_x(\alpha) \neq \tau_x(\alpha), \\ w\left(y_x(\tau_x(\alpha),\alpha)\right), & t_x(\alpha) = \tau_x(\alpha). \end{cases}$$

Setting

$$T_\delta(p) \;=\; \inf\left\{t \geq 0 : \operatorname{dist}\left(y_p(t,\alpha), \partial(\mathcal{S} \setminus \mathcal{T})\right) < \delta, \alpha \in \mathcal{A}\right\} \quad \forall\; p \in \mathbb{R}^N,\; \delta > 0,$$

we define $x_1 := x$, $\tau_1 := T_1(x_1)$ when $T_1(x_1) < +\infty$, and $\tau_1{:=}10$ when $T_1(x_1) = +\infty$. Since $w \in C(\Omega)$ and $\overline{\mathcal{S} \setminus \mathcal{T}} \subseteq \Omega$, we can use (5.3) of Lemma 5.2 to get an α_1 such that $w(x_1) \geq I(x_1,\tau_1,\alpha_1) - \varepsilon/4$. Note that $y_{x_1}(\tau_1,\alpha_1) \in \mathcal{S} \setminus \mathcal{T}$. By induction, we define

$$x_k \;:=\; y_{x_{k-1}}(\tau_{k-1},\alpha_{k-1}) \;\in\; \mathcal{S} \setminus \mathcal{T} \quad \text{for} \quad k = 2,3,\dots, \tag{5.8}$$

where

$$\tau_k \;:=\; \begin{cases} T_{1/k}(x_k) & \text{if } T_{1/k}(x_k) < +\infty, \\ 10^k, & \text{otherwise.} \end{cases}$$

Since $x_k \in \mathcal{S} \setminus \mathcal{T}$, we can use (5.3) to get an $\alpha_k \in \mathcal{A}$ such that

$$w(x_k) \;\geq\; I(x_k,\tau_k,\alpha_k) \;-\; 2^{-(k+1)}\varepsilon \;\;\forall\;\; k. \tag{5.9}$$

We also set $\sigma_k := \tau_1 + \cdots + \tau_k$, $\bar{\sigma} = \limsup_k \sigma_k$, and, for an arbitrary $\bar{a} \in A$,

$$\bar{\alpha}(s) := \begin{cases} \alpha_1(s) & \text{if } 0 \leq s < \sigma_1, \\ \alpha_2(s-\sigma_1) & \text{if } \sigma_1 \leq s < \sigma_2, \\ \vdots & \\ \alpha_k(s-\sigma_{k-1}) & \text{if } \sigma_{k-1} \leq s < \sigma_k, \\ \vdots & \\ \bar{a} & \text{if } \bar{\sigma} \leq s, \end{cases}$$

with the last line used if $\bar{\sigma} < +\infty$. From the definitions of x_k, P, and $\bar{\alpha}$, we know that

$$y_x(s,\bar{\alpha}) \;=\; y_{x_k}(s-\sigma_{k-1},\alpha_k) \;\in\; \mathcal{S}\setminus\mathcal{T}\subseteq P \quad \text{when} \quad s<\bar{\sigma} \tag{5.10}$$

and

$$\int_0^{\tau_k}\ell(y_{x_k}(s,\alpha_k),\alpha_k(s))\,ds \;=\; \int_{\sigma_{k-1}}^{\sigma_k}\ell(y_x(s,\bar{\alpha}),\bar{\alpha}(s))ds \;\geq\; 0 \;\forall\; k. \tag{5.11}$$

Applying (5.9) repeatedly, we therefore get

$$\begin{aligned} w(x) &\geq \int_0^{\tau_1}\ell(y_x(s,\bar{\alpha}),\bar{\alpha}(s))\,ds + w(x_2) - \frac{\varepsilon}{4} \\ &\geq \int_0^{\sigma_2}\ell(y_x(s,\bar{\alpha}),\bar{\alpha}(s))\,ds + w(x_3) - \varepsilon\left(\frac{1}{4}+\frac{1}{8}\right) \\ &\geq \cdots \\ &\geq I(x,\sigma_k,\bar{\alpha}) - \frac{\varepsilon}{2}\left(1-\frac{1}{2^k}\right) \qquad \forall k. \end{aligned} \tag{5.12}$$

By (5.8) and the boundedness of $\mathcal{S}\in\{\Omega_j\}$, we know that $\{x_k\}$ is bounded and therefore clusters. Let $\bar{x}$ be a cluster point of the x_k's, and assume without loss of generality (w.l.o.g.) that $x_k\to\bar{x}$ (by passing to a subsequence without relabeling). Then $\bar{x}\in\overline{\mathcal{S}\setminus\mathcal{T}}$. We will need the following minimality property of $\bar{x}$.

PROPOSITION 5.5. *In the above notation,* $\bar{\tau} := \inf\{\tau_{\bar{x}}(\alpha) : \alpha\in\mathcal{A}\} \;\leq\; \limsup_k \tau_k$.

Proof. First assume $\bar{\tau}<\infty$. Let $\delta>0$ be given, and suppose that, for k as large as desired, we have $\tau_k < \bar{\tau}-\delta$. Passing to a subsequence, we assume that $\tau_k\to z\in[0,\bar{\tau}-\delta]$. There would then exist a sequence $\tilde{\tau}_k\to z$ and a control $u\in\mathcal{A}^r$ such that

$$\operatorname{dist}(y^r_{\bar{x}}(z,u),\partial(\mathcal{S}\setminus\mathcal{T})) \;\leftarrow\; \operatorname{dist}(y_{x_k}(\tilde{\tau}_k,u_k),\partial(\mathcal{S}\setminus\mathcal{T})) \;\leq\; \frac{1}{k}\to 0 \;\text{ as }\; k\to+\infty. \tag{5.13}$$

The u_k's and the $\tilde{\tau}_k$'s are chosen using the definition of the infima τ_k, and u is a weak-$\star$ limit of some subsequence of the u_j's. The existence of such a control u follows from Lemma 5.3 with c chosen to be some upper bound of the $\tilde{\tau}_k$'s. To check (5.13), note that

$$||y^r_{\bar{x}}(z,u)-y_{x_k}(\tilde{\tau}_k,u_k)|| \;\leq\; ||y^r_{\bar{x}}(z,u)-y^r_{\bar{x}}(\tilde{\tau}_k,u)|| \;+\; ||y^r_{\bar{x}}(\tilde{\tau}_k,u)-y_{x_k}(\tilde{\tau}_k,u_k)|| \;\to\; 0$$

and that $\operatorname{dist}(\cdot,\partial(\mathcal{S}\setminus\mathcal{T}))$ is continuous. The standard trajectories $y_{\bar{x}}(\cdot,u_k)$ converge uniformly to $y^r_{\bar{x}}(\cdot,u)$ on $[0,z+1]$ (by Lemma 5.3), and (5.13) gives

$$y^r_{\bar{x}}(\tilde{\tau}_k,u) \quad\to\quad y^r_{\bar{x}}(z,u) \quad\in\quad \partial(\mathcal{S}\setminus\mathcal{T}) \quad \text{as} \quad k\;\to\;\infty. \tag{5.14}$$

Therefore, for large k, we know that $y_{\bar{x}}(\tilde{\tau}_k,u_k)$ lies in $\mathcal{S}\setminus\mathcal{T}$ (since we are supposing that $\tilde{\tau}_k<\bar{\tau}$ for all k) and can be brought to $\partial(\mathcal{S}\setminus\mathcal{T})$ by some standard control $\tilde{u}$ in time less than $\delta/2$ (by the SSTC condition and (5.14)). If we concatenate a control u_k for such a k and a corresponding control $\tilde{u}$, we get a (standard) trajectory which brings $\bar{x}$ to $\partial(\mathcal{S}\setminus\mathcal{T})$ in time $\leq\bar{\tau}-\delta/4$, which stands in contradiction to the definition of $\bar{\tau}$. We conclude that $\tau_k\geq\bar{\tau}-\delta$ for k large enough, which establishes the result

if $\bar\tau < \infty$. If $\bar\tau = +\infty$, then replace $\bar\tau - \delta$ in the argument above with any positive number to get the result. ∎

Passing to a further subsequence without relabeling, fix $l \geq \bar\tau$ so that $\tau_k \uparrow l \in [0, +\infty]$. By the estimate $(\mathcal{E}_1)$, we know that $\bar\tau = 0$ iff $\bar x \in \partial(\mathcal{S} \setminus \mathcal{T})$. (Indeed, if $\bar x \notin \partial(\mathcal{S} \setminus \mathcal{T})$, then $\bar x \in \mathcal{S} \setminus \mathcal{T}$, so $B(\{\bar x\}, \mu) \subseteq \mathcal{S} \setminus \mathcal{T}$ for some $\mu > 0$. By $(\mathcal{E}_1)$, there is a $\gamma > 0$ so that all trajectories which start at $\bar x$ and run for time $\leq \gamma$ stay in $B(\{\bar x\}, \mu)$, and thus $\bar\tau \geq \gamma > 0$. The converse is trivial.) We use the following corollary, which is a consequence of Proposition 5.5.

COROLLARY 5.6. *With the above notation, we have* $\bar x \in \partial(\mathcal{S} \setminus \mathcal{T})$.

Proof. Let $M \in (0, l)$, and let $\tilde\alpha \in \mathcal{A}^r$ be a weak-$\star$ limit of a subsequence of the α_k's in $\mathcal{A}^r$ on $[0, M]$, which we assume to be the sequence itself for brevity (cf. Lemma 5.3). We conclude from (5.11) that

$$0 \leftarrow \int_{\sigma_{k-1}}^{\sigma_k \wedge \{\sigma_{k-1}+M\}} \ell(y_x(s, \bar\alpha), \bar\alpha(s))\, ds = \int_0^{\tau_k \wedge M} \ell(y_{x_k}(s, \alpha_k), \alpha_k(s))\, ds \;\to\; \int_0^{l \wedge M} \ell^r(y^r_{\bar x}(s, \tilde\alpha), \tilde\alpha(s))\, ds. \tag{5.15}$$

The left arrow is by the divergence test applied to the last integral in (5.12), since w is bounded below on $\bar{\mathcal{S}}$ and $x_k \in \mathcal{S} \setminus \mathcal{T} \subseteq P$ for all k. To justify the right arrow, apply Lemma 5.3 to get an $\tilde\alpha \in \mathcal{A}^r$ such that

$$y^r_{x_k}(s, \alpha_k) \;\to\; y^r_{\bar x}(s, \tilde\alpha) \quad \text{uniformly on} \quad [0, M]. \tag{5.16}$$

Let $\alpha_{k,s}(\cdot)$ (resp., $\tilde\alpha_s(\cdot)$) denote the Radon measures $\alpha_k(s)$ (resp., $\tilde\alpha(s)$) for each k. Then,

$$\begin{aligned} &\left| \int_0^M \left[\ell^r(y^r_{\bar x}(s, \tilde\alpha), \tilde\alpha(s)) - \ell^r(y^r_{x_k}(s, \alpha_k), \alpha_k(s)) \right] ds \right| \\ &\leq \left| \int_0^M \int_A \ell(y^r_{\bar x}(s, \tilde\alpha), a)\, d\tilde\alpha_s(a)\, ds - \int_0^M \int_A \ell(y^r_{\bar x}(s, \tilde\alpha), a)\, d\alpha_{k,s}(a)\, ds \right| \\ &\quad + \left| \int_0^M \int_A \left[\ell(y^r_{\bar x}(s, \tilde\alpha), a) - \ell(y^r_{x_k}(s, \alpha_k), a) \right] d\alpha_{k,s}(a)\, ds \right|. \end{aligned}$$

The first term on the right-hand side (RHS) $\to 0$ because $\alpha_k \to \tilde\alpha$ weak-$\star$ on the interval $[0, M]$ and because we can set $(h(s))(a) := \ell(y^r_{\bar x}(s, \tilde\alpha), a)$ in (2.1).[4] The second RHS term $\to 0$ by (5.16) and the assumption (A_3). This justifies the right arrow in (5.15), because $\tau_k \wedge M = M$ for large k, since $M < l$ and $\tau_k \to l$.

If we had $\int_0^{\bar\tau} \ell^r(y^r_{\bar x}(s, \tilde\alpha), \tilde\alpha(s))\, ds > 0$, then we have $\int_0^G \ell^r(y^r_{\bar x}(s, \tilde\alpha), \tilde\alpha(s))\, ds > 0$ for some $G \in (0, \bar\tau)$. Since $l \geq \bar\tau$ (cf. Proposition 5.5), we would reach a contradiction by putting $M = G$ in (5.15). Therefore, $\bar\tau = 0$, or $\bar x$ is not in P. If $\bar\tau = 0$, then $\bar x \in \partial(\mathcal{S} \setminus \mathcal{T})$, as explained above. Assume $\bar x \notin P$. Since $\bar x \in \overline{\mathcal{S} \setminus \mathcal{T}}$ by construction, and $\mathcal{S} \setminus \mathcal{T} \subset P$ by assumption, we again conclude that $\bar x \in \partial(\mathcal{S} \setminus \mathcal{T})$, as needed. ∎

We can therefore find $\bar n$ such that for each $k > \bar n$ there exists a $\beta_k \in \mathcal{A}$ which drives x_k to a point $\tilde x \in \partial(\mathcal{S} \setminus \mathcal{T})$ (depending on k) and which is such that

$$\int_0^{\tau_{x_k}(\beta_k)} \ell(y_{x_k}(s, \beta_k), \beta_k(s))\, ds \;<\; \varepsilon/4. \tag{5.17}$$

[4] We are using the fact that ℓ is continuous in the control set value.

To see why, let F be a bounded set containing $\{y_{x_k}(s,\beta) : \beta \in \mathcal{A}, 0 \leq s \leq 1, k \in \mathbb{N}\}$. Such a set exists by the estimate $(\mathcal{E}_2)$ and the boundedness of the x_k's. Set $\bar{\kappa} = 1 + \sup \ell\lceil [F \times A]$. This is finite since A is compact and ℓ is continuous. By the SSTC assumption and the fact that $\{x_k\}$ converges to a point in $\partial(\mathcal{S} \setminus \mathcal{T})$, we can find $\bar{n} \in \mathbb{N}$ so that if $k > \bar{n}$, then x_k can be brought to $\partial(\mathcal{S} \setminus \mathcal{T})$ in time $< \varepsilon/(4\bar{\kappa}) \wedge 1$. This establishes (5.17) for $k \geq \bar{n}$.

For $k > \bar{n}$, the construction (5.12) therefore gives

$$w(x) \ \geq \ I(x, \sigma_{k-1}, \bar{\alpha}) \ + \ \int_0^{\tau_{x_k}(\beta_k)} \ell(y_{x_k}(s,\beta_k), \beta_k(s))\, ds \ - \ \varepsilon/2\,(1 - 2^{-(k-1)} + 1/2). \tag{5.18}$$

Since $\partial(\mathcal{S}\setminus\mathcal{T}) \subseteq \mathcal{T} \cup (\partial\mathcal{S}\setminus\mathcal{T})$, we now separately consider the case in which $\tilde{x} \in \mathcal{T}$ and the case in which $\tilde{x} \in (\partial\mathcal{S}\setminus\mathcal{T})$. For large k, it follows that if $\tilde{x} \in \mathcal{T}$, then

$$w(x_k) \ + \ \varepsilon/4 \ \ > \ \ w(\tilde{x}). \tag{5.19}$$

To see why, first recall that $w \in C(\Omega)$ and $\mathcal{T} \subseteq \Omega$. By choosing small enough running times for the paths from x_k to $\tilde{x} \in \partial(\mathcal{S} \setminus \mathcal{T})$, we can use the estimate $(\mathcal{E}_1)$ from section 2 to ensure that $||x_k - \tilde{x}|| < \delta$, where $\delta > 0$ is chosen so that $|w(p) - w(a)| < \varepsilon/4$ for all $a \in \mathcal{T} \cap \bar{\mathcal{S}}$ and $p \in B(\mathcal{T} \cap \bar{\mathcal{S}}, \delta)$. By the SSTC assumption, these running times can be made as small as desired by taking k to be large. Such a δ exists since w is continuous on the compact set $\bar{\mathcal{S}} \cap \mathcal{T}$. The estimate (5.19) now follows by choosing $a = \tilde{x}$ and $p = x_k$.

If, on the other hand, we have $\tilde{x} \in \partial\mathcal{S} \setminus \mathcal{T}$, then we get $w(x_k) \geq 1/2\,(w(x) + \omega_0)$ for large k, since $w(x) < \omega_o$ and, by the compatibility condition, we can assume that the j we choose is so large that $w(p) > \frac{1}{4}w(x) + \frac{3}{4}w_0$ for $p \in \mathcal{S} \setminus \mathcal{T}$ sufficiently close to $\partial\mathcal{S} \setminus \mathcal{T}$. (The closeness of x_k to $\tilde{x}$ is achieved by choosing k large, as explained in the previous paragraph.) In the former case, we add and subtract $\varepsilon/4$ in (5.18). We can therefore take $\hat{\alpha}$ to be the concatenation

$$\hat{\alpha}(s) \ := \ \begin{cases} \bar{\alpha}(s) & \text{if } \ 0 \leq s \leq \sigma_{k-1}, \\ \\ \beta_k(s - \sigma_{k-1}) & \text{if } \ \sigma_{k-1} \leq s < \infty \end{cases}$$

for k large enough.

Thus, when $\omega_0 \in \mathbb{R}$, our construction always gives us a control $\hat{\alpha}$ which satisfies the condition $\tau_x(\hat{\alpha}) < +\infty$ and which satisfies the requirement (5.7). Since $\varepsilon > 0$ was arbitrary, we conclude that

$$w(x) \geq \inf_{\alpha} \left\{ \int_0^{\tau_x(\alpha)} \ell(y_x(s,\alpha), \alpha(s)) ds \ + \ \lambda_x(\alpha) : \ \ \tau_x(\alpha) \ < \ +\infty \right\}. \tag{5.20}$$

If $\alpha \in \mathcal{A}$ is such that $\tau_x(\alpha) \neq t_x(\alpha)$, then the corresponding infimand in (5.20) is

$$\int_0^{\tau_x(\alpha)} \ell(y_x(s,\alpha), \alpha(s))\, ds \ + \ \frac{1}{2}\,(\omega_0 + w(x)) \ \ > \ \ w(x),$$

since $w(x) < w_0$ for all $x \in \Omega$ and $x \in \mathcal{S} \setminus \mathcal{T} \subseteq P$, and thus such a control is irrelevant for the infimum. Since $g = w = v$ on $\mathcal{T}$, this establishes that $w \geq v$ on Ω for the case in which $\omega_o < \infty$.

The proof for the $w_0 = +\infty$ case is similar to this. We view a fixed $x \in \Omega \setminus \mathcal{T}$ as a member of $\Omega_j \setminus \mathcal{T}$, where j is large enough so that values of w on $\partial(\Omega_j) \setminus \mathcal{T}$ majorize $w(x)$, and then we construct a trajectory that reaches $\partial(\Omega_j)$ or $\mathcal{T}$ in finite time. By the above, the controls α for which $\tau_x(\alpha) \neq t_x(\alpha)$ are irrelevant for the calculation of the infimum in (5.20), and thus $w \geq v$ as before. This gives $w = v$ on Ω.

5.2.2. Proof of Corollary 3.4. It remains to prove the sublevel set characterization $\Omega = \{x \in \mathbb{R}^N : v(x) < \omega_o\}$ under the added assumption (♯). The inclusion "$\subseteq$" holds, since $w \equiv v$ on Ω and $w < \omega_o$ on Ω. Suppose $x \notin \Omega$ and $v(x) < \omega_o$, for the sake of obtaining a contradiction. Using the definition of the infimum, we can then find a $\bar{\beta} \in \mathcal{A}$ and a $K < \omega_o$ so that $t_x(\bar{\beta}) < \infty$ and so that

$$\int_0^{t_x(\bar{\beta})} \ell(y_x(s,\bar{\beta}),\bar{\beta}(s))ds \;+\; g(y_x(t_x(\bar{\beta}),\bar{\beta})) \;\;=\;\; K.$$

Since $\mathcal{T} \subset \Omega$ is closed and Ω is open, we have

$$\bar{t} := \sup\big\{t \in [0, t_x(\bar{\beta})) : y_x(t,\bar{\beta}) \in \partial\Omega\big\} \;\;<\;\; t_x(\bar{\beta}).$$

Set $z_n := y_x(\bar{t} + 1/n, \bar{\beta})$ for $n \in \mathbb{N}$ and let $\{\Omega_j\}$ be the controllability sequence from the compatibility hypothesis. We can assume $\bar{t} + 1/n < t_x(\bar{\beta})$ for all n. Since $z := y_x(\bar{t},\bar{\beta}) \notin \mathcal{T}$ and $\Omega \cap P \supseteq \Omega \setminus \mathcal{T} \ni z_n \to z$, it follows that for each sufficiently large n there is an $m(n) \in \mathbb{N}$ such that $z_n \in \Omega_{m(n)} \setminus \mathcal{T}$. Recall that $\{\Omega_j\}$ is an increasing sequence. If $\{m(n)\}$ can be taken to be bounded, then it follows that $\{z_n\}$ accumulates in some $\overline{\Omega_{\bar{m}}}$, and thus $z \in \overline{\Omega_{\bar{m}}} \setminus \mathcal{T} \subseteq \Omega$. Since $z \in \partial\Omega$, this contradicts the fact that Ω is open.

Therefore, for each n, we can find a $k(n) \in (0, \frac{1}{n})$ so that $z_n' := y_x(\bar{t} + k(n), \bar{\beta}) \in \partial\left(\Omega_{m(n)}\right) \setminus \mathcal{T}$, namely, $k(n) := \inf\{t \in [0, 1/n] : y_x(\bar{t} + t, \bar{\beta}) \in \partial(\Omega_{m(n)})\}$. By the definition of compatibility,

$$\limsup_{n\to\infty} w(z_n') \;\geq\; \omega_o. \tag{5.21}$$

On the other hand, by setting $\bar{\alpha}_n(\cdot) = \bar{\beta}(\cdot + \bar{t} + k(n))$ and applying Lemma 5.1 as before, we get

$$\begin{aligned} w(z_n') &\leq \int_0^{t_{z_n'}(\bar{\alpha}_n)} \ell(y_{z_n'}(s,\bar{\alpha}_n),\bar{\alpha}_n(s))ds + g(y_x(t_x(\bar{\beta}),\bar{\beta})) \\ &\leq \int_0^{t_{z_n'}(\bar{\alpha}_n)} \ell(y_{z_n'}(s,\bar{\alpha}_n),\bar{\alpha}_n(s))ds + g(y_x(t_x(\bar{\beta}),\bar{\beta})) \\ &\quad + \int_0^{\bar{t}} \ell(y_x(s,\bar{\beta}),\bar{\beta}(s))\,ds \leq K \end{aligned}$$

for n sufficiently large. The second inequality follows from (♯), and the last one follows from the fact that $\Omega \setminus \mathcal{T} \subseteq P$ and from the definition of $z \in \mathcal{T}^c$, which guarantees that

$$\int_{\bar{t}}^{\bar{t}+k(n)} \ell(y_x(s,\bar{\beta}),\bar{\beta}(s))\,ds \;\geq\; 0,$$

as a limit of the nonnegative running costs from points $y_x(\bar{t} + 1/m, \bar{\beta}) \in \Omega \setminus \mathcal{T} \subseteq P$ as $m \to \infty$. Therefore,

$$\limsup_{n\to\infty}\, w(z_n') \;\;\leq\;\; K \;\;<\;\; \omega_o,$$

which contradicts (5.21). This proves the sublevel set characterization. Corollary 3.4 now follows.

6. Extensions. We close with remarks on how to extend the results of section 3. We show how to relax the requirements that the control set A be bounded and that $\Omega \setminus \mathcal{T} \subseteq P$. We also show how to improve the theorem to get a uniqueness characterization within a class of functions satisfying more general notions of properness.

6.1. A further property of controllability sequences. For the Fuller example, there is a controllability sequence Ω_j for which $\mathcal{T} \subseteq \partial(\Omega_j)$ for all j. However, Theorem 3.1 allows controllability sequences for which this condition fails. On the other hand, if Ω, w, and so forth are as in the hypotheses of the theorem, and if $x \in \Omega \setminus \mathcal{T}$ and k are chosen so that $x \in \Omega_k$ and so that $\min\{w(p) : p \in \partial(\Omega_k) \setminus \mathcal{T}\} > w(x)$, then the argument of Theorem 3.1 gives a trajectory from x to $\mathcal{T}$ which lies completely in $\bar{\Omega}_k$. Indeed, given a point $x \in \Omega \setminus \mathcal{T}$, the hypotheses of Theorem 3.1 guarantee the existence of a trajectory which remains in $\bar{\Omega}_k$ and which reaches $\partial(\Omega_k) \setminus \mathcal{T}$ or $\mathcal{T}$ in finite time. The proof shows that if there is no trajectory whose exit time from $\Omega_k \setminus \mathcal{T}$ is an exit time from $\mathcal{T}^c$, then all the infimands in (5.20) are $> w(x)$, which is a contradiction. In particular, $\bar{\Omega}_k \cap \mathcal{T} \neq \emptyset$.

6.2. Generalized properness notions. As mentioned in section 2, the uniqueness conclusion "$w = v$" of Theorem 3.1 remains true if the compatibility assumption of the theorem is replaced by the assumption that w satisfies an even more general properness assumption (and all the other hypotheses are kept the same). For example, we can replace the compatibility requirement with the following:

(REG) Condition STC$\mathcal{T}$ holds, and for each $x \in \Omega \cap P$ there is a bounded open set $\Omega_x \subseteq \Omega \setminus \mathcal{T}$ which contains x and a $\omega_{o,x} \in \mathbb{R} \cup \{+\infty\}$ such that
- (REG_1) $w(p) < \omega_{o,x}$ for all $p \in \Omega_x$,
- (REG_2) $w(p) \to \omega_{o,x}$ as $\Omega_x \ni p \to x_o$ for all $x_o \in \partial(\Omega_x)$, and
- (REG_3) STC(Ω_x^c) and $\bar{\Omega}_x \setminus \mathcal{T} \subseteq \Omega \cap P$.

The proof follows from the argument we gave above, once we replace $\mathcal{S} = \Omega_j$ with the set Ω_x, and the argument will show that $\bar{\Omega}_x \cap \mathcal{T} \neq \emptyset$ for all x. The novelty here is that there may be uncountably many Ω_x's and that one does not need subcollections of the Ω_x's to form increasing sets (cf. Definition 2.3).

6.3. Problems with unbounded control sets. One can relax the compactness requirement (A_0) in Theorem 3.1 in various ways. One way to do this is to replace the inputs $t \mapsto \alpha(t)$ with vector field valued inputs $[0, \infty[\ni t \mapsto f(\cdot, \alpha(t)) \times \ell(\cdot, \alpha(t))$ and then add assumptions on f and ℓ which guarantee that the control set for the vector field valued inputs is compact. The details are as follows. We assume that A is a closed subset of a Euclidean space, for simplicity, but the argument also holds for more general control sets. We equip $C(\mathbb{R}^N, \mathbb{R}^{N+1})$ with the *metrizable* topology of compact convergence. We make the following assumptions on f and ℓ:

- (NC_1) $M_\ell := \sup\{\ell(0, a) : a \in A\} < \infty$.
- (NC_2) $\{f(\cdot, a) \times \ell(\cdot, a) : a \in A\}$ is a closed subset of $C(\mathbb{R}^N, \mathbb{R}^{N+1})$.
- (NC_3) $f : \mathbb{R}^N \times A \to \mathbb{R}^N$ satisfies condition (A_1) and is bounded on $B_R(0) \times A$ for all $R > 0$.
- (NC_4) $\ell : \mathbb{R}^N \times A \to \mathbb{R}$ is continuous and bounded below, and there is a modulus ω such that for all $x, y \in \mathbb{R}^N$ and all $a \in A$, $|\ell(y, a) - \ell(x, a)| \leq \omega(|| x - y ||)$.

We will also find it convenient to assume the additional condition:

(NC_5) $f(\cdot,a)\times\ell(\cdot,a)\equiv f(\cdot,b)\times\ell(\cdot,b)\Rightarrow a=b$, i.e., the mapping $a\overset{\mu}{\mapsto} f(\cdot,a)\times\ell(\cdot,a)$ is one-to-one.

However, (NC_5) is not necessary if we restate the condition $\Omega\setminus\mathcal{T}\subseteq P$ in terms of vector field valued inputs (see the discussion below). In (NC_3), a modulus is a continuous nondecreasing function $\omega:\mathbb{R}_+\to\mathbb{R}_+$ which satisfies $\omega(0)=0$. The vector field $\ell(0,\cdot)$ in (NC_1) can be replaced by $\ell(p,\cdot)$ for any $p\in\mathbb{R}^N$ (by assumption (NC_4)). Notice that under assumption (NC_1), the mappings f^r and ℓ^r defined in (2.2) are still well-defined functions. Also, (NC_3) guarantees that $y_x(\cdot,\alpha)$ is defined on $[0,\infty)$ for all $\alpha\in\mathcal{A}$ and $x\in\mathbb{R}^N$ and satisfies ($\mathcal{E}_1$)–($\mathcal{E}_2$) (cf. [2]). All the weak-$\star$ convergence in what follows takes place in $\mathcal{K}^r$ (i.e., the relaxed controls on K), where K is a suitable compact set of vector fields defined below, and we topologize K^r and $\mathcal{K}^r$ as before. In particular, K^r is topologized as a subset of the dual space of $C(K)$, and $\mathcal{K}^r\ni\alpha_n\to\bar\alpha\in\mathcal{K}^r$ means that (2.1) holds, with A replaced by K, for all Lebesgue integrable functions $h:[0,\tau]\to C(K)$ and all $\tau>0$. Condition (NC_5) guarantees that $\mu\lceil\ \overline{B_R(0)}\ $ has a continuous inverse for each $R>0$, where we write $B_R(0)$ to mean $B_R(0)\cap A$, to simplify the notation. (Indeed, assume $a_n,a\in B_R(0)$ for all n and $\mu(a_n)\to\mu(a)$. If $\varepsilon>0$ is such that $||a_{n'}-a||\geq\varepsilon$ along some sequence $a_{n'}$, then we can pass to a further subsequence without relabeling to get $a_{n'}\to\bar a\in\ \overline{B_R(0)}$. Since μ is continuous, we get $\mu(a_{n'})\to\mu(\bar a)$. By ($NC_5$), $\bar a=a$, which is a contradiction.)

In particular, if $S\subseteq A$ is closed, then

$$\begin{aligned}\mu(S) &= \bigcup_{n=1}^{\infty}\mu\left(S\bigcap\left[\ \overline{B_n(0)}\ \setminus\ B_{n-1}(0)\ \right]\right)\\ &= \bigcup_{n=1}^{\infty}\left\{\left[\mu\lceil\ \overline{B_n(0)}\ \right]^{-1}\right\}^{-1}\left(S\bigcap\left[\ \overline{B_n(0)}\ \setminus\ B_{n-1}(0)\ \right]\right)\end{aligned}$$

is a union of closed sets and therefore measurable. It follows that μ sends Borel subsets of A to Borel subsets of K, and in particular, we conclude that $m\circ\mu\in A^r$ for all $m\in K^r$. We can therefore topologize A^r (differently from before) by declaring that

$$\mathcal{O}\subset A^r\text{ is open}\quad\Longleftrightarrow\quad\{m\in K^r:m\circ\mu\in\mathcal{O}\}\text{ is open.}$$

The choice of the topology on A^r will guarantee that $s\mapsto k(s)\mapsto k(s)\circ\mu\in A^r$ is measurable when $s\mapsto k(s)\in K^r$ is measurable. This will allow us to apply the condition in the definition of the set P to relaxed controls in $\mathcal{K}^r$ (cf. (6.1) below). We remark that condition (NC_5) can be dropped if we replace the positivity set P with the analogue for vector field valued inputs (i.e., if we replace ℓ and $\mathcal{A}^r$ with Λ and $\mathcal{K}^r$, respectively, in the definition of P, where the notation is as defined below).

We now show how to prove an analogue of Theorem 3.1 for problems with noncompact control sets. We assume that $A\subset\mathbb{R}^M$ is closed and that conditions (A_2) from section 2 and (NC_1)–(NC_5) are satisfied. Let w satisfy the assumptions of the theorem. Let m_ℓ be a lower bound for ℓ. Let $\hat K$ be any collection of continuous functions $\phi\times\lambda:\mathbb{R}^N\to\mathbb{R}^N\times\mathbb{R}$ each of whose members satisfies

1. $\sup\{||\phi(x)||:x\in B_R(0)\}\leq\sup\{||f(x,a)||:a\in A,x\in B_R(0)\}$ for all $R>0$,
2. $||\phi(x)-\phi(y)||\ \leq\ L\,||x-y||$ for all $x,y\in\mathbb{R}^N$,
3. $|\lambda(x)-\lambda(y)|\ \leq\ \omega(||x-y||)$ for all $x,y\in\mathbb{R}^N$,
4. $m_\ell\ \leq\ \lambda(0)\ \leq\ M_\ell$.

These conditions guarantee that $\hat K$ is equicontinuous and pointwise bounded. Applying Ascoli's theorem, it follows that $\hat K$ is precompact in $C(\mathbb{R}^N,\mathbb{R}^{N+1})$. It now follows

from assumption (NC_2) that $K := \{h_a(\cdot) := f(\cdot,a) \times \ell(\cdot,a) : a \in A\}$ is a compact subset of the metric space $C(\mathbb{R}^N, \mathbb{R}^{N+1})$. We use K as our new control set. Set

$$\Phi(x,k) = (\Pi_{\mathbb{R}^N} \circ k)(x) \quad \text{and} \quad \Lambda(x,k) := (\Pi_{\mathbb{R}} \circ k)(x) \ \ \forall \ \ k \in K,$$

where $\Pi_{\mathbb{R}^N}$ (resp., $\Pi_{\mathbb{R}}$) is the projection from K to its first N components (resp., last component). Then Φ and Λ satisfy (A_1) and (A_3), with f replaced by Φ, ℓ replaced by Λ, and A replaced by K. Set

$$\tilde{\mathcal{A}} \ = \ \{[0,\infty) \ni t \mapsto h_{\alpha(t)} : \alpha \in \mathcal{A}\}.$$

Since $a \mapsto h_a$ is continuous, this is a subset of the measurable mappings of $[0,\infty)$ into K.

Note that for each $p \in \mathbb{R}^N$, the trajectories of Φ with controls in $\tilde{\mathcal{A}}$ starting at p are exactly the trajectories of f with controls in $\mathcal{A}$ starting at p. The proof of Theorem 3.1 under the new hypotheses is then exactly as before but with ℓ replaced by Λ (using controls in $\tilde{\mathcal{A}}$ that bring points to $\mathcal{T}$ in finite time) up to (but not including) the proof that $\bar{x} \in \partial(\mathcal{S} \setminus \mathcal{T})$. We then follow the argument of Corollary 5.6, except that the weak-$\star$ limits k^r we obtain are K^r-valued. Then $\bar{x} \in \partial(\mathcal{S} \setminus \mathcal{T})$ follows since

$$\int_0^t \int_K \Lambda(\phi_x^r(s,\beta),k)\, d\beta_s(k)\, ds \ = \ \int_0^t \int_A \ell(y_x^r(s,\beta\circ\mu),a)\, d(\beta_s \circ \mu)(a)\, ds \ > \ 0 \tag{6.1}$$

for all $\beta \in \mathcal{K}^r$, $x \in P$, and $t > 0$, where $\phi_x^r(\cdot,\beta)$ is the relaxed trajectory for the relaxed dynamics Φ^r and the control $s \mapsto \beta_s$ which starts at x. We leave the proof that $\phi_x^r(\cdot,\beta) \equiv y_x^r(\cdot,\beta\circ\mu)$ to the reader.

Another approach to extending Theorem 3.1 to problems with noncompact control sets is as follows. We assume that $0 \in A \subseteq \mathbb{R}^M$, with A closed but not necessarily compact. Following [3] and [7], we also add the following hypotheses on f and ℓ, which we assume in addition to (A_1) and (A_3):

- (A_4) There exists a $\sigma \geq 1$ and, for each compact subset $\mathcal{F} \subset \mathbb{R}^N$, a constant $f_{\mathcal{F}} > 0$ so that
$$||f(x,a)|| \ \leq \ f_{\mathcal{F}}(1 + ||a||^\sigma) \ \ \forall \ \ (x,a) \in \mathcal{F} \times A.$$
- (A_5) There exist $\ell_o > 0$, $C_o \geq 0$, $\beta \in (0,1]$, $\delta_2 \geq 0$, $\bar{\ell} \geq 0$, and $\delta_1 \geq 0$ such that the following conditions hold for all $x, y \in \mathbb{R}^N$ and all $a \in A$:
  (a) $\ell(x,a) \ \geq \ \ell_o||a||^{\delta_1} - C_o$,
  (b) $|\ell(x,a) - \ell(y,a)| \ \leq \ \bar{\ell}||x-y||^\beta \left(1 + ||a||^{\delta_1} + ||x||^{\delta_2} + ||y||^{\delta_2}\right)$.

We further assume that $\delta_1 > \sigma$. Assumptions (A_4)–(A_5) are satisfied for exit time problems with linear-quadratic (LQ) data. These assumptions penalize the use of control set values of large norm.

Let $\Omega \subset \mathbb{R}^N$ be an open set containing $\mathcal{T}$, let $w \in QG(\Omega)$ be a viscosity solution of (3.1), and assume that conditions (A_1)–(A_5) are satisfied. Since A is not necessarily compact, we cannot use the weak-$\star$ convergence argument from the proof of Theorem 3.1. Instead of replacing A with a compact set of vector fields as we did above, we will apply Lemma 5.3 to compact subsets of A. We use the positivity set

$$P^\sharp \ = \ \left\{ \begin{array}{l} x \in \mathbb{R}^N : \ \int_0^t \ell^r(y_x^r(s,\alpha),\alpha(s))\, ds > 0 \ \ \text{for all} \\ t \in (0,\infty], \ \alpha \in \mathcal{K}^r, \ \text{and compact sets } K \subseteq A \end{array} \right\}$$

in place of P, and we assume that w is $(\Omega \cap P^\sharp, \mathcal{T}, \omega_o)$-compatible for some $\omega_o \in \mathbb{R} \cup \{+\infty\}$ and has locally bounded subdifferentials (cf. [2]). We recall that the *subdifferential* of a function w on a set $S \subseteq \mathbb{R}^N$ is the set-valued map $D^- w : S \to \mathcal{P}(\mathbb{R}^N)$ defined by

$$D^- w(x) \;=\; \left\{ p \in \mathbb{R}^N : \liminf_{S \ni y \to x} \frac{w(y) - w(x) - p \cdot (y - x)}{||x - y||} \geq 0 \right\},$$

and we say w has locally bounded subdifferentials, provided that for each compact set $K_1 \subseteq S$ there is a compact set $K_2 \subseteq \mathbb{R}^N$ for which $\cup \{D^- w(x) : x \in K_1\} \subseteq K_2$. Let us remark that locally Lipschitz continuous functions have locally bounded subdifferentials (cf. [2]). Let $\{\Omega_j\}$ denote the corresponding controllability sequence. We next need the following lemma, which is an immediate consequence of the proof of Proposition 2.1 of [3].

LEMMA 6.1. *Let $A \subset \mathbb{R}^M$ be a closed set containing $\vec{0}$, and assume that conditions* (A_1)–(A_5) *hold. Let w be an $(\Omega \cap P^\sharp, \mathcal{T}, \omega_o)$-compatible viscosity solution of* (3.1) *with locally bounded subdifferentials, and let $\{\Omega_j\}$ denote the associated controllability sequence. Then for each $j \in \mathbb{N}$ there exists a compact subset $K_j \subseteq A$ such that w is also a viscosity supersolution of*

$$\begin{cases} \sup\limits_{a \in K_j} \{-f(x,a) \cdot Dw(x) - \ell(x,a)\} = 0, & x \in \Omega_j \setminus \mathcal{T}, \\ w(x) = g(x), & x \in \mathcal{T}, \end{cases}$$

for each j (i.e., a supersolution of $H_{K_j}(x, Dw(x)) = 0$ on $\Omega_j \setminus \mathcal{T}$ equalling g on $\mathcal{T}$, in the notation of section 2) .

This lemma follows since there are compact sets $K_j \subseteq A$ such that $H_A \lceil [\Omega_j \times \mathcal{D}_j] \equiv H_{K_j} \lceil [\Omega_j \times \mathcal{D}_j]$, where $\mathcal{D}_j$ is a compact set large enough to contain the subdifferentials of $w \lceil \Omega_j$ (cf. section 2 for the notation). Now let $x \in \left[\Omega \cap P^\sharp\right] \setminus \mathcal{T}$, and choose a $j \in \mathbb{N}$ so that $x \in \Omega_j$ and so that the values of w near $\partial(\Omega_j) \setminus \mathcal{T}$ majorize $w(x)$, as in the proof of Theorem 3.1. Assume further that (NC_3)–(NC_4) hold, and thus $(\mathcal{E}_1)$–$(\mathcal{E}_2)$ also hold. The argument now proceeds exactly as in the "$w \geq v$" part of that proof, up through the proof of $\bar{\tau} = 0$ but with the control set A replaced by the compact set K_j, so now all the weak-$\star$ limits are well-defined K_j^r-valued relaxed controls. Then we follow the part of the proof of that theorem after the $\bar{\tau} = 0$ proof as originally stated, to get $w(x) \geq v(x)$. Noting that the compactness assumption on A can be replaced by (NC_3)–(NC_4) in the proof of Proposition 5.4 (cf. [2]), we can therefore summarize our results for problems with noncompact control sets as follows.

COROLLARY 6.2. *Let* (A_2) *hold, and let $A \subseteq \mathbb{R}^M$ be a closed set containing $\vec{0}$. Let $\Omega \subseteq \mathbb{R}^N$ be an open set containing $\mathcal{T}$, let $\omega_o \in \mathbb{R} \cup \{+\infty\}$, and let $w \in BC_{\omega_o}(\Omega)$ be a viscosity solution of* (3.1) *with locally bounded subdifferentials. Assume the following:*

1. *$\Omega \setminus \mathcal{T} \subseteq P^\sharp$.*
2. *w is $(\Omega \cap P^\sharp, \mathcal{T}, \omega_0)$-compatible, and (NC_3)–(NC_4) hold.*
3. *Either (NC_1), (NC_2), and (NC_5) hold, or else* (A_1)–(A_5) *hold with $\delta_1 > \sigma$.*

Then $w \equiv v$ on Ω.

Remark 6.3. Just as before, we can relax the hypothesis that $w \in BC_{\omega_o}(\Omega)$ to the requirements that $w \in QG(\Omega)$ and $w < \omega_o$ on Ω. Note that even if w is $(\Omega \cap P^\sharp, \mathcal{T}, \omega_o)$-compatible with the controllability sequence $\{\Omega_j\}$, the condition $(SSTC_2)$ can fail if we replace the control set A with some compact set $\tilde{K} \subset A$. On the other hand, one can show that if $SSTC(P^\sharp \cap \Omega, \mathcal{T})$ holds with A replaced by some compact set $J \subseteq A$, and if the other assumptions of Corollary 6.2 are satisfied with (A_1)–(A_5) and $\delta_1 > \sigma$,

then the restriction of v to any of the sets Ω_j coincides with the value function $v_{K_j \cup J}$ gotten by replacing the control set A with $K_j \cup J$, where the K_j are as in Lemma 6.1. Indeed, the condition $SSTC(P^\sharp \cap \Omega, \mathcal{T})$ holds for controls in $K_j \cup J$. Therefore, we can carry out the "$w \geq v$" part of the proof of Theorem 3.1 with A replaced by $K_j \cup J$ to get $w \geq v_{K_j \cup J}$ on Ω_j. Also, we have $w \leq v$ on Ω_j as before. Therefore, we get $v \geq v_{K_j \cup J} \geq v$ on Ω_j, and thus $v \equiv v_{K_j \cup J}$ on Ω_j, as claimed.

6.4. Problems with negative Lagrangians. The equality "$w = v$" in Theorem 3.1 remains true if we drop the assumption that $\Omega \setminus \mathcal{T} \subseteq P$, as long as the Lagrangian ℓ is "not very negative" and the remaining assumptions of the theorem are satisfied. By "not very negative," we mean that the following additional conditions are satisfied:

- (NVN_1) If $x \in \mathbb{R}^N$, if $\alpha \in \mathcal{A}^f(x)$, if $\{t \geq 0 : t \leq t_x(\alpha) \text{ and } y_x(t,\alpha) \notin \Omega\} \neq \emptyset$, and if we set $\lambda := \sup\{t \geq 0 : \ t \leq t_x(\alpha) \ \text{ and } \ y_x(t,\alpha) \notin \Omega\}$, then $\int_0^\lambda \ell(y_x(s,\alpha),\alpha(s))\, ds$ is nonnegative.
- (NVN_2) For each $x \in \Omega \setminus [P \cup \mathcal{T}]$, there is a bounded open set $B \subseteq \Omega$ containing x so that $\bar{B} \subseteq \Omega \setminus \mathcal{T}$ and a positive number

$$\Psi \ < \ \inf_{\alpha \in \mathcal{A}} \left\{ t > 0 : \ \mathrm{dist}(y_x(t,\alpha), \partial B) \leq \frac{1}{2} \mathrm{dist}(x, \partial B) \right\}$$

such that $y_x(\Psi,\alpha) \in P \cap \Omega$ and $\int_0^\Psi \ell(y_x(s,\alpha),\alpha(s))\, ds \geq 0$ for all $\alpha \in \mathcal{A}$.

Condition (NVN_1) takes the place of the condition $\Omega \setminus \mathcal{T} \subseteq P$ in the proof of the inequality "$w \leq v$" (cf. Proposition 5.4). The details are as follows. In the left-hand side of (5.6), we can replace $w(x)$ with $w(x) - \delta$ for some $\delta > 0$, and condition (NVN_1) guarantees that the last integral in (5.6) is $\geq -\delta$ for large enough n. This gives $w(x) \geq w(z_n)$ for large n, and then the contradiction is as before.

Condition (NVN_2) roughly means that for each $x \in \Omega \setminus [P \cup \mathcal{T}]$, there is a time Ψ such that all trajectories starting at x are in $P \cap \Omega$ at time Ψ. The proof of the reverse inequality "$w \geq v$" for cases in which $\Omega \setminus \mathcal{T} \not\subseteq P$ relies on (NVN_2). Indeed, the proof of Theorem 3.1 shows that $w(x) \geq v(x)$ for all $x \in [\Omega \cap P] \setminus \mathcal{T}$. On the other hand, if $x \in \Omega \setminus [P \cup \mathcal{T}]$, we write

$$w(x) \ = \ [w(x) \ - \ w(y_x(\Psi,\alpha'))] \ + \ w(y_x(\Psi,\alpha'))$$

for some $\alpha' \in \mathcal{A}$ and $\Psi > 0$ such that $y_x(\Psi,\alpha') \in P \cap \Omega$ and such that

$$w(x) \ - \ w(y_x(\Psi,\alpha')) \ \geq \ \int_0^\Psi \ell(y_x(s,\alpha'),\alpha'(s))\, ds \ - \ \varepsilon$$

(using Lemma 5.2). The $\hat{\alpha}$ we choose to satisfy requirement (5.7) is then the concatenation of $\alpha' \lceil [0,\Psi]$, followed by the control constructed in the proof of Theorem 3.1 for the point $y_x(\Psi,\alpha') \in P$. Note that the condition (NVN_1) holds vacuously if $\Omega = \mathcal{R} = \mathbb{R}^N$. In this way, we can extend Theorem 3.1 to cases in which $\Omega \setminus \mathcal{T} \subseteq P$ is not satisfied but the Lagrangian is "not very negative."

Acknowledgements. I would like to thank Professor H. J. Sussmann for suggesting these problems, and I would like to thank all the members of my doctoral dissertation committee, Professors D. J. Ocone, H. M. Soner, E. D. Sontag, and H. J. Sussmann, for helpful discussions and good advice.

REFERENCES

[1] Z. Artstein, *Relaxed controls and the dynamics of control systems*, SIAM J. Control Optim., 16 (1978), pp. 689–701.

[2] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser Boston, Cambridge, MA, 1997.

[3] M. Bardi and F. Da Lio, *On the Bellman equation for some unbounded control problems*, NoDEA Nonlinear Differential Equations Appl., 4 (1997), pp. 276–285.

[4] M. Bardi, M. Falcone, and P. Soravia, *Numerical methods for pursuit-evasion games via viscosity solutions*, in Stochastic and Differential Games: Theory and Numerical Methods, M. Bardi, T. E. S. Raghavan, and T. Parthasarathy, eds., Birkhäuser Boston, Cambridge, MA, 1999, pp. 105–175.

[5] M. Bardi and P. Soravia, *Hamilton-Jacobi equations with singular boundary conditions on a free boundary and applications to differential games*, Trans. Amer. Math. Soc., 325 (1991), pp. 205–229.

[6] J.-D. Benamou, *Big ray tracing: Multivalued travel time field computation using viscosity solutions of the eikonal equation*, J. Comput. Phys., 128 (1996), pp. 463–474.

[7] F. Da Lio, *On the Bellman equation for infinite horizon problems with unbounded cost functional*, Appl. Math. Optim., 41 (1999), pp. 171–197.

[8] L. C. Evans and H. Ishii, *Differential games and nonlinear first order PDE in bounded domains*, Manuscripta Math., 49 (1984), pp. 109–139.

[9] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer, New York, 1993.

[10] P.-L. Lions and P. E. Souganidis, *Differential games, optimal control and directional derivatives of viscosity solutions of Bellman's and Isaacs' equations*, SIAM J. Control Optim., 23 (1985), pp. 566–583.

[11] P.-L. Lions, E. Rouy, and A. Tourin, *Shape from shading, viscosity solutions and edges*, Numer. Math., 64 (1993), pp. 323–353.

[12] M. Malisoff, *Viscosity solutions of the Bellman equation for exit time optimal control problems with non-Lipschitz dynamics*, ESAIM Control Optim. Calc. Var., 6 (2001), pp. 415–441.

[13] R. T. Rockafeller, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[14] B. Piccoli and H. J. Sussmann, *Regular synthesis and sufficiency conditions for optimality*, SIAM J. Control Optim., 39 (2000), pp. 359–410.

[15] E. D. Sontag, *Mathematical Control Theory, Deterministic Finite Dimensional Systems*, 2nd ed., Springer-Verlag, New York, 1998.

[16] P. Soravia, *Pursuit–evasion problems and viscosity solutions of Isaacs equations*, SIAM J. Control Opt., 31 (1993), pp. 604–23.

[17] P. Soravia, *Optimality principles and representation formulas for viscosity solutions of Hamilton-Jacobi equations* I: *Equations of unbounded and degenerate control problems without uniqueness*, Adv. Differential Equations, 4 (1999), pp. 275–296.

[18] P. Soravia, *Optimality principles and representation formulas for viscosity solutions of Hamilton-Jacobi equations* II: *Equations of control problems with state constraints*, Differential Integral Equations, 12 (1999), pp. 275–293.

[19] P. Souganidis, *Two-player, zero-sum differential games and viscosity solutions*, in Stochastic and Differential Games: Theory and Numerical Methods, M. Bardi, T. E. S. Raghavan, and T. Parthasarathy, eds., Birkhäuser Boston, Cambridge, MA, 1999, pp. 69–104.

[20] H. J. Sussmann, *A general theorem on local controllability*, SIAM J. Control Optim., 25 (1987), pp. 158–194.

[21] H. J. Sussmann, *From the Brachystochrone to the maximum principle*, in Proceedings of the IEEE Conference on Decision and Control, Kobe, Japan, 1996, IEEE Publications, New York, 1996, pp. 1588–1594.

[22] H. J. Sussmann, *Geometry and optimal control*, in Mathematical Control Theory, J. Baillieul and J. C. Willems, eds., Springer-Verlag, New York, 1998, pp. 140–198.

[23] H. J. Sussmann and B. Piccoli, *Regular presynthesis and synthesis, and optimality of families of extremals*, in Proceedings of the IEEE Conference on Decision and Control, Phoenix, AZ, 1999, IEEE Publications, New York, 1999, pp. 3352–3357.

[24] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

[25] J. J. Ye, *Discontinuous solutions of the Hamilton–Jacobi equation for exit time problems*, SIAM J. Control Optim., 38 (2000), pp. 1067–1085.

[26] M. I. Zelikin and V. F. Borisov, *Theory of Chattering Control with Applications to Astronautics, Robotics, Economics, and Engineering*, Birkhäuser Boston, Cambridge, MA, 1994.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 5, pp. 1384–1405

A SINGULAR PERTURBATION APPROACH TO A RECURSIVE DECONVOLUTION PROBLEM*

FABIO FAGNANI† AND LUCIANO PANDOLFI†

Abstract. Following the approach described by A. V. Kryazhimskii and Yu. S. Osipov, we present a *recursive* algorithm which can be applied to solve the deconvolution problem of linear finite dimensional input output systems. The method gives an on line approximation of the unknown input, based on approximate samples of the output. Key features of this approach are the introduction of an associated singularly perturbed system and the use of a *quasi canonical* form due to Morse.

Key words. approximation, deconvolution, input identification, inverse problems, linear systems, singular perturbations

AMS subject classifications. 45L05, 93C05, 93E11

PII. S0363012900368259

1. Description of the problem. In this paper we study a recursive version of the deconvolution problem for a linear finite dimensional system,

$$\dot{x} = Ax + Bu\,, \qquad y = Cx, \tag{1}$$

where $x \in \mathbb{R}^q$, $u \in \mathbb{R}^m$, and $y \in \mathbb{R}^p$. The matrices have consistent dimensions and are constant. The point of view of this paper is as follows: we assume that the initial condition $x(0)$ is known. Instead, *the input function u represents an unknown input, often a disturbance, which must be (at least approximately) identified on the basis of measurements taken on the output y during a time interval* $[0, T]$. Measurements of y are available only at times $\tau_k = kT/n$, $1 \le k \le n-1$, where n is a given number, and they may be corrupted by noise. The information at time τ_k will be written as $\xi_k = y(\tau_k) + \theta_k$, where θ_k is an (unavoidable) error. We assume that the tolerance h of the error is known, i.e., that $||\theta_k|| < h$ for every k.

The input output relation which is obtained from system (1), with the initial condition $x(0) = 0$, is

$$y(t) = \int_0^t H(t,s)u(s)\ ds\,, \qquad H(t,s) = Ce^{A(t-s)}B,$$

so that our deconvolution problem fits into the framework described in [24] (see p. 3, but we shall not assume smoothness of the input u). An analogous problem is also described in [5, 18] in the context of medical applications. In fact, the deconvolution problem is ubiquitous from medical to astrophysical or geological applications.

Our analysis, however, will quite depart from what is done in the above-mentioned literature. On one hand, our assumptions are more stringent: the input output map is causal (this rules out, for example, applications as image reconstruction; see [4]); and we are assuming that the weight function is the impulse response of a time

*Received by the editors February 22, 2000; accepted for publication (in revised form) May 24, 2001; published electronically January 9, 2002. This research was supported in part by the Italian Ministero dell'Università e della Ricerca Scientifica e Tecnologica and in part by INTAS under project 96-0816. It fits the program of GNAMPA.

http://www.siam.org/journals/sicon/40-5/36825.html

†Politecnico di Torino, Dipartimento di Matematica, Corso Duca degli Abruzzi 24, I–10129 Torino, Italy (fagnani@calvino.polito.it, lucipan@polito.it).

invariant *finite dimensional* system. Our class of systems is thus smaller than the one considered, for instance, in [18, 24, 21], where more general convolution kernels are considered. Actually, if $C = I$, full state observation, the method in this paper can also be applied to some classes of distributed systems; see [14, 15]. Instead, in this paper, we are particularly interested in the case $C \neq I$, i.e., partial state observation. In our setting we will be able to use the very rich geometrical structure of linear time invariant systems, which will be exploited, in particular, in section 5. On the other hand, while in the literature the deconvolution problem is in general solved off line as a smoothing problem, we will here enforce *time recursivity*. Namely, we will impose that the reconstruction of the input u at a certain instant t will use only samples of y up to time t and, moreover, will not be updated in the future. We thus expect new and more precise results which, however, will be mostly applicable in the context of signal and control theory (see also [3, 16] and the recent paper [20]).

The recursive identification method that we are going to present is inspired by ideas of the Russian school; see [8, 9, 10, 12, 13]. Two distinctive features of this method are the introduction of a *model system* which is used in order to test candidate approximations of the input and the use of a *penalization technique*. Model systems have been used since the very beginning of identification theory (see [1, 11]). On the other hand, penalization techniques have been widely used in solving the deconvolution problem; see [18, 19, 22, 23, 24]. Differently than in [10], we shall not make any assumption on the unknown input u, a part square integrability.

Input identification is possible (at least with observations at *each* time) if and only if the map $u \to y$ is injective and this condition is equivalent to the following geometric property: the maximal controllability subspace $\mathcal{R}_*$ contained in $\ker C$ is $\{0\}$ (which implies *unknown input* state *reconstructibility*; see [3, p. 233]) and $\ker B = \{0\}$ (which allows the reconstruction of the input from the state). From now on, we will assume that these two conditions are satisfied. We notice that the subspace $\mathcal{R}_*$ can be computed using known algorithms.

In the remainder of this section we give a brief illustration of the identification algorithm we will study. We start with a "model" of our system:

$$\dot{w} = Aw + Bv\,, \qquad z = Cw\,. \tag{2}$$

As usual in identification problems, we shall compare the output z of the model to the measures ξ_k's taken on the output of the system in order to make an "optimal" choice for a function v, which should approximate the input u. The definition of v is recursive and "natural": on $[\tau_k, \tau_{k+1})$ it is defined to minimize a suitable functional of $||\xi_k - w(\tau_{k+1})||$ and $\alpha||v||_2$ (the penalization term). This is described in section 3. We stress the fact that we shall not make any a priori assumption on the magnitude or regularity of the unknown input.

A key issue which we will study in this paper is *consistency*. Our identification algorithm produces a signal v depending on n, h, and the *penalization parameter* α. We will study whether when h converges to 0, n to ∞, and α to 0, then v converges to the real input. More precisely, we will distinguish two different limiting processes: The first one is the limit for $n \to +\infty$ and $h \to 0$ and is treated in section 3. The second one for $\alpha \to 0$ is instead considered in section 4, which also contains explicit convergence estimates for the two limits. We prove, in particular, that both limits always exist, and, in particular, the second one exists thanks to the presence of a key *singular perturbation* in the underlying differential equations. (This explains the title of the paper.) This limit will be the unknown input for the special class of systems

of relative degree 1, i.e., such that $\ker CB = \{0\}$. The general case will be treated in section 5: it consists of an iteration of the basic deconvolution procedure described in sections 3 and 4, and it uses a special "quasi canonical form" proposed by Morse in [17].

2. The robustness issue. The consistency condition discussed in previous section is a fundamental property of any identification algorithm. In our setting, it ensures robustness with respect to the perturbation in the observation and to the sampling. The fact that robustness is indeed a delicate issue here can easily be seen by taking $A = 0$ in our system. The deconvolution problem then reduces to the problem of numerical differentiation, which is well known to be an ill posed problem (see [2]).

To emphasize the robustness issue, we examine in the next example the following apparently natural algorithm. Assume that (1) is controllable, and consider the following method: on the interval $[\tau_k, \tau_{k+1})$ we take the input v with minimal norm among those inputs driving the output of the model from ξ_k to ξ_{k+1}.

Example 1. We apply the proposed algorithm to the system

$$\dot{x} = u, \qquad y = x, \qquad x(0) = 0.$$

There is a unique input u which produces the given output. Let the observation interval be $[0, 1]$, and let the input, unknown to us, be $u(t) \equiv 0$. We observe at times $\tau_k = k/n$, and we get the observations $\xi_k = x(\tau_k) + \theta_k = \theta_k$, where θ_k is a disturbance unknown to us but such that $||\theta_k|| \leq 1/n$. In fact, the "real" output is $y(t) \equiv 0$. We define v_n on $[\tau_k, \tau_{k+1})$ as that control of minimal norm which transfers ξ_k to ξ_{k+1}.

It is easily computed that

$$v_{n|_{[\tau_k, \tau_{k+1})}}(t) = n[\theta_{k+1} - \theta_k]\,.$$

Now we compute v_n explicitly in the following cases: the case that $\theta_k = 1/n$ for each k or $\theta_k = (-1)^k/n$. In the first case we have

$$v_n(t) \equiv 0\,.$$

This is exactly the sought-for input u, so it seems that the proposed identification procedure is very efficient. However, in the second case we obtain

$$v_n(t) = \begin{cases} -2 & \text{on } [\tau_{2k}, \tau_{2k+1}), \\ +2 & \text{on } [\tau_{2k+1}, \tau_{2k+2}), \end{cases}$$

and this sequence of functions does not converge either pointwise or in the L^2-norm to $u(t) \equiv 0$. In fact, it converges to $u(t) \equiv 0$ but only weakly in $L^2(0, 1)$.

3. The construction of a sequence of inputs. For clarity, we collect here the standing assumptions (already described). The system we consider is

$$\dot{x} = Ax + Bu\,, \qquad y = Cx. \tag{3}$$

The matrices A, B, C are constant and known, and the unknown input u is square integrable. The initial condition x_0 is known, so it is not restrictive to assume $x_0 = 0$. Observations are available only at times $\tau_k = kT/n$ and are denoted by

$$\xi_k = y(\tau_k) + \theta_k, \qquad ||\theta_k|| < h.$$

The map from u to y is injective; i.e., the explicitly computable geometric conditions $\mathcal{R}_* = \{0\}$ (only used in section 5) and $\ker B = \{0\}$ hold. We shall also impose the (nonrestrictive) condition that C is surjective.

We now *fix* a positive number $\alpha < 1$ which, will be the penalization parameter. The quantities that we are going to construct do depend on n, h, and α. However, for the sake of simplicity, in this section dependence on α, which is fixed, will not be explicitly noted. It is also convenient to introduce $\delta = \delta_n = T/n$.

We use M in order to denote a generic constant which does not depend on n, h, α. We shall introduce also two particular constants, which do not depend on n, h, α, but which play a special role: they will be denoted by ω and η. (The constant η is introduced in the next section.)

In this section we present an algorithm which produces a candidate approximation v of the unknown input u. In order to use a simpler notation, v_n instead of $v_{n,h}$, we relate h to n, $h = h_n$. We assume $\lim h_n = 0$.

Remark 2. Notice that we are not assuming any particular functional relation between h and n since $\{h_n\}$ can be *any* infinitesimal sequence. Hence our analysis of the behavior of v is done for $n \to +\infty$ and $h \to 0$ independently. The convergence estimates at the end of section 4 will depend on n and h separately.

As said already, we are going to compare the observed quantities ξ_k's and the output z of the model system

$$\dot{w} = Aw + Bv\,, \qquad z = Cw\,. \tag{4}$$

Let us assume that we are at the instant $\tau_0 = 0$. We have the information that $x(0) = 0$ so that we shall also impose $w(0) = 0$. No further information will be available until time τ_1, and, consequently, the choice of $v_{|_{(0,\tau_1)}} \in L^2(0,\tau_1)$ will be

$$v_{|_{(0,\tau_1)}} = \arg\min \left\{ ||z(\tau_1) - 0||^2 + \alpha \int_0^{\tau_1} ||v(s)||^2 \, ds \right\}.$$

It is clear that $v_{|_{(0,\tau_1)}}$=0. We now proceed recursively. Assume we already constructed v on $(0,\tau_k)$; we *extend* it to the next interval (τ_k,τ_{k+1}) as follows: we feed v to (4) on the interval $[0,\tau_k]$, and we compute $w(\tau_k)$. Then we define v in the next interval by

$$v_{|_{(\tau_k,\tau_{k+1})}} = \arg\min \left\{ ||z(\tau_{k+1}) - \xi_k||^2 + \alpha \int_{\tau_k}^{\tau_{k+1}} ||v(s)||^2 \, ds \right\}. \tag{5}$$

It is clear that the number α is a penalization constant, as in [19, 22, 23].

In this way we construct a function v on $[0,T]$. This function does depend on n and h. As we said before, we denote it simply as v_n.

At the same time as v_n, we also construct the functions w_n and $z_n = Cw_n$, respectively, the state and output functions of the model system (4).

In order to better describe the sequences $\{v_n\}$ and $\{w_n\}$, we introduce the operators $\Lambda_{(k)}$: $L^2(\tau_k,\tau_{k+1}) \to \mathbb{R}^p$ and the adjoints $\Lambda^*_{(k)}$: $\mathbb{R}^p \to L^2(\tau_k,\tau_{k+1})$:

$$\Lambda_{(k)} v = \int_{\tau_k}^{\tau_{k+1}} Ce^{A(\tau_{k+1}-s)} Bv(s)\, ds, \qquad (\Lambda^*_{(k)} z)(t) = B^* e^{A^*(\tau_{k+1}-t)} C^* z.$$

In order to simplify some notation, we denote by $v_{(k)}$ the restriction of v_n to (τ_k,τ_{k+1}).

It is clear that $z(\tau_{k+1}) = Ce^{A\delta}w(\tau_k) + \left(\Lambda_{(k)}v_{(k)}\right)(\tau_{k+1})$ so that

$$v_{(k)} = -[\alpha I + \Lambda^*_{(k)}\Lambda_{(k)}]^{-1}\Lambda^*_{(k)}[Ce^{A\delta}w(\tau_k) - \xi_k]. \tag{6}$$

The following lemma is proved by direct computation.

LEMMA 3. *The following formulas hold:*

1. *the operator $\Lambda^*_{(k)}\Lambda_{(k)}$ acts from $L^2(\tau_k, \tau_{k+1})$ into itself and is given by*

$$[\Lambda^*_{(k)}\Lambda_{(k)}v](t) = B^*e^{A^*(\tau_{k+1}-t)}C^* \int_{\tau_k}^{\tau_{k+1}} Ce^{A(\tau_{k+1}-s)}Bv(s)\ \mathrm{d}s;$$

2. *the operator $\Lambda_{(k)}\Lambda^*_{(k)}$ acts from $\mathbb{R}^q$ into itself and is given by*

$$\Lambda_{(k)}\Lambda^*_{(k)} = \int_0^\delta Ce^{As}BB^*e^{A^*s}C^*\ \mathrm{d}s = R,$$

   independent of k *(but it does depend on n since it depends on $\delta = T/n$, $R = R_n$);*

3. $[\alpha I + \Lambda^*_{(k)}\Lambda_{(k)}]^{-1}\Lambda^*_{(k)} = \Lambda^*_{(k)}[\alpha I + \Lambda_{(k)}\Lambda^*_{(k)}]^{-1}$.

With the notation just introduced, the equation of the model on (τ_k, τ_{k+1}) is given by

$$\dot{w}_n = Aw_n - B[\alpha I + \Lambda^*_{(k)}\Lambda_{(k)}]^{-1}\Lambda^*_{(k)}[Ce^{A\delta}w_n(\tau_k) - \xi_k], \qquad z_n(t) = Cw_n(t). \tag{7}$$

We analyze the behavior of the sequence $\{w_n\}$ at the sampling times τ_k.

$$\begin{aligned}
&w_n(\tau_{k+1}) = e^{A\delta}w_n(\tau_k)\\
&-\textstyle\int_{\tau_k}^{\tau_{k+1}} e^{A(\tau_{k+1}-s)}B\left([\alpha I + \Lambda^*_{(k)}\Lambda_{(k)}]^{-1}\Lambda^*_{(k)}[Ce^{A\delta}w_n(\tau_k) - \xi_k]\right)(s)\ ds\\
&= e^{A\delta}w_n(\tau_k)\\
&-\textstyle\int_{\tau_k}^{\tau_{k+1}} e^{A(\tau_{k+1}-s)}B\left(\Lambda^*_{(k)}[\alpha I + \Lambda_{(k)}\Lambda^*_{(k)}]^{-1}[Ce^{A\delta}w_n(\tau_k) - \xi_k]\right)(s)\ ds\\
&= e^{A\delta}w_n(\tau_k)\\
&-\left(\textstyle\int_{\tau_k}^{\tau_{k+1}} e^{A(\tau_{k+1}-s)}BB^*e^{A^*(\tau_{k+1}-s)}C^*\ ds\right)\cdot\left([\alpha I + R]^{-1}[Ce^{A\delta}w_n(\tau_k) - \xi_k]\right)\\
&= \{I - KC^*[\alpha I + R]^{-1}C\}e^{A\delta}w_n(\tau_k) - KC^*(\alpha I + R)^{-1}\xi_k,
\end{aligned} \tag{8}$$

where K is the controllability operator,

$$K = \int_0^\delta e^{As}BB^*e^{A^*s}\ ds.$$

We note that K, as well as R, depends on δ, i.e., on n and that $||K|| < M \cdot \delta = M/n$.

We write the equality of the expression in (8) as follows:

$$w_n(\tau_{k+1}) = Hw_n(\tau_k) + f_k, \qquad \text{where}$$

$$\begin{cases} H &= \{I - KC^*[\alpha I + R]^{-1}C\}e^{A\delta},\\ f_k &= -KC^*(\alpha I + R)^{-1}\xi_k. \end{cases}$$

The crucial fact is that $\sup_k ||\xi_k|| < M$, independent of n, h, α. This implies that we can write

$$||f_k|| \le \frac{M}{n\alpha} \qquad \forall\, n\,,\ \alpha. \tag{9}$$

In order to estimate H, we consider the following inequality which will be repeatedly used in what follows:

$$||e^{A^* s} - I|| \leq \frac{M}{n} \qquad \forall s \in (0, 1/n). \tag{10}$$

We obtain

$$\begin{cases} ||H|| \leq 1 + \frac{\omega}{\alpha n} & \text{in general,} \\ ||H|| \leq 1 + \frac{M}{n} & \text{if } C = I. \end{cases} \tag{11}$$

The first inequality follows from (10), and the second follows from (10) and the fact that when $C = I$ the matrix $\{I - KC^*[\alpha I + R]^{-1}C\} = \{I - K[\alpha I + K]^{-1}\}$ is a contraction.

We are now ready to prove the following result.

THEOREM 4. *For every fixed $\alpha < 1$, the sequence $\{w_n(\tau_k)\}$ is uniformly bounded in n and k, and, for every $k < n$, we have the following estimates, where M and ω do not depend on n, h, α:*

$$\begin{cases} ||w_n(\tau_k)|| \leq M e^{\omega/\alpha} & \textit{in general,} \\ ||w_n(\tau_k)|| \leq M/\alpha & \textit{if } C = I. \end{cases} \tag{12}$$

Proof. We prove the inequality in the general case. We have

$$||w_n(\tau_k)|| \leq \big(\sup_k ||f_k||\big) \cdot \sum_{j=0}^{k} ||H||^j.$$

Using inequalities (11) and (9), we obtain

$$\begin{aligned} ||w_n(\tau_k)|| &\leq \frac{M}{\alpha n} \cdot \sum_{j=0}^{k} \Big(1 + \frac{\omega}{\alpha n}\Big)^j = \frac{M}{\alpha n} \cdot (\alpha n) \cdot \frac{(1 + \frac{\omega}{\alpha n})^{k+1} - 1}{\omega} \\ &\leq \frac{M}{\omega} \cdot \Big(1 + \frac{\omega}{\alpha n}\Big)^{k+1} \leq \frac{M}{\omega} \cdot \Big(1 + \frac{\omega}{\alpha n}\Big)^{n} \leq \frac{M}{\omega} e^{\omega/\alpha}, \end{aligned}$$

as wanted.

The proof in the case $C = I$ follows in a similar way by using the second inequality in (11). □

Remark 5.

(a) Estimate (12) in the case when $C \neq I$ is very weak. It will be used only for theoretical reasons since we shall prove in section 5 that the general case $C \neq I$ can be reduced to the study of a chain of simplest cases in which $C = I$. Hence, for practical purposes only, the second inequality in (12) is relevant.

(b) Inequality (12) was given only for $k < n$. If $k = n$, a further factor $[1+\omega/(\alpha n)]$ appears which, however, will not create any problem since the case of interest is when α converges to zero slowly (see Remarks 12 and 26).

In order to study the asymptotics of the whole sequence $\{w_n\}$, it is convenient to introduce further operators and to establish a number of continuity results for them.

We define Λ_n: $L^2(0,T) \to \mathbb{R}^{nq}$ as

$$(\Lambda_n f) = \mathrm{col}[\Lambda_{(0)} f_{|_{(0,\tau_1)}}, \ldots, \Lambda_{(n-1)} f_{|_{(\tau_{n-1},T)}}],$$

and we compute Λ_n^*: $\mathbb{R}^{nq} \to L^2(0,T)$:

$$\Lambda_n^*(x_0,\dots,x_{n-1}) = \bigwedge_{k=0}^{n-1} \Lambda_{(k)}^* x_k,$$

where the wedge $\wedge$ denotes concatenation. It follows the equality

$$(\Lambda_n^*\Lambda_n f)_{|_{(\tau_k,\tau_{k+1})}} = \Lambda_{(k)}^*\Lambda_{(k)} f_{|_{(\tau_k,\tau_{k+1})}}.$$

Next we introduce the space $\mathcal{K}_n(0,T)$ of the q-vector valued functions which are piecewise constant, with n jumps at most, located at the points τ_k. This space is a subspace of $L^2(0,T)$ isomorphic to $\mathbb{R}^{nq}$. We now introduce the operator $\Gamma_n : \mathcal{K}_n(0,T) \to \mathbb{R}^{qn}$, defined as follows. We fix arbitrary points $s_k \in (\tau_k, \tau_{k+1})$, and we put

$$\Gamma_n f = \mathrm{col}[f(s_0), f(s_1), \dots, f(s_{n-1})]. \tag{13}$$

We note that the space over which the operator Γ_n is defined is itself a function of n and that we have

$$||\Gamma_n f||_{\mathbb{R}^{nq}} = \sqrt{n}||f||_2. \tag{14}$$

We now study the properties of these operators. We have the following lemma.

LEMMA 6. *Consider a sequence of functions $\{g_n\}$ with $g_n \in \mathcal{K}_n(0,T)$. We have the following:*

1. *if the sequence $\{g_n\}$ is uniformly bounded on $[0,T]$, then the sequence $\{\Lambda_n^*\Gamma_n g_n\}$ is also uniformly bounded on $[0,T]$;*
2. *if there exists a bounded function g such that $\sup_t ||g_n(t) - g(t)|| \le M_0$, then*

$$\sup_t ||(\Lambda_n^*\Gamma_n g_n)(t) - B^*C^*g(t)|| \le M \cdot \left\{ \frac{1}{n} + M_0 \right\}.$$

Proof. We prove item 1. We have, from (13),

$$\begin{aligned} &\sup_t ||\left(\Lambda_n^*\Gamma_n g_n\right)(t)|| = \sup_k \sup_{t\in(\tau_k,\tau_{k+1})} ||\Lambda_{(k)}^* g_n(s_k)|| \\ &= \sup_k \sup_{t\in(\tau_k,\tau_{k+1})} ||B^* e^{A^*(\tau_{k+1}-t)} C^* g_n(s_k)|| \le M \cdot ||B|| \cdot ||C|| \sup_{s\in(0,\delta)} ||e^{A^*s}||, \end{aligned}$$

as wanted.

Now we prove item 2. We have

$$\begin{aligned} &||(\Lambda_n^*\Gamma_n g_n)(t) - B^*C^*g(t)|| \le ||(\Lambda_n^*\Gamma_n g_n)(t) - B^*C^*g_n(t)|| \\ &+||B^*C^*[g_n(t) - g(t)]||. \end{aligned} \tag{15}$$

The first addendum is equal to $||B^*(e^{A^*(\tau_{k+1}-t)} - I)C^*g_n(t)||$. (We recall that g_n is constant on (τ_k, τ_{k+1}).) Using (15) and (10) and the fact that $\{g_n\}$ is uniformly bounded, we obtain that, on each interval (τ_k, τ_{k+1}),

$$||(\Lambda_n^*\Gamma_n g_n)(t) - B^*C^*g(t)|| \le M \cdot \left\{ \frac{1}{n} + \sup_{t\in[0,T]} ||g_n(t) - g(t)|| \right\}, \tag{16}$$

as wanted. ∎

LEMMA 7. *The operators* $\Lambda_n : L^2(0,T) \to \mathbb{R}^{nq}$ *satisfy the following:*

1. $||\Lambda_n|| \to 0$.
2. *There exists a constant* M *such that* $\sup_{n,t} ||(\Lambda_n^* X)(t)|| \le M||X||$ *for every vector* $X \in \mathbb{R}^{nq}$. *The constant* M *does not depend on* n, h, α.

Proof. In order to prove item 1, we estimate $||\Lambda_n^*||$ first. Let $X = \mathrm{col}(x_0, x_1, \ldots, x_{n-1})$ be a unitary vector. We have

$$||\Lambda_n^* X||_2^2 = \sum_{k=0}^{n-1} ||\Lambda_{(k)}^* x_k||_2^2 = \sum_{k=0}^{n-1} \int_{\tau_k}^{\tau_{k+1}} ||B^* e^{A^*(\tau_{k+1}-s)} C^* x_k||^2 \, ds$$

$$\le \delta \sum_{k=0}^{n-1} \sup_{s\in[\tau_k,\tau_{k+1}]} \left(||B^* e^{A^*(\tau_{k+1}-s)} C^*||\right)^2 \cdot ||x_k||^2 \le M \cdot \delta \cdot ||X||^2 . \tag{17}$$

This yields

$$||\Lambda_n|| = ||\Lambda_n^*|| \le \frac{M}{n} \to 0 . \tag{18}$$

The second statement has an analogous proof. □

Finally, we prove a further pointwise estimate.

LEMMA 8. *Let* $\{g_n\}$ *be a sequence of functions,* $g_n \in \mathcal{K}_n(0,T)$. *We have the following:*

1. *If the sequence* $\{g_n\}$ *is uniformly bounded, then* $[\alpha I + \Lambda_n^* \Lambda_n]^{-1} g_n$ *is a piecewise continuous function and*

$$\sup_t ||([\alpha I + \Lambda_n^* \Lambda_n]^{-1} g_n)(t)|| < M \cdot \left(\frac{1}{\alpha} + \frac{1}{\sqrt{n}\alpha^2}\right) \cdot ||g_n||_2 .$$

   The constant M *does not depend on* n, h, α.
2. *If* $\{g_n\}$ *uniformly converges to* $g \in C(0,T)$, *then the sequence* $\{[\alpha I + \Lambda_n^* \Lambda_n]^{-1} g_n\}$ *converges to* $\frac{1}{\alpha} g$ *uniformly on* $[0,T]$.

Proof. Let $\psi_n = [\alpha I + \Lambda_n^* \Lambda_n]^{-1} g_n$. We have

$$\alpha \psi_n + \Lambda_n^* \Lambda_n \psi_n = g_n \tag{19}$$

so that ψ_n is a piecewise continuous function.

Taking the L^2 inner product of (19) by ψ_n and using the positivity of $\Lambda_n^* \Lambda_n$, we get

$$||\psi_n||_2 \le \frac{||g_n||_2}{\alpha}.$$

If $t \in (\tau_k, \tau_{k+1})$, we have

$$||(\Lambda_n^* \Lambda_n \psi_n)(t)|| = \left|\left| B^* e^{A^*(\tau_{k+1}-t)} C^* \int_{\tau_k}^{\tau_{k+1}} C e^{A(\tau_{k+1}-s)} B \psi_n(s) \, ds \right|\right|$$

$$\le M \cdot \sqrt{\delta} \cdot ||\psi_n||_2 \le \frac{M \cdot \sqrt{\delta}}{\alpha} ||g_n||_2. \tag{20}$$

It follows from (19) and (20) that

$$||\psi_n(t)|| \le M \left(\frac{1}{\alpha} + \frac{\sqrt{\delta}}{\alpha^2}\right) ||g_n||_2.$$

This proves item 1.

To prove item 2, we evaluate

$$\alpha\psi_n - g = [\alpha\psi_n - g_n] + [g_n - g] = -[\Lambda_n^*\Lambda_n\psi_n] + [g_n - g]. \tag{21}$$

It follows from (20), considering that $||g_n||_2$ is bounded, that

$$\sup_t ||(\Lambda_n^*\Lambda_n\psi_n)(t)|| \le \frac{M}{\alpha\sqrt{n}}. \tag{22}$$

From (21) and (22), we obtain the explicit estimate

$$\left|\left| [\,(\alpha I + \Lambda_n^*\Lambda_n)^{-1}g_n](t) - \frac{1}{\alpha}g(t)\right|\right| \le \frac{M}{\alpha^2\sqrt{n}} + \frac{1}{\alpha}\sup_t ||g_n(t) - g(t)||. \tag{23}$$

The right-hand side converges to zero for $n \to +\infty$ for each fixed α. □

Now we can study the sequence $\{w_n\}$. We prove the following theorem.

THEOREM 9. *The sequence $\{w_n\}$ is relatively compact in $C(0,T)$.*

Proof. We consider again equality (7). This can be written as

$$\dot{w}_n = Aw_n - B[\alpha I + \Lambda_n^*\Lambda_n]^{-1}\Lambda_n^*\Gamma_n f_n, \tag{24}$$

where $f_n \in \mathcal{K}_n(0,T)$ is given by

$$f_{n\,|_{(\tau_k,\tau_{k+1})}} = Ce^{A\delta}w_n(\tau_k) - \xi_k. \tag{25}$$

We know from Theorem 4 that $\{w_n(\tau_k)\}$ is bounded (for fixed α) so that $\{f_n\}$ is bounded too. Using Lemma 6, item 1, and Lemma 8, item 1, we see that the affine term in (24) is bounded. This implies that the sequence $\{w_n\}$ and, therefore, also $\{\dot{w}_n\}$, is uniformly bounded. Compactness then follows from the Ascoli–Arzelà theorem. □

For future use we note the explicit inequality

$$||\dot{w}_n(t)|| \le K = K(\alpha). \tag{26}$$

Finally, we prove the following theorem.

THEOREM 10. *The sequence $\{w_n\}$ converges uniformly to the unique solution w^α of the problem*

$$\dot{w} = \left[A - \frac{BB^*C^*C}{\alpha}\right]w + \frac{BB^*C^*C}{\alpha}x, \qquad w(0) = 0. \tag{27}$$

Proof. We show that every limit point in $C(0,T)$ of $\{w_n\}$ coincides with w^α. We consider a subsequence of $\{w_n\}$, still denoted $\{w_n\}$, which converges to a limit w.

Since $w_n(0) = 0$, we can write

$$w_n(t) = \int_0^t e^{A(t-s)}\left[\frac{B}{\alpha}\left[I + \frac{1}{\alpha}\Lambda_n^*\Lambda_n\right]^{-1}\Lambda_n^*\Gamma_n f_n\right]\,ds.$$

(The sequence $\{f_n\}$ is defined in (25).)

Consider now the bounded sequence $\{f_n\}$. We prove that

$$\lim_n \sup_{t\in[0,T]} ||f_n(t) - Cw(t) - y(t)|| = 0\,. \tag{28}$$

In fact, the supremum in (28) is less than

$$\sup_k \sup_{t\in(\tau_k,\tau_{k+1})} ||C[e^{A\delta}w_n(\tau_k) - w(t)]|| + \sup_k \sup_{t\in(\tau_k,\tau_{k+1})} ||\xi_k - y(t)||\,.$$

On one hand, if t belongs to (τ_k, τ_{k+1}), we have

$$||e^{A\delta}w_n(\tau_k) - w(t)|| \le ||e^{A\delta}w_n(\tau_k) - w_n(t)|| + ||w_n(t) - w(t)||,$$

and both right-hand terms converge uniformly to zero because $w_n \to w$ uniformly by assumption, $e^{A\delta} = e^{A/n} \to I$, and the sequence of functions $\{w_n\}$ is bounded and equicontinuous, thanks to (26).

On the other hand,

$$||\xi_k - y(t)|| \le ||\xi_k - y(\tau_k)|| + ||y(\tau_k) - y(t)||\,.$$

The first addendum converges to zero since it is dominated by $h_n \to 0$. The second term converges to zero because the integral $C\int_0^t e^{A(t-s)}Bu(s)\ ds$ is absolutely continuous on $[0,T]$. This proves (28).

We recall that we are considering a convergent subsequence (in the uniform norm), still denoted $\{w_n\}$, which converges to w. We apply Lemma 6, and we see that

$$\Lambda_n^*\Gamma_n f_n \to B^*C^*[Cw - y]$$

uniformly. Hence, from Lemma 7, we have that the integral converges to the solution of (27), as wanted. □

We can also give a pointwise estimate of the convergence illustrated in Theorem 10 as follows.

LEMMA 11. *For each $\alpha < 1$ and for each $t \in [0,T]$ we have*

$$\text{(29)}\qquad \begin{cases} ||w_n(t) - w^\alpha(t)|| \le \frac{Me^{\omega/\alpha}}{\sqrt{n\alpha}} & \text{in general,} \\ ||w_n(t) - w^\alpha(t)|| \le \frac{M}{\sqrt{n\alpha^2}} & \text{if } C = I. \end{cases}$$

Proof. We compute

$$w_n(t) - w^\alpha(t) = \lim_m \int_0^t e^{A(t-s)}B\big\{[\alpha I + \Lambda_n^*\Lambda_n]^{-1} - [\alpha I + \Lambda_m^*\Lambda_m]^{-1}\big\}\Lambda_n^*\Gamma_n f_n\ ds + \int_0^t e^{A(t-s)}B[\alpha I + \Lambda_m^*\Lambda_m^*]^{-1}\left\{\Lambda_n^*\Gamma_n f_n - \Lambda_m^*\Gamma_m f_n\right\}\ ds.$$

Let $s_k \in (\tau_k, \tau_{k+1})$ as in (13). For $m > n$ we have the following estimates (we use (18) and (14), (25), and Theorem 4):

$$\begin{aligned} &\left|\left|\int_0^t e^{A(t-s)}B\left\{[\alpha I + \Lambda_n^*\Lambda_n]^{-1} - [\alpha I + \Lambda_m^*\Lambda_m]^{-1}\right\}\Lambda_n^*\Gamma_n f_n\ ds\right|\right| \\ &\le M\big\{||(\alpha I + \Lambda_n^*\Lambda_n)^{-1}|| + ||[\alpha I + \Lambda_m^*\Lambda_m]^{-1}||\big\}||\Lambda_n^*\Gamma_n f_n||_2 \\ &\le \frac{2M}{\alpha}\frac{1}{n}||\Gamma_n f_n||_2 \le \frac{2M}{\alpha\sqrt{n}}e^{\omega/\alpha}, \end{aligned}$$

where we used $||w_n(\tau_k)|| + ||\xi_k|| \le Me^{\omega/\alpha}$ for $\alpha < 1$.

The second addendum is estimated as follows:

$$\left|\left|\int_0^t e^{A(t-s)}B[\alpha I+\Lambda_m^*\Lambda_m]^{-1}\{\Lambda_n^*\Gamma_n f_n-\Lambda_m^*\Gamma_m f_m\}\,ds\right|\right|$$
$$\leq \frac{M}{\alpha n}\{||\Gamma_n f_n||+||\Gamma_m f_m||\}\leq \frac{2M}{\sqrt{n}\alpha}e^{\omega/\alpha}.$$

In both the estimates we replace $e^{\omega/\alpha}$ with $1/\alpha$ if $C=I$. The required estimates follow. □

Remark 12. The previous lemma implies, in particular, that if α converges to zero slowly in such a way that $\alpha^2\sqrt{n}>M>0$, then, in the case when $C=I$, the sequence $\{w_n\}$ remains bounded. This strengthens the result in Theorem 4.

We now consider the sequence $\{v_n\}$. Taking into account the definition (6) of $v_{(k)}$ and the definitions of Γ_n, Λ_n, we can write

$$v_n=-[\alpha I+\Lambda_n^*\Lambda_n]^{-1}\Lambda_n^*\Gamma_n f_n. \tag{30}$$

It follows from Lemma 6 that, for $n\to+\infty$, $\Lambda_n^*\Gamma_n f_n\to B^*C^*C[w^\alpha-x]$, and it follows from Lemma 7 that $||\Lambda_n^*\Lambda_n||\to 0$. This yields the following result.

THEOREM 13. *We have*

$$\lim_n v_n=v_\alpha,\qquad \textit{where}\qquad v_\alpha=-\frac{1}{\alpha}B^*C^*C[w^\alpha-x]\,. \tag{31}$$

Convergence in the previous theorem is meant to be in $L^2(0,T)$, but later we shall also give pointwise estimates.

4. Input identification in a special case. In this section we consider the convergence of the solutions of system (27) when $\alpha\to 0$. We put

$$\tilde{J}=BB^*C^*C.$$

With this notation, the "error" $e^\alpha=w^\alpha-x$ satisfies the system with *singular perturbations*

$$\dot{e}^\alpha=\dot{w}^\alpha-\dot{x}=\left(A-\frac{\tilde{J}}{\alpha}\right)e^\alpha-Bu. \tag{32}$$

We study the convergence properties of e^α when $\alpha\to 0$.

It is easy to see that the matrix $\tilde{J}=BB^*C^*C$ is diagonalizable and its eigenvalues are nonnegative. With no loss of generality, we thus assume that $\tilde{J}$ is diagonal, and we decompose

$$\tilde{J}=\begin{bmatrix} J & 0\\ 0 & 0\end{bmatrix},$$

where the matrix J is diagonal and has *positive* determinant. We represent, accordingly,

$$A=\begin{bmatrix} A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix},\qquad B=\begin{bmatrix} B_+\\ B_-\end{bmatrix}.$$

We recall the Wintner–Ważeski inequalities (see [6, p. 25]): if G is any square matrix and σ_-, σ_+ are the minimum and maximum eigenvalues of $(G+G^*)/2$, then we have, for each $t \geq 0$,

$$e^{\sigma_- t} \leq ||e^{Gt}|| \leq e^{\sigma_+ t}.$$

First we prove the following lemma.

LEMMA 14. *There exists a positive number M such that*

$$||e^{\alpha}(t)|| \leq M \qquad \forall t \in [0,T], \quad \forall \alpha > 0.$$

Proof. We apply the Wintner–Ważeski inequalities to the matrix $G = A - \frac{1}{\alpha}\tilde{J}$. The largest eigenvalue of $[(A-\frac{1}{\alpha}\tilde{J})+(A-\frac{1}{\alpha}\tilde{J})^*]/2$ is smaller than the largest eigenvalue of the matrix $(A+A^*)/2$, call it μ, for every α. Hence we have

$$||e^{[A-\frac{1}{\alpha}\tilde{J}]t}|| \leq e^{\mu t} \qquad \forall t \in [0,T], \quad \forall \alpha > 0. \tag{33}$$

The result follows since the input u is fixed. □

We decouple (32) as

$$\begin{cases} \dot{e}_+^{\alpha} = (A_{11} - \frac{1}{\alpha}J)e_+^{\alpha} + A_{12}e_-^{\alpha} - g_+, \\ \dot{e}_-^{\alpha} = A_{21}e_+^{\alpha} + A_{22}e_-^{\alpha} - g_-, \end{cases} \quad \text{where} \quad g = \begin{bmatrix} g_+ \\ g_- \end{bmatrix} = Bu. \tag{34}$$

We now investigate the behavior of e_+^{α} and of e_-^{α} for $\alpha \to 0$. We shall repeatedly use the next result.

LEMMA 15. *There exists a number $\eta > 0$ such that for each $t > 0$, $\alpha > 0$ small enough, we have*

$$||e^{(A_{11}-J/\alpha)t}|| \leq e^{-(\eta/\alpha)t}.$$

Proof. We apply the Wintner–Ważeski inequalities to the matrix $A_{11} - J/\alpha$. Notice that $\frac{1}{2}\{[A_{11}-J/\alpha]+[A_{11}-J/\alpha]^*\} = \frac{1}{\alpha}[-J+\alpha(A_{11}+A_{11}^*)]$ is *negative definite* for small α, and its maximum eigenvalue $-\mu$ can be estimated as follows: $-\mu < -\eta/\alpha$, where $\eta > 0$ is any number less than all the eigenvalues of J. This gives the required estimate. □

Now we prove the following lemma.

LEMMA 16.

$$\lim_{\alpha \to 0} e_+^{\alpha}(t) = 0$$

uniformly in $[0,T]$.

Proof. It follows from Lemmas 14 and 15 and from an application of the Schwarz inequality that

$$||e_+^{\alpha}(t)|| = \left|\left| \int_0^t e^{[A_{11}-J/\alpha](t-s)}[A_{12}e_-^{\alpha}(s) - g_+(s)]\, ds \right|\right| \leq M \cdot \sqrt{\alpha}. \tag{35}$$

□

Now define e_-^0 as the solution of

$$\dot{e} = A_{22}e - g_-, \qquad e(0) = 0. \tag{36}$$

We have the following result.

LEMMA 17.

$$\lim_{\alpha\to 0} e_-^\alpha = e_-^0, \qquad \lim_{\alpha\to 0} \dot{e}_-^\alpha = \dot{e}_-^0$$

uniformly on $[0,T]$.

Proof. It follows from Lemma 16 that

$$e_-^\alpha(t) = \int_0^t e^{A_{22}(t-s)}\{A_{21}e_+^\alpha(s) - g_-(s)\} \to -\int_0^t e^{A_{22}(t-s)}g_-(s)\ ds. \tag{37}$$

This proves that e_-^α converges uniformly to e_-^0 on $[0,T]$. Convergence of the derivatives follows (37), Lemma 16, and (34). □

We now focus on the limit of $\frac{1}{\alpha}e^\alpha(t)$. The following is the main result of this section.

THEOREM 18. *We have*

$$\lim_{\alpha\to 0}\left[-\frac{1}{\alpha}Je_+^\alpha\right] = g_+ - A_{12}e_-^0.$$

The convergence is

- *in* $L^2(0,T)$ *if* u *is square integrable,*
- *uniform on* $[\epsilon,T]$ *for every* $\epsilon > 0$ *if* u *is absolutely continuous,*
- *uniform on* $[0,T]$ *if* u *is absolutely continuous and* $u(0) = 0$.

Proof. We proceed in a number of steps, first for absolutely continuous inputs and then for square integrable inputs.

Absolutely continuous inputs. In this case u and, consequently, g are *bounded* on $[0,T]$.

We start with a preliminary lemma.

LEMMA 19. *Let* g *be bounded. Then there exists a constant* M *such that*

$$\left|\left|\frac{1}{\alpha}e_+^\alpha(t)\right|\right| \le M \qquad \forall t\in[0,T], \quad \forall\alpha > 0.$$

Consequently,

$$||\dot{e}_+^\alpha(t)|| \le M \qquad \forall t\in[0,T], \quad \forall\alpha > 0.$$

Proof. In the following computation we use Lemma 15.

$$\begin{aligned}\left|\left|\frac{1}{\alpha}e_+^\alpha(t)\right|\right| &= \left|\left|\int_0^t \frac{1}{\alpha}e^{(A_{11}-\frac{1}{\alpha}J)(t-s)}[A_{12}e_-^\alpha(s) - g_+(s)]\ ds\right|\right| \\ &\le \int_0^t \frac{1}{\alpha}e^{-\eta(t-s)/\alpha}||A_{12}e_-^\alpha(s) - g_+(s)||\ ds \\ &\le \frac{1}{k}[1-e^{-\eta t/\alpha}]\sup_{s\in[0,T]}||A_{12}e_-^\alpha(s) - g_+(s)||.\end{aligned}$$

This is bounded since g_+ is bounded by assumption and e_-^α is bounded on $[0,T]$ from Lemma 14. Finally, boundedness of the derivative follows from (34). □

Remark 20. Lemma 19 and (37) give, when u is bounded,

$$||e_-^\alpha(t) - e_-^0(t)|| \le M\cdot\alpha \qquad \forall t\in[0,T], \quad \forall\alpha > 0. \tag{38}$$

Now we are ready to prove Theorem 18 for absolutely continuous inputs. We write the equation of e_+^α as

$$\dot{e}_+^\alpha = -\frac{1}{\alpha} J e_+^\alpha + \{-g_+ + A_{11} e_+^\alpha + A_{12} e_-^\alpha\}.$$

We have

$$\begin{aligned}
&\frac{1}{\alpha} J e_+^\alpha(t) = \int_0^t \left[\frac{d}{ds} e^{-\frac{1}{\alpha} J(t-s)}\right] \left[A_{11} e_+^\alpha(s) + A_{12} e_-^\alpha(s) - g_+(s)\right] \, ds \\
&= A_{11} e_+^\alpha(t) + A_{12} e_-^\alpha(t) - g_+(t) + e^{-\frac{1}{\alpha} Jt} g_+(0) \\
&- \int_0^t e^{-\frac{1}{\alpha} J(t-s)} \left[A_{11} \dot{e}_+^\alpha(s) + A_{12} \dot{e}_-^\alpha(s) - \dot{g}_+(s)\right] \, ds. \qquad (39)
\end{aligned}$$

As in the proof of Lemma 16, we see that the integral converges to zero uniformly on $[0,T]$. On the other hand, we know from Lemma 16 that e_+^α converges to zero uniformly on $[0,T]$ and, from Lemma 17, that e_-^α converges uniformly to e_-^0. Hence we have convergence to $-g_+ + A_{12} e_-^0$ uniformly on $[\epsilon, T]$ for any $\epsilon > 0$ and on $[0,T]$ if $u(0)$; hence $g_+(0)$, is zero.

The previous computation, combined with Lemma 19 and (38), gives the following pointwise estimate when $u \in W^{1,2}$:

$$\left|\left| \frac{1}{\alpha} J e_+^\alpha(t) + g_+(t) - A_{12} e_-^0(t) \right|\right| \leq M \cdot \left(\sqrt{\alpha} + e^{-\eta t/\alpha} ||g_+(0)||\right). \qquad (40)$$

Square integrable inputs. We now prove Theorem 18 for inputs which are assumed only to be in $L^2(0,T)$. We introduce the functions

$$(41) \begin{cases} z^\alpha(t) = -g_+(t) + A_{11} e_+^\alpha(t) + A_{12} e_-^\alpha(t), \quad t \in [0,T], \quad z^\alpha(t) = 0 \text{ otherwise;} \\ z^0(t) = -g_+(t) + A_{12} e_-^0(t), \quad t \in [0,T], \quad z^0(t) = 0 \text{ otherwise;} \\ \zeta^\alpha(t) = \frac{1}{\alpha} \int_0^t J e^{-\frac{J}{\alpha}(t-s)} z^\alpha(s) ds, \qquad t \in \mathbb{R}. \end{cases}$$

By definition, z^α converges to z^0 in $L^2(\mathbb{R})$, and ζ^α is the convolution of z^α with the function which is zero for $t < 0$ and $\frac{1}{\alpha} J e^{-Jt/\alpha}$ for $t > 0$. The function ζ^α belongs to $L^2(\mathbb{R})$ since the eigenvalues of J are *positive* and $\alpha > 0$.

We must prove that ζ^α converges to z^0 in $L^2(0,T)$, and this is implied by the $L^2(-\infty, +\infty)$ convergence of $\chi\zeta^\alpha$ to χz^0, where $\chi(t) = e^{-rt}$, since both $\zeta^\alpha(t)$ and $z^0(t)$ are zero for $t < 0$. Here r is a *fixed* positive number.

In order to prove $L^2(-\infty, +\infty)$ time domain convergence, we prove equivalently $L^2(-\infty, +\infty)$ convergence of the trace on the imaginary axis of the Laplace transforms. The trace of the Laplace transform of $\chi\zeta^\alpha$ is the function

$$\frac{1}{\alpha} J \left[(i\omega + r) I + \frac{J}{\alpha}\right]^{-1} \hat{z}^\alpha(i\omega + r).$$

We must prove

$$\lim_{\alpha \to 0} \int_{-\infty}^{+\infty} \left|\left| \mathcal{L}[e^{-r\cdot} \zeta^\alpha(\cdot) - e^{-r\cdot} z^0(\cdot)] \right|\right|^2 \, d\omega = 0. \qquad (42)$$

We have

$$\left[\int_{-\infty}^{+\infty} \left|\left|\mathcal{L}[e^{-r\cdot}\zeta^\alpha(\cdot) - e^{-r\cdot}z^0(\cdot)]\right|\right|^2 \, d\omega\right]^{1/2} \tag{43}$$

$$\leq \left[\int_{-\infty}^{+\infty} \left|\left| \frac{1}{\alpha} J \left[(i\omega + r)I + \frac{J}{\alpha}\right]^{-1} \left[\hat{z}^\alpha(i\omega + r) - \hat{z}^0(i\omega + r)\right] \right|\right|^2 \, d\omega\right]^{1/2} \tag{44}$$

$$+\left[\int_{-\infty}^{+\infty} \left|\left| \left[\frac{1}{\alpha} J \left[(i\omega + r)I + \frac{J}{\alpha}\right]^{-1} - I\right] \hat{z}^0(i\omega + r) \right|\right|^2 \, d\omega\right]^{1/2}. \tag{45}$$

We prove a lemma first.

LEMMA 21. *There exists a number M such that for every λ with $\Re e\, \lambda > 0$ and every $\alpha > 0$ we have*

$$\left|\left| \frac{1}{\alpha} J \left[\lambda I + \frac{J}{\alpha}\right]^{-1} \right|\right| \leq M. \tag{46}$$

Proof. We use a reference system where J is diagonal. (We recall that the eigenvalues of J are positive.) In this reference system we have $||\frac{1}{\alpha} J \left[\lambda I + \frac{J}{\alpha}\right]^{-1}|| < 1$, as it is simply seen because the matrix $\frac{1}{\alpha} J \left[\lambda I + \frac{J}{\alpha}\right]^{-1}$ is diagonal, its entries being of the form $j/(j + \lambda\alpha)$, where j is *positive* since it is an eigenvalue of J. □

It follows from Lemma 21 and the fact that $\hat{z}^\alpha(\cdot + r)$ converges to $\hat{z}^0(\cdot + r)$ in $L^2(-\infty, +\infty)$ (see (41) and Lemmas 16 and 17) that the first integral (44) converges to zero for $\alpha \to 0$.

In order to prove that the second integral (45) converges to zero, we need a further lemma.

LEMMA 22. *We have*

$$\lim_{\alpha \to 0} \frac{1}{\alpha} J \left[\lambda I + \frac{J}{\alpha}\right]^{-1} = I\,,$$

uniformly for λ in compact sets of $\Re e\, \lambda > 0$.

Proof. We note that

$$\frac{1}{\alpha} J \left[\lambda I + \frac{J}{\alpha}\right]^{-1} - I = [-\lambda\alpha J^{-1}] \frac{1}{\alpha} J \left[\lambda I + \frac{J}{\alpha}\right]^{-1}.$$

The result now follows easily from Lemma 21. □

Now we split (45) into the sum of the two integrals

$$\int_{\mathbb{R} - [h,h]} \left|\left| \left[\frac{1}{\alpha} J \left[(i\omega + r)I + \frac{J}{\alpha}\right]^{-1} - I\right] \right|\right| \cdot ||\hat{z}^0(i\omega + r)||^2 \, d\omega, \tag{47}$$

$$\int_{-h}^{+h} \left|\left| \left[\frac{1}{\alpha} J \left[(i\omega + r)I + \frac{J}{\alpha}\right]^{-1} - I\right] \right|\right| \cdot ||\hat{z}^0(i\omega + r)||^2 \, d\omega. \tag{48}$$

Using standard techniques and Lemma 22, we obtain that both integrals converge to 0. This proves (42) and *completes the proof* of Theorem 18.

It is clear from Theorem 18 that, in general, Bv_α does not converge to Bu since $A_{12}e^0_-$ in general will not be zero. We have the following result.

THEOREM 23. *The following conditions are equivalent:*
(a) $\lim_{\alpha\to 0} Bv_\alpha = Bu$ *and* $\lim_{\alpha\to 0} w_\alpha = x$ *for every input* u.
(b) $B_- = 0$.
(c) $\ker CB = \{0\}$.

Proof. (a)⇒(b). It immediately follows from the definition of v_α and the structure of the matrix $\hat{J}$.

(b)⇒(a). If (b) holds, then from (36) we see that $e_-^0 = 0$. The result then follows from Theorem 18.

(b)⇒(c). Since $\ker B = \{0\}$ and $B_- = 0$, we have that $\ker B_+ = \{0\}$. Multiplying $\hat{J}$ on the left by B^* and on the right by B, we get

$$[B^*B]B^*C^*[CB] = B_+^* J B_+.$$

Now $B_+^* J B_+$ is invertible since $\ker B_+ = \{0\}$ and $J = J^* > 0$. Hence CB is injective.

(c)⇒(b). We recall that $\hat{J} = BB^*C^*C = \operatorname{diag}[J, 0]$ and that $B = \operatorname{col}[B_+, B_-]$. We must prove $B_- = 0$. This is true because

$$B[B^*C^*CB] = [BB^*C^*C]B = \begin{bmatrix} JB_+ \\ 0 \end{bmatrix}.$$

As CB is injective, we have that B^*C^*CB is invertible so that

$$B = \begin{bmatrix} JB_+ \\ 0 \end{bmatrix} \cdot [B^*C^*CB]^{-1} = \begin{bmatrix} ? \\ 0 \end{bmatrix},$$

which was our claim. □

Remark 24. Property (c) in Theorem 23 is a condition of relative degree one on the system. This condition has a relevant role in robustness analysis; see [7]. When it is not satisfied, we need to carry on a more refined analysis, which is the content of next section.

Theorem 23 gives a necessary and sufficient condition under which the proposed recursive identification method really identifies both the input u and the evolution x of system (3). Notice, however, that we described the identification process as a double limiting process: first a limit with respect to n and h and next a limit with respect to α. In fact, there exists a small value of α such that $||Bv_\alpha - Bu||_2$ is less than $\epsilon/2$ for a prescribed ϵ; with such an α, we have that $||Bv_{n,\alpha} - Bv_\alpha||_2 < \epsilon/2$ for large n and small h. In this way we get that $Bv_{n,\alpha}$ approximate Bu within a tolerance ϵ (we denote by $v_{n,\alpha}$ the function defined in (30)). We present explicit convergence estimates in the case when u is differentiable. These estimates, which we prove in the special case $\ker CB = \{0\}$, will be crucial for the extension to the general case presented in the next section. Our standing assumption is $\ker B = \{0\}$ so that we can equivalently give an estimate for $g = Bu$.

We put ourselves in the case described by Theorem 23, and we assume that the input u, hence also $g = Bu$, is differentiable. In this case we can give estimates which explicitly show the dependence on n, h, and α. As we said, we denote by $v_{n,\alpha}$ the input which was simply denoted by v_n in section 3 and by $w_{n,\alpha}$ the corresponding solution to (27). The expression of $v_{n,\alpha}$ is given by (30).

We use

$$||Bv_{n,\alpha}(t) - Bu(t)|| \le ||Bv_{n,\alpha}(t) - Bv_\alpha(t)|| + ||Bv_\alpha(t) - Bu(t)||, \tag{49}$$

and we give asymptotic estimates for both terms on the right-hand side.

We use the pointwise estimate (40) (which concerns $||Bv_\alpha(t) - Bu(t)||$ since e_-^0 is now zero), and we get

$$||Bv_\alpha(t) - Bu(t)|| \leq M \cdot \left(e^{-\frac{\eta t}{\alpha}} ||u(0)|| + \sqrt{\alpha} \right). \tag{50}$$

The value of the constant depends on the values of u and $\dot{u}$.

Now we observe (see (30) and (31)) that

$$\begin{aligned} & v_{n,\alpha}(t) - v_\alpha(t) \\ & = -\left\{ [\alpha I + \Lambda_n^* \Lambda_n]^{-1} \Lambda_n^* \Gamma_n C e^{A\delta} w_{n,\alpha}(\tau_k) - \frac{1}{\alpha} B^* C^* C w^\alpha(t) \right\} \qquad (51) \\ & + \left\{ [\alpha I + \Lambda_n^* \Lambda_n]^{-1} \Lambda_n^* \Gamma_n \xi_k - \frac{1}{\alpha} B^* C^* C x(t) \right\}. \qquad (52) \end{aligned}$$

□

We want to give an estimate of this difference for α fixed and large n.

The sequence of piecewise constant functions Ξ_n, $\Xi_n(t) = \xi_k$ for $t \in [\tau_k, \tau_{k+1})$, converges uniformly to the Lipschitz function Cx (the output of the system) on $[0, T]$ and

$$||\Xi_n(t) - Cx(t)|| < h + \frac{M}{n}. \tag{53}$$

Hence, from the explicit estimates (16), we have

$$||(\Lambda_n^* \Gamma_n \Xi_n)(t) - B^* C^* C x(t)|| \leq M \cdot \left[h + \frac{1}{n} \right].$$

We use now the explicit estimate (23), and we get

$$\left|\left| \left([\alpha I + \Lambda_n^* \Lambda_n]^{-1} \Lambda_n^* \Gamma_n \Xi_n \right)(t) - \frac{1}{\alpha} B^* C^* C x(t) \right|\right| \leq M \cdot \left(\frac{1}{\alpha^2 \sqrt{n}} + \frac{1}{\alpha n} + \frac{h}{\alpha} \right).$$

We estimate (51) with the following sum:

$$\begin{aligned} & \left|\left| [\alpha I + \Lambda_n^* \Lambda_n]^{-1} \Lambda_n^* \Gamma_n C [e^{A\delta} - I] w_{n,\alpha}(\tau_k) \right|\right| \\ & + \left|\left| [\alpha I + \Lambda_n^* \Lambda_n]^{-1} \Lambda_n^* \Gamma_n C w_{n,\alpha}(\tau_k) - \frac{1}{\alpha} B^* C^* C w^\alpha(\tau_k) \right|\right| \\ & + \frac{1}{\alpha} \left|\left| B^* C^* C [w^\alpha(\tau_k) - w^\alpha(t)] \right|\right|. \end{aligned}$$

The inequalities in Lemma 8, Theorem 4, and (10) show that the first addendum is less than

$$\frac{M}{n^{3/2}} \left(\frac{1}{\alpha} + \frac{1}{\sqrt{n} \alpha^2} \right) e^{\omega/\alpha}.$$

The second addendum is estimated from the inequalities (23), (29). It is less than

$$M \left(\frac{1}{\alpha^2 \sqrt{n}} + \frac{1}{\alpha n} + \frac{1}{\alpha^2 \sqrt{n}} e^{\omega/\alpha} \right).$$

In order to estimate the third term, we use (32), Lemma 14, and the boundedness of u. We get that the third term is less than

$$\frac{M}{n\alpha}.$$

According to Theorem 4 and Lemma 11, the factor $e^{\omega/\alpha}$ in the previous inequalities must be replaced by $1/\alpha$ if $C = I$.

We collect the estimates above and we get the following theorem.

THEOREM 25. *Let the unknown input u be of class $W^{1,2}$ and let the condition $\ker CB = \{0\}$ hold. Then there exist numbers M, η, and ω, which do not depend on α, n, h, such that, for every $t \in [0,T]$,*

$$\begin{aligned} ||Bv_{n,\alpha}(t) - Bu(t)|| \le M \Big\{ &\left(e^{-\frac{\eta t}{\alpha}} ||u(0)|| + \sqrt{\alpha}\right) \\ &+ \left(\tfrac{1}{\alpha n} + \tfrac{1}{\alpha^2\sqrt{n}} + \tfrac{h}{\alpha}\right) + \left(\tfrac{1}{\alpha^2\sqrt{n}} + \tfrac{1}{\alpha n^{3/2}} + \tfrac{1}{\alpha^2 n^2}\right) \cdot e^{\omega/\alpha} \Big\}. \end{aligned} \tag{54}$$

In the special case when $C = I$ and $u(0) = 0$, the previous inequality is replaced by

$$||Bv_{n,\alpha}(t) - Bu(t)|| \le M \left(\sqrt{\alpha} + \frac{1}{\alpha n} + \frac{1}{\alpha^2\sqrt{n}} + \frac{h}{\alpha} + \frac{1}{\alpha^3\sqrt{n}}\right). \tag{55}$$

Remark 26.

(a) We observe that the explicit estimate shows that, naturally, if we want a good estimate, the penalization coefficient α should converge to zero not too quickly. Moreover, the right-hand sides of (54) and (55) show explicitly the dependence on n and h separately.

(b) The book [10] studies an even more general input identification problem for nonlinear systems under a condition which, in the linear case, is exactly $\ker CB = \{0\}$ (see, in particular, section 18.4). However, it is assumed there that the unknown input takes values in a known compact set explicitly used in the solution. Moreover, in contrast with (55), only L^2 estimates are given.

5. Input identification in the general case. In this section we extend the identification method described in previous sections to the general case when $\ker CB \neq \{0\}$ (maintaining, of course, the assumptions $\mathcal{R}_* = \{0\}$ and $\ker B = \{0\}$).

To clarify our approach, we first consider an example.

Example 27. Assume that the system matrices have the following special form:

$$A = \left[\begin{array}{ccc|ccc} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ a_{3,1} & a_{3,2} & a_{3,3} & a_{3,4} & a_{3,5} & a_{3,6} \\ \hline 0 & 0 & 0 & 0 & 1 & 0 \\ a_{5,1} & a_{5,2} & a_{5,3} & a_{5,4} & a_{5,5} & a_{5,6} \\ \hline a_{6,1} & a_{6,2} & a_{6,3} & a_{6,4} & a_{6,5} & a_{6,6} \end{array}\right], \quad B = \left[\begin{array}{ccc} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ \hline 0 & 0 & 0 \\ 0 & 1 & 0 \\ \hline 0 & 0 & 1 \end{array}\right], \tag{56}$$

$$C = \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right].$$

A system which has the previous representation but with $a_{3,i} = 0$, $a_{5,i} = 0$, $a_{6,i} = 0$ for each i is just a tandem connection of integrators. It is said to be in "prime" canonical form; see [17].

We represent $x = \mathrm{col}[x_i]$, $u = \mathrm{col}[u_i]$, and $y = \mathrm{col}[y_i]$ consistently with the structure of the matrices.

We now repeatedly apply the identification method described in previous sections. Consider first the subsystem

$$\left\{ \begin{array}{rcl} \dot{x}_1 &=& x_2, \\ \dot{x}_4 &=& x_5, \end{array} \right. \qquad y = \left[\begin{array}{c} x_1 \\ x_4 \end{array} \right].$$

Our recursive identification method gives an estimate of "inputs" x_2, x_5 *because the equivalent conditions of Theorem* 23 *are satisfied.* We observe that the "inputs" x_2, x_5 are differentiable and zero for $t = 0$. Hence Theorem 25 gives a pointwise estimate for the error, which holds *uniformly* on $[0, T]$.

We use the estimate given in Theorem 25 as a new "tolerance" h, and we consider the subsystem

$$\dot{x}_2 = x_3, \qquad y_1 = x_2.$$

We proceed analogously, and we approximate x_3, which will be considered as a new "input" in the last step. The approximation to x_3 is uniform on $[0, T]$.

In the last step we consider the subsystem

$$\left\{ \begin{array}{rcllll} \dot{x}_3 &=& a_{3,3}x_3 + a_{3,5}x_5 + a_{3,6}x_6 & +\rho_1, & & \\ \dot{x}_5 &=& a_{5,3}x_3 + a_{5,5}x_5 + a_{5,6}x_6 & +\rho_2, & y_2 = & \left[\begin{array}{c} x_3 \\ x_5 \\ x_6 \end{array} \right] \\ \dot{x}_6 &=& a_{6,3}x_3 + a_{6,5}x_5 + a_{6,6}x_6 & +\rho_3, & & \end{array} \right.$$

and

$$\rho = \left[\begin{array}{c} \rho_1 \\ \rho_2 \\ \rho_3 \end{array} \right] = \left[\begin{array}{c} a_{3,1}x_1 + a_{3,2}x_2 + a_{3,4}x_4 \\ a_{5,1}x_1 + a_{5,2}x_2 + a_{5,4}x_4 \\ a_{6,1}x_1 + a_{6,2}x_2 + a_{6,4}x_4 \end{array} \right] + \left[\begin{array}{c} u_1 \\ u_2 \\ u_3 \end{array} \right].$$

At this point, the "output" y_2 is known, within a certain tolerance, and the assumptions of Theorem 23 are satisfied. Hence we can apply our procedure for the identification of the "input" ρ, i.e., of the real input u, since we already estimated the first addendum of ρ within a certain tolerance.

We observe that we do not need to perform the procedure just described sequentially in three steps hence in a nonrecursive way. In fact, let n be fixed. Once the first observation time τ_1 is elapsed, we have an estimate for the new "outputs" $x_2(t)$, $x_5(t)$ for $t \in [0, \tau_1]$. Hence we can start the identification procedure for x_3, and, at time τ_3, we can start the identification procedure for u. We also observe that each step of the previous procedure studies a system for which $C = I$. Hence the exponential term $e^{\omega/\alpha}$ does not affect the rate of convergence.

The previous example may look quite special. In fact, it is completely general. In order to see this, we introduce a special "quasi canonical" form due to Morse; see [17]. This form resembles Kalman decomposition, but it goes far deeper. It is an extension of Brunovski canonical form and, in particular, it shows precisely the coupling between inputs and outputs.

According to [17], in the special case $\mathcal{R}_* = \{0\}$, there exist a state feedback F, an input injection L, and suitable reference frames in the state, input, and output spaces such that $A + BF + LC$, B, and C take the following forms:

$$A+BF+LC = \left[\begin{array}{ccc} A_0 & 0 & 0 \\ 0 & A_1 & 0 \\ 0 & 0 & A_2 \end{array} \right], \qquad B = \left[\begin{array}{c} 0 \\ 0 \\ B_2 \end{array} \right], \qquad C = \left[\begin{array}{ccc} 0 & C_1 & 0 \\ 0 & 0 & C_2 \end{array} \right]. \tag{57}$$

Moreover, the triple A_2, B_2, C_2 is in "prime" canonical form. This means

$$A_2 = \operatorname{diag}[N_i]\,, \qquad B_2 = \operatorname{diag}[b_i]\,, \qquad c_2 = \operatorname{diag}[c_i]\,,$$

$$N_i = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}, \qquad b_i = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \qquad c_i = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \end{bmatrix}.$$

If $\mathcal{R}_* \neq \{0\}$, then a further block would appear in the matrices $A + BF + LC$, and a further column would appear in the matrix B. This column would correspond to inputs that cannot be reconstructed.

Morse form can be used as follows in the deconvolution problem. We represent system (3) as

$$\begin{aligned} \dot{x} &= \{(A + BF)x + Ly\} + Bu - (BFx + Ly) \\ &= (A + BF + LC)x + Bu - BFx - Ly\,, \qquad y = Cx, \end{aligned}$$

and we put ourselves in that reference system in which the matrices have the form described above.

The matrices F and L (which can be explicitly computed) have the block forms

$$F = \begin{bmatrix} F_0 & F_1 & F_2 \end{bmatrix}, \qquad L = \begin{bmatrix} L_0' & L_0'' \\ L_1' & L_1'' \\ L_2' & L_2'' \end{bmatrix}.$$

Hence the system takes the following form:

$$\text{(58)} \qquad \begin{cases} \dot{x}_0 = A_0x_0 - \{L_0'C_1x_1 + L_0''C_2x_2\} \\ \qquad\;\; = A_0x_0 - \{L_0'y_1 + L_0''y_2\}, \\ \qquad\qquad\qquad\qquad\qquad\qquad y_1 = C_1x_1, \\ \dot{x}_1 = A_1x_1 - \{L_1'C_1x_1 + L_1''C_2x_2\} \\ \qquad\;\; = A_1x_1 - \{L_1'y_1 + L_1''y_2\} \end{cases}$$

and

$$\text{(59)} \qquad \begin{aligned} \dot{x}_2 &= A_2x_2 - \{L_2'C_1x_1 + L_2''C_2x_2\} - B_2Fx + B_2u \\ &= (A_2x_2 - B_2F_2)x_2 - \{L_2'y_1 + L_2''y_2\} \\ &+ B_2\{u - F_0x_0 - F_1x_1\}\,, \qquad\qquad y_2 = C_2x_2. \end{aligned}$$

The functions y_1 and y_2 are measured (at the sample times τ_k and with known error). They enter as regular perturbations in the previous differential equations so that we can compute x_0 and x_1 on $[0, T]$ (with a tolerance which tends to zero for $h \to 0$ and $\delta \to 0$).

Now we look more closely to the subsystem (59). This system has a block form similar to the one in Example 27. In fact, due to the prime-type structure of the matrices A_2 and B_2, the matrix $A_2 + B_2F_2$ has the same form as the matrix A in (56) (in general with more blocks) and B_2, C_2 have forms similar to those in (56), with more blocks. Hence we can apply the identification algorithm to systems of the form

$$\dot{x}_i = x_{i+1} - \mu_i,$$

where μ_i, due to the addendum $-\{L_1' y_1 + L_2'' y_2\}$, is approximately known. So the next component x_{i+1} can be approximately identified.

We do this for the system of the first components of the vectors which correspond to blocks of dimensions larger then 1. We consider the next component in a second step, as in Example 27, until we remain with the vectors whose entries correspond to the last rows of the blocks of $A_2 + B_2 F$. We get a system of differential equations, each one of the form

$$\dot{x}^k_{r_k} = a^{kk}_{k,r_k} x^k_{r_k} + \{\mu_r + \nu_r + u_r\} + \sum_{(i,j)\neq(k,r_k)} a^k_{ij} x^k_{ij},$$

where μ_r is due to the output injection (hence it is directly measured); ν_r denotes the contribution of x_0, x_1, which can be computed; the terms of the last sum are computed in the preceding steps. Hence all these terms are (approximately) known. Hence a last application of the identification algorithm gives an estimate of the unknown input $u = \mathrm{col}[u_r]$.

We observe that in order to iterate the first step on the second block of variables, we do not need to wait until time T is elapsed: once that time $t_0 > \tau$ is elapsed, we can apply the process to the second block of variables on the interval $[0, t_0]$. Hence also the identification method in the general case is recursive.

In conclusion, a recursive identification process can be applied to the most general case in which input identification is possible.

We do not insist on giving explicit convergence estimates for each step of the iteration. These estimates can be obtained by a repeated application of Theorem 25. We observe the importance of the pointwise estimate (55). In fact, even if we could have perfect observation, at the second step the new "output" is the estimate of the variables $x_{i,2}$ obtained in the previous step with an error given by (55).

We note that at each step, with the exception of the last one, *the pointwise estimates are uniform on* $[0, T]$ *even if the unknown input* u *is not differentiable. In fact, the fictitious inputs that must be estimated at each step, except the last one, are differentiable and zero for* $t = 0$*; and, moreover, at each step we work with a system for which* $C = I$ *so that the exponential factor* $e^{\omega/\alpha}$ *is replaced by* $1/\alpha$ *in the estimate for the convergence rate.*

Acknowledgments. We thank the referees for the careful reading of the paper. Their observations helped us in putting the paper in the proper perspective of the deconvolution problem.

REFERENCES

[1] M. Aoki and C. Yue, *On certain convergence questions in systems identification*, SIAM J. Control, 8 (1970), pp. 239–256.
[2] V. V. Arestov, *Best approximation of differentiation operators*, Mat. Zametki, 1 (1967), pp. 149–154 (in Russian).
[3] G. Basile and G. Marro, *Controlled and Conditioned Invariants in Linear Systems Theory*, Prentice Hall, Englewood Cliffs, NJ, 1992.
[4] M. Bertero, T. A. Poggio, and V. Torre, *Ill-posed problems in early vision*, Proc. IEEE, 76 (1988), pp. 869–889.
[5] D. Commenges, *The deconvolution problem: Fast algorithm including the preconditioned conjugate-gradient to compute a MAP estimator*, IEEE Trans. Automat. Control, 20 (1984), pp. 229–243.
[6] R. Conti, *Linear Differential Equations and Control*, Academic Press, London, 1976.
[7] V. Dragan and A. Halanay, *Stabilization of Linear Systems*, Birkhäuser Boston, Boston, 1999.

[8] N. N. Krasovskii and A. I. Subbotin, *Game-Theoretical Control Problems*, Springer-Verlag, New York, Berlin, 1988.
[9] A. V. Kryazhimskii and Yu. S. Osipov, *Modelling of a control in a dynamic system*, Engrg. Cybernetics, 21 (1983), pp. 51–60 (in Russian).
[10] A. V. Kryazhimskii and Yu. S. Osipov, *Inverse Problems for Ordinary Differential Equations: Dynamical Solutions*, Gordon and Breach, London, 1995.
[11] F.-M. Lee, I.-K. Fong, and L.-C. Fu, *Stable on line parameter identification algorithm for systems with nonparametric uncertainties and disturbances*, Internat. J. Control, 65 (1996), pp. 329–345.
[12] V. I. Maksimov, *On the stable solution of inverse problems for nonlinear distributed systems* I, Differential Equations, 26 (1990), pp. 1537–1546 (in Russian).
[13] V. I. Maksimov, *On the stable solution of inverse problems for nonlinear distributed systems* II, Differential Equations, 27 (1991), pp. 416–421 (in Russian).
[14] V. Maksimov and L. Pandolfi, *Dynamical reconstruction of inputs for contraction semigroup systems: The boundary input case*, J. Optim. Theory Appl., 103 (1999), pp. 401–420.
[15] V. Maksimov and L. Pandolfi, *The problem of dynamical reconstruction of Dirichlet boundary control in semilinear hyperbolic equations.* Inverse Ill-Posed Probl., 8 (2000), pp. 399–420.
[16] M. Morf, G. S. Sidhu, and T. Kailath, *Some new algorithms for recursive estimation in constant linear discrete-time systems*, IEEE Trans. Automat. Control, 19 (1974), pp. 315–323.
[17] A. S. Morse, *Structural invariants of linear multivariable systems*, SIAM J. Control, 11 (1973), pp. 446–465.
[18] G. De Nicolao, G. Sparacino, and C. Cobelli, *Nonparametric input estimation in physiological systems: Problems, methods and case studies*, Automatica J. IFAC, 33 (1997), pp. 851–870.
[19] D. L. Phillips, *A technique for the numerical solution of certain integral equations of the first kind*, J. ACM, 9 (1962), pp. 84–97.
[20] A. Saberi, A.A. Stoorvogel, and P. Sannuti, *Inverse filtering and deconvolution*, Internat. J. Robust Nonlinear Control, 11 (2001), pp. 131–156.
[21] L. A. Sakhnovich, *Integral Equations with Difference Kernels on Finite Intervals*, Birkhäuser, Basel, 1996.
[22] A. N. Tikhonov, *On the solution of ill-posed problems and the method of regularization*, Soviet Math. Dokl., 4 (1963), pp. 1035–1038.
[23] A. N. Tikhonov and V. Y. Arsenin, *Solution of Ill-Posed Problems*, Winston and Wiley, Washington, D.C., New York, 1977.
[24] G. M. Wing, *A Primer on Integral Equations of the First Kind*, SIAM, Philadelphia, 1991.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 5, pp. 1406–1430

LOSSLESS AND DISSIPATIVE DISTRIBUTED SYSTEMS*

HARISH K. PILLAI† AND JAN C. WILLEMS‡

Abstract. This paper deals with linear shift-invariant distributed systems. By this we mean systems described by constant coefficient linear partial differential equations. We define dissipativity with respect to a quadratic differential form, i.e., a quadratic functional in the system variables and their partial derivatives. The main result states the equivalence of dissipativity and the existence of a storage function or a dissipation rate. The proof of this result involves the construction of the dissipation rate. We show that this problem can be reduced to Hilbert's 17th problem on the representation of a nonnegative rational function as a sum of squares of rational functions.

Key words. quadratic differential forms, linear multidimensional systems, behavioral theory, polynomial matrices, lossless systems, positivity, dissipativeness, storage functions

AMS subject classifications. 93A30, 93C20, 13P05, 35G05, 37L99, 35L65

PII. S0363012900368028

1. Introduction.

One of the very useful concepts in systems theory is the notion of a dissipative system. It lies at the root of most of the stability results and on the synthesis of robust controllers. The theory of dissipative systems has been developed until now as a system theoretic concept for dynamical systems, i.e., for systems in which the independent variable is time. However, many if not most models of physical systems are distributed, involving both time and space variables. The purpose of this paper is to develop the theory of dissipative systems for systems described by partial differential equations.

The central problem in the theory of dissipative systems is the construction of an internal function called the storage function. Instances of functions that play the role of storage functions are Lyapunov functions in stability analysis, the internal energy, and entropy in thermodynamics, etc. The construction of storage functions for dynamical systems is reasonably well understood [23, Part 1] for general nonlinear systems and in much detail for linear systems with quadratic supply rates [23, Part 2] [25]. As we shall see, analogous results may be obtained, as far as existence is concerned, for distributed systems described by linear constant coefficient partial differential equations and with quadratic differential forms (QDFs) as supply rates. However, there are important differences in the resulting theory, the most important one being the fact that for distributed systems the storage functions need to be (in general) a function of unobservable ("hidden") latent variables.

Several recent papers [2, 12, 13] dealing with conservative and dissipative systems have been brought to our notice. In these papers, the authors consider an input/state/output framework for the multidimensional systems involved. The results in these papers are clearly related to the results presented in this paper. While the

*Received by the editors February 18, 2000; accepted for publication (in revised form) June 5, 2001; published electronically January 9, 2002.
http://www.siam.org/journals/sicon/40-5/36802.html

†ISIS Research Group, Department of Electronics and Computing Science, University of Southampton, SO17 1QP, Southampton, UK. Current address: Department of Electrical Engineering, Indian Institute of Technology, Bombay, Powai, Mumbai 400076, India (hp@ee.iitb.ac.in).

‡Institute for Mathematics and Computing Science, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands. Current address: Department of Electrical Engineering, ESAT/SISTA, Univeristy of Leuven, B-3001 Leuven-Haverlee, Belgium (Jan.Willems@esat.kuleuven.ac.be).

results in [2, 12, 13] are more general (in the sense that they consider more general signal spaces—Hilbert spaces), they are far less structured (in the sense that they tackle only problems that admit a type of state formulation—the Roesser model). On the other hand, the results in this paper are more structured in the sense that it deals with systems that arise as solutions of constant coefficient partial differential equations (without assuming "states," etc.), though the signal spaces used are not as general. The mathematics involved in the two approaches are also substantially different.

An interesting feature of the results presented in this paper is the mathematics that underlies the construction of the storage function (for linear systems with quadratic supply rates). In the context of lumped dynamical systems the construction of a storage function involves, as we shall see, the factorization of a real polynomial matrix Φ in one indeterminate into the product $\Phi(\xi) = F^T(-\xi)F(\xi)$ with F also a real polynomial matrix. This factorization is readily seen to be possible if and only if $\Phi(\xi) = \Phi^T(-\xi)$ and $\Phi(i\omega) \geq 0$ for all $\omega \in \mathbb{R}$. However, in the case of distributed systems, Φ is a polynomial matrix in n indeterminates. In this case, the factorization $\Phi(\xi) = F^T(-\xi)F(\xi)$ is not always possible with F as a real polynomial matrix but it is possible with F as a matrix of rational functions. This factorization, it turns out, is known as *Hilbert's* 17*th problem*, and it is most stimulating indeed to see this problem emerge in a basic system theoretic question!

First, a few words about notation. We use the standard notation $\mathbb{R}^n$, $\mathbb{R}^{n_1 \times n_2}$, etc., for finite-dimensional vectors and matrices. When the dimension is not specified (but, of course, finite), we write $\mathbb{R}^\bullet$, $\mathbb{R}^{n \times \bullet}$, $\mathbb{R}^{\bullet \times \bullet}$, etc. In order to enhance readability, we typically use the notation $\mathbb{R}^{\mathtt{w}}$ when functions taking their values in that vector space are denoted by w. Real polynomials in the indeterminates $\xi = (\xi_1, \xi_2, \ldots, \xi_n)$ are denoted by $\mathbb{R}[\xi]$ and real rational functions by $\mathbb{R}(\xi)$, with obvious modifications for the matrix case. The space of infinitely differentiable functions with domain $\mathbb{R}^n$ and codomain $\mathbb{R}^{\mathtt{w}}$ is denoted by $\mathfrak{C}^\infty(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}})$ and its subspace containing elements with compact support by $\mathfrak{D}(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}})$.

The proofs of the results are collected in the appendix.

2. Multidimensional systems. We view a system as a family of trajectories mapping a set of "independent" variables into a set of "dependent" variables. See [20] for an elaboration of this with examples. Thus a *system* Σ is defined as a triple $\Sigma = (\mathbb{T}, \mathbb{W}, \mathfrak{B})$, where $\mathbb{T}$ is the indexing set, the set of independent variables, $\mathbb{W}$ is the signal space, the set of dependent variables, and $\mathfrak{B} \subset \mathbb{W}^{\mathbb{T}}$ is the behavior. In the present paper we consider systems with $\mathbb{T} = \mathbb{R}$ (we call these *lumped dynamical* systems or one-dimensional (1D) systems) and systems with $\mathbb{T} = \mathbb{R}^n$ (we call these *distributed* systems—they are commonly called nD systems). Also, we assume throughout that $\mathbb{W}$ is a finite-dimensional real vector space, $\mathbb{W} = \mathbb{R}^{\mathtt{w}}$.

A system $\Sigma = (\mathbb{R}^n, \mathbb{R}^{\mathtt{w}}, \mathfrak{B})$ is said to be *linear* if $\mathfrak{B}$ is a linear subspace of $(\mathbb{R}^{\mathtt{w}})^{\mathbb{R}^n}$ and *shift-invariant* if $\mathfrak{B} = \sigma^{x'}\mathfrak{B}$ for all $x' = (x'_1, \ldots, x'_n) \in \mathbb{R}^n$, where $\sigma^{x'} : (\mathbb{R}^{\mathtt{w}})^{\mathbb{R}^n} \to (\mathbb{R}^{\mathtt{w}})^{\mathbb{R}^n}$ denotes the x'-shift defined for $x' = (x'_1, \ldots, x'_n)$ by $\sigma^{x'} f(x_1, \ldots, x_n) = f(x_1 + x'_1, \ldots, x_n + x'_n)$. We call Σ a linear shift-invariant differential system if $\mathfrak{B}$ is the solution set of a system of linear constant coefficient partial differential equations. More precisely, if there exists a real polynomial matrix $R \in \mathbb{R}^{\bullet \times \mathtt{w}}[\xi]$ in n indeterminates, $\xi = (\xi_1, \ldots, \xi_n)$, such that $\mathfrak{B}$ consists of the $\mathfrak{C}^\infty(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}})$-solutions of

$$R\left(\frac{d}{d\mathbf{x}}\right) w = 0, \tag{1}$$

where $\frac{d}{d\mathbf{x}} = (\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \ldots, \frac{\partial}{\partial x_n})$. The assumption that we consider only $\mathfrak{C}^\infty$-solutions is made for the ease of exposition, and the results remain valid for other solution concepts—for example, for distributions. We denote the family of linear shift-invariant differential systems $\Sigma = (\mathbb{R}^n, \mathbb{R}^{\mathtt{w}}, \mathfrak{B})$ as $\mathfrak{L}_n^{\mathtt{w}}$. We also denote $(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}}, \mathfrak{B}) \in \mathfrak{L}_n^{\mathtt{w}}$ as $\mathfrak{B} \in \mathfrak{L}_n^{\mathtt{w}}$ since the indexing set and the signal space are then obvious from the context.

A system $\mathfrak{B} \in \mathfrak{L}_n^{\mathtt{w}}$ is uniquely specified by its annihilators, defined by

$$\mathfrak{N}_{\mathfrak{B}} = \left\{ p \in \mathbb{R}^{1\times \mathtt{w}}[\xi] \mid p\left(\frac{d}{d\mathbf{x}}\right)\mathfrak{B} = 0 \right\}.$$

It is easy to see that $\mathfrak{N}_{\mathfrak{B}}$ is a submodule of $\mathbb{R}^{1\times \mathtt{w}}[\xi]$ viewed as a module over $\mathbb{R}[\xi]$. In fact, there is a one-to-one relation between $\mathfrak{L}_n^{\mathtt{w}}$ and the submodules of $\mathbb{R}^{1\times \mathtt{w}}[\xi]$. Thus, whereas $R \in \mathbb{R}^{\bullet\times \mathtt{w}}[\xi]$ uniquely specifies a behavior $\mathfrak{B} \in \mathfrak{L}_n^{\mathtt{w}}$ through (1) with $\mathfrak{N}_{\mathfrak{B}}$ the module generated by the rows of R, any other polynomial matrix whose rows generate the same submodule define the same behavior.

The family of systems $\mathfrak{L}_n^{\mathtt{w}}$ enjoys many convenient properties, and this has been studied in detail in [19]. An important feature is the elimination theorem, which is the consequence of the following. Let $F \in \mathbb{R}^{\mathtt{w}_1\times \mathtt{w}_2}[\xi]$. Then $\mathfrak{B}_2 \in \mathfrak{L}_n^{\mathtt{w}_2}$ implies $F(\frac{d}{d\mathbf{x}})\mathfrak{B}_2 \in \mathfrak{L}_n^{\mathtt{w}_1}$ and $\mathfrak{B}_1 \in \mathfrak{L}_n^{\mathtt{w}_1}$ implies $(F(\frac{d}{d\mathbf{x}}))^{-1}\mathfrak{B}_1 \in \mathfrak{L}_n^{\mathtt{w}_2}$. This, in particular, implies that if $\mathfrak{B}_1, \mathfrak{B}_2 \in \mathfrak{L}_n^{\mathtt{w}}$, then $\mathfrak{B}_1 \cap \mathfrak{B}_2 \in \mathfrak{L}_n^{\mathtt{w}}$ and $\mathfrak{B}_1 + \mathfrak{B}_2 \in \mathfrak{L}_n^{\mathtt{w}}$. It also implies the elimination theorem that states that, for any $\mathfrak{B} \in \mathfrak{L}_n^{\mathtt{w}_1+\mathtt{w}_2}$, the set

$$\{w_1 \in \mathfrak{C}^\infty(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}_1}) \mid \exists w_2 \in \mathfrak{C}^\infty(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}_2}) : (w_1, w_2) \in \mathfrak{B}\}$$

is itself an element of $\mathfrak{L}_n^{\mathtt{w}_1}$. The elimination theorem and its variations follow from the important *fundamental principle* that states that the system of partial differential equations

$$A\left(\frac{d}{d\mathbf{x}}\right) f = g,$$

with $A \in \mathbb{R}^{\mathtt{w}_1\times \mathtt{w}_2}[\xi]$ and $g \in \mathfrak{C}^\infty(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}_1})$ given, is solvable for $f \in \mathfrak{C}^\infty(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}_2})$ if and only if whenever $p \in \mathbb{R}^{1\times \mathtt{w}_1}[\xi]$ satisfies $pA = 0$, then there must hold that $p(\frac{d}{d\mathbf{x}})g = 0$.

Whereas we have *defined* the behavior of a system in $\mathfrak{L}_n^{\mathtt{w}}$ as the set of solutions of a system of partial differential equations in the system variables, often, in practical applications, the specification of the behavior involves other, auxiliary variables, which we call *latent variables*. Specifically, consider the system of partial differential equations

$$R\left(\frac{d}{dx}\right) w = M\left(\frac{d}{dx}\right) \ell \tag{2}$$

with $w \in \mathfrak{C}^\infty(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}})$ and $\ell \in \mathfrak{C}^\infty(\mathbb{R}^n, \mathbb{R}^\ell)$ and with $R \in \mathbb{R}^{\bullet\times \mathtt{w}}[\xi]$ and $M \in \mathbb{R}^{\bullet\times \ell}[\xi]$ polynomial matrices with the same number of rows. The set

$$\mathfrak{B}_f = \{(w, \ell) \in \mathfrak{C}^\infty(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}+\ell}) \mid (2) \text{ holds}\} \tag{3}$$

obviously belongs to $\mathfrak{L}_n^{\mathtt{w}+\ell}$. It immediately follows from the elimination theorem that the set

$$\{w \in \mathfrak{C}^\infty(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}}) \mid \exists \ell \in \mathfrak{C}^\infty(\mathbb{R}^n, \mathbb{R}^\ell) : (w, \ell) \in \mathfrak{B}_f\} \tag{4}$$

belongs to $\mathfrak{L}_n^{\mathbb{w}}$. We call (2) a latent variable representation, with manifest variables w and latent variables ℓ, of the system with full behavior (3) and manifest behavior (4). Correspondingly, we call (1) a kernel representation of the system with the behavior $\ker(R(\frac{d}{d\mathbf{x}}))$. We shall soon meet another sort of representation, the image representations, in the context of controllability.

3. Controllability and observability. Two very influential classical properties of dynamical systems are those of controllability and observability. In [24] these properties have been lifted to lumped dynamical systems in a behavioral setting, while in [19] generalizations to distributed systems have been introduced. We discuss these concepts here exclusively in the context of systems described by linear constant coefficient partial differential equations.

DEFINITION 1. *A system $\mathfrak{B} \in \mathfrak{L}_n^{\mathbb{w}}$ is said to be* controllable *if for all $w_1, w_2 \in \mathfrak{B}$ and for all sets $U_1, U_2 \subset \mathbb{R}^n$ with disjoint closure, there exists a $w \in \mathfrak{B}$ such that $w\,|_{U_1} = w_1\,|_{U_1}$ and $w\,|_{U_2} = w_2\,|_{U_2}$.*

Thus controllable partial differential equations are those in which the solutions can be "patched up" from solutions on subsets: in a sense there is no "action of a distance." There are a number of characterizations of controllability. In terms of its submodule of annihilators, $\mathfrak{N}_{\mathfrak{B}}$, $\mathfrak{B} \in \mathfrak{L}_n^{\mathbb{w}}$, is controllable if and only if the module $\mathbb{R}^{1\times \mathbb{w}}[\xi]/\mathfrak{N}_{\mathfrak{B}}$ is torsion-free [19].

More useful for our purposes is the equivalence of controllability with the existence of an image representation. Consider the following special latent variable representation:

$$w = M\left(\frac{d}{d\mathbf{x}}\right)\ell \tag{5}$$

with $M \in \mathbb{R}^{\mathbb{w}\times \ell}[\xi]$. Obviously, by the elimination theorem, its manifest behavior $\mathfrak{B} \in \mathfrak{L}_n^{\mathbb{w}}$. Such special latent variable representations often appear in physics, where the latent variables involved in such a representation are called *potentials*. Obviously, $\mathfrak{B} = \mathrm{im}(M(\frac{d}{d\mathbf{x}}))$ with $M(\frac{d}{d\mathbf{x}})$ viewed as a map from $\mathfrak{C}^\infty(\mathbb{R}^n, \mathbb{R}^\ell)$ to $\mathfrak{C}^\infty(\mathbb{R}^n, \mathbb{R}^{\mathbb{w}})$. For this reason, we call (5) an *image representation* of its manifest behavior. Whereas every $\mathfrak{B} \in \mathfrak{L}_n^{\mathbb{w}}$ allows (by definition) a kernel representation and hence trivially a latent variable representation, not every $\mathfrak{B} \in \mathfrak{L}_n^{\mathbb{w}}$ allows an image representation. In fact, see the following theorem.

THEOREM 2. *$\mathfrak{B} \in \mathfrak{L}_n^{\mathbb{w}}$ admits an image representation if and only if it is controllable.*

We denote the set of controllable systems in $\mathfrak{L}_n^{\mathbb{w}}$ by $\mathfrak{L}_{n,\mathrm{cont}}^{\mathbb{w}}$.

Observability is the property of systems that have two kinds of variables; the first set of variables are the "observed" set of variables, and the second set of variables are the ones that are "to-be-deduced" from the observed variables. Every variable that can be deduced uniquely from the manifest variables of a given behavior will be called an *observable*. So observability is not an intrinsic property of a given behavior. One has to be given a partition of the variables in the behavior into two classes before one can say whether one class of variables in the behavior can actually be deduced from the other class of variables (which were observed).

DEFINITION 3. *Let $w = (w_1, w_2)$ be a partition of the variables in $\Sigma = (\mathbb{R}^n, \mathbb{R}^{\mathbb{w}_1+\mathbb{w}_2}, \mathfrak{B})$. Then w_2 is said to be* observable *from w_1 in $\mathfrak{B}$ if given any two trajectories $(w_1', w_2'), (w_1'', w_2'') \in \mathfrak{B}$ such that $w_1' = w_1''$; then $w_2' = w_2''$.*

A natural situation to use observability is when one looks at the latent variable representation of a behavior. Then one may ask whether the latent variables are

observable from the manifest variables. If this is the case, then we call the latent variable representation *observable*.

As we have already mentioned, every controllable behavior has an image representation. In the case of 1D systems, it can be shown that every controllable behavior has an observable image representation. This is not true for nD systems.

4. QDFs. In [25, 26] a theory was developed for linear (1D) differential systems and quadratic functionals associated with these systems. It was shown that for systems described by *one-variable* polynomial matrices, the appropriate tool to express quadratic functionals are *two-variable* polynomial matrices. In the same vein, in this paper we will use polynomial matrices in $2n$ variables to express quadratic functionals for functions of n variables.

For convenience, let ζ denote $(\zeta_1, \dots, \zeta_n)$, and let η denote $(\eta_1, \dots, \eta_n)$. Let $\mathbb{R}^{\mathtt{w}_1 \times \mathtt{w}_2}[\zeta, \eta]$ denote the set of real polynomial matrices in the $2n$ indeterminates ζ and η. We will consider quadratic forms of the type $\Phi \in \mathbb{R}^{\mathtt{w}_1 \times \mathtt{w}_2}[\zeta, \eta]$. Explicitly,

$$\Phi(\zeta, \eta) = \sum_{\mathbf{k},\mathbf{l}} \Phi_{\mathbf{k},\mathbf{l}} \zeta^{\mathbf{k}} \eta^{\mathbf{l}}.$$

The sum above ranges over all nonnegative multi-indices $\mathbf{k} = (k_1, k_2, \dots, k_n), \mathbf{l} = (l_1, l_2, \dots, l_n) \in \mathbb{N}^n$, and the sum is assumed to be finite. Moreover, $\Phi_{\mathbf{k},\mathbf{l}} \in \mathbb{R}^{\mathtt{w}_1 \times \mathtt{w}_2}$. The polynomial matrix Φ induces a *bilinear differential form* (BLDF), that is, the map

$$L_\Phi : \mathfrak{C}^\infty(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}_1}) \times \mathfrak{C}^\infty(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}_2}) \to \mathfrak{C}^\infty(\mathbb{R}^n, \mathbb{R})$$

defined by

$$L_\Phi(v, w)(\mathbf{x}) := \sum_{\mathbf{k},\mathbf{l}} \left(\frac{d^{\mathbf{k}} v}{d\mathbf{x}^{\mathbf{k}}}(\mathbf{x}) \right)^T \Phi_{\mathbf{k},\mathbf{l}} \left(\frac{d^{\mathbf{l}} w}{d\mathbf{x}^{\mathbf{l}}}(\mathbf{x}) \right),$$

where $\frac{d^{\mathbf{k}}}{d\mathbf{x}^{\mathbf{k}}} = \frac{\partial^{k_1}}{\partial x_1^{k_1}} \frac{\partial^{k_2}}{\partial x_2^{k_2}} \cdots \frac{\partial^{k_n}}{\partial x_n^{k_n}}$ and analogously for $\frac{d^{\mathbf{l}}}{d\mathbf{x}^{\mathbf{l}}}$. Note that ζ corresponds to differentiation of terms to the left, and η refers to differentiation of the terms to the right.

If $\mathtt{w}_1 = \mathtt{w}_2 = \mathtt{w}$, then Φ induces the QDF

$$Q_\Phi : \mathfrak{C}^\infty(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}}) \to \mathfrak{C}^\infty(\mathbb{R}^n, \mathbb{R})$$

defined by

$$Q_\Phi(w) := L_\Phi(w, w).$$

Define the * operator

$$^* : \mathbb{R}^{\mathtt{w} \times \mathtt{w}}[\zeta, \eta] \to \mathbb{R}^{\mathtt{w} \times \mathtt{w}}[\zeta, \eta]$$

by

$$\Phi^*(\zeta, \eta) := \Phi^T(\eta, \zeta).$$

If $\Phi = \Phi^*$, then Φ is called *symmetric*. For the purposes of QDFs induced by polynomial matrices, it suffices to consider the symmetric QDFs since $Q_\Phi = Q_{\Phi^*} = Q_{\frac{1}{2}(\Phi + \Phi^*)}$.

We also consider vectors $\Psi \in (\mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta,\eta])^n$, i.e., $\Psi = (\Psi_1,\dots,\Psi_n)$. Analogous to the QDF induced by Φ, Ψ induces a *vector of quadratic differential forms* (VQDF)

$$Q_\Psi(w) : \mathfrak{C}^\infty(\mathbb{R}^n,\mathbb{R}^{\mathtt{w}}) \to (\mathfrak{C}^\infty(\mathbb{R}^n,\mathbb{R}))^n$$

defined by $Q_\Psi = (Q_{\Psi_1},\dots,Q_{\Psi_n})$.

Finally, we define the "div" (divergence) operator that associates with the VQDF induced by Ψ, the scalar QDF:

$$(\text{div } Q_\Psi)(w) := \frac{\partial}{\partial x_1} Q_{\Psi_1}(w) + \cdots + \frac{\partial}{\partial x_n} Q_{\Psi_n}(w).$$

The theory of QDFs has been developed in much detail in [25, 26] for 1D systems. In the next section, we put forward those aspects which are useful in the construction of storage function for distributive systems.

5. Path independence. Consider the integral

$$\int_\Omega Q_\Phi(w)d\mathbf{x}, \tag{6}$$

where Ω is a closed bounded subset of $\mathbb{R}^n$ with a nonempty interior. This integral is said to be independent of the "path" w (or a *path integral*) if the integral depends only on the value of w and its derivatives on the boundary of Ω, denoted by $\partial\Omega$. More precisely, if for any $w_1, w_2 \in \mathfrak{C}^\infty(\mathbb{R}^n,\mathbb{R}^{\mathtt{w}})$ such that $\frac{d^{\mathbf{k}}w_1}{d\mathbf{x}^{\mathbf{k}}}(\mathbf{x}) = \frac{d^{\mathbf{k}}w_2}{d\mathbf{x}^{\mathbf{k}}}(\mathbf{x})$ for all $\mathbf{x}\in\partial\Omega$ and all $\mathbf{k}\in\mathbb{N}^n$, there holds

$$\int_\Omega Q_\Phi(w_1)d\mathbf{x} \;=\; \int_\Omega Q_\Phi(w_2)d\mathbf{x}.$$

Instead of some $\Omega\subset\mathbb{R}^n$, if we consider the integral (6) over all of $\mathbb{R}^n$, then the integral need not be well defined for all $w\in\mathfrak{C}^\infty(\mathbb{R}^n,\mathbb{R}^{\mathtt{w}})$. We can overcome this by considering it only for w's of compact support. This yields the functional

$$\int Q_\Phi : \mathfrak{D}(\mathbb{R}^n,\mathbb{R}^{\mathtt{w}}) \to \mathbb{R}$$

defined by

$$\int Q_\Phi(w) := \int_{\mathbb{R}^n} Q_\Phi(w)d\mathbf{x},$$

which evaluates the integral over all of $\mathbb{R}^n$.

The following theorem gives several conditions that are equivalent to path independence.

THEOREM 4. *Let $\Phi\in\mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta,\eta]$. Then the following statements are equivalent:*

1. *$\int_\Omega Q_\Phi$ is independent of path for all closed bounded subsets Ω of $\mathbb{R}^n$.*
2. *$\int Q_\Phi = 0$.*
3. *$\Phi(-\xi,\xi) = 0$.*
4. *There exist $\Psi_1,\dots,\Psi_n\in\mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta,\eta]$ such that*

$$\Phi(\zeta,\eta) = (\zeta_1+\eta_1)\Psi_1(\zeta,\eta) + \cdots + (\zeta_n+\eta_n)\Psi_n(\zeta,\eta).$$

5. *There exists a* $\Psi \in (\mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta,\eta])^n$ *such that*

$$\operatorname{div} Q_\Psi = Q_\Phi$$

for all $w \in \mathfrak{C}^\infty(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}})$.

At this point we would like to point out an important difference for the cases $n = 1$ and $n > 1$. Although the above theorem holds for all values of n, more can be said in the case when $n = 1$. In the case when $n = 1$, the last condition of the above theorem can be strengthened to state that there exists a unique Ψ such that $\frac{d}{dt}Q_\Psi = Q_\Phi$ (assuming t is the independent variable). This uniqueness of Ψ does not hold when $n > 1$. This will become clear from the subsequent proposition, which will help us in classifying this nonuniqueness. If Ψ_1 and Ψ_2 induce two VQDFs such that

$$Q_\Phi = \operatorname{div} Q_{\Psi_1} = \operatorname{div} Q_{\Psi_2}, \tag{7}$$

then $\Psi = \Psi_1 - \Psi_2$ defines a VQDF such that $\operatorname{div} Q_\Psi(w) = 0$ for all $w \in \mathfrak{C}^\infty(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}})$. Such a VQDF is said to have *null divergence*. Thus it is obvious that given a $\Phi \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta,\eta]$ which defines a path integral and a VQDF induced by $\Psi \in (\mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta,\eta])^n$ such that $\operatorname{div} Q_\Psi(w) = Q_\Phi(w)$, it is possible to obtain other VQDFs that satisfy this property by adding VQDFs that have null divergence to the already obtained VQDF Ψ. We now characterize those VQDFs that have null divergence.

PROPOSITION 5. *A VQDF induced by* $\Psi = (\Psi_1, \dots, \Psi_n) \in (\mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta,\eta])^n$ *has null divergence if and only if there exists a family of* n^2 *QDFs induced by* $\Delta_{ij} \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta,\eta]$, $i = 1, \dots, n$, $j = 1, \dots, n$, *with* $\Delta_{ij} = -\Delta_{ji}$ *such that*

$$\Psi_i = (\zeta_1 + \eta_1)\Delta_{i1} + (\zeta_2 + \eta_2)\Delta_{i2} + \cdots + (\zeta_n + \eta_n)\Delta_{in}.$$

From the above proposition, it is clear that $\Delta_{ii} = 0$. Thus for 1D systems, the QDF induced by Δ_{11} is the zero QDF, and so there exists no nonzero 1D (V)QDFs that have null divergence. Hence the Ψ obtained in Theorem 4 for 1D systems is unique [26, Theorem 3.1]. In fact, $\Psi(\zeta,\eta) = \frac{\Phi(\zeta,\eta)}{\zeta+\eta}$ in 1D systems. In nD systems with $n > 1$, the Ψ obtained in Theorem 4 is no longer unique since there exist nonzero VQDFs that give rise to null divergences. The above proposition completely classifies the nonuniqueness of these VQDFs. Hence, for every path independent QDF induced by $\Phi \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta,\eta]$, one obtains an equivalence class of VQDFs such that (7) holds. The members of an equivalence class are exactly those that differ by a VQDF that has null divergence.

6. Lossless systems. In this section, we study the notion of path independence generalized to controllable systems $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}_{n,\mathrm{cont}}$. We cast this in the context of conservative systems.

Let $\Phi = \Phi^* \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta,\eta]$ and $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}_{n,\mathrm{cont}}$. Now consider the QDF $Q_\Phi(w)$ for trajectories $w \in \mathfrak{B}$. We consider $Q_\Phi(w)(\mathbf{x})$ (with $\mathbf{x} \in \mathbb{R}^n$) as the rate of supply of some physical quantity (for example, energy) delivered to the system at the point $\mathbf{x}$ (whence positive when the system absorbs supply).

DEFINITION 6. *The system* $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}_{n,\mathrm{cont}}$ *is said to be* lossless *with respect to the supply rate* Q_Φ *induced by* $\Phi = \Phi^* \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta,\eta]$ *if* $\int_{\mathbb{R}^n} Q_\Phi(w)d\mathbf{x} = 0$ *for all* $w \in \mathfrak{B} \cap \mathfrak{D}(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}})$.

The interpretation of this condition is that $\int_{\mathbb{R}^n} Q_\Phi(w)d\mathbf{x}$ denotes the net amount of supply that the system absorbs integrated over "time" and "space." Whence the system is lossless if this integral is zero: any supply absorbed at some time or place is temporarily stored but eventually recovered perhaps at some other time or place.

A related notion is that of path independence along a behavior. Let Ω be a closed and bounded subset of $\mathbb{R}^n$. The integral $\int_\Omega Q_\Phi(w)d\mathbf{x}$ is said to be independent of path for trajectories $w \in \mathfrak{B}$ if whenever $w_1, w_2 \in \mathfrak{B}$ and $\frac{d^{\mathbf{k}}w_1}{d\mathbf{x}^{\mathbf{k}}}(\mathbf{x}) = \frac{d^{\mathbf{k}}w_2}{d\mathbf{x}^{\mathbf{k}}}(\mathbf{x})$ for $\mathbf{x} \in \partial\Omega$ and all $\mathbf{k} \in \mathbb{N}^n$, then

$$\int_\Omega Q_\Phi(w_1)d\mathbf{x} \;=\; \int_\Omega Q_\Phi(w_2)d\mathbf{x}.$$

Define the * operator mapping from $\mathbb{R}^{\mathtt{w}_1\times\mathtt{w}_2}[\xi]$ to $\mathbb{R}^{\mathtt{w}_2\times\mathtt{w}_1}[\xi]$ by $X^\star(\xi) := X^T(-\xi)$. In other words, if we look at $X(\frac{d}{d\mathbf{x}})$ as a partial differential operator, then $X^\star(\frac{d}{d\mathbf{x}})$ is the (formal) adjoint operator.

The following theorem gives a number of equivalent conditions for a system to be lossless.

THEOREM 7. *Let $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}_{n,\mathrm{cont}}$. Let $R \in \mathbb{R}^{\bullet\times\mathtt{w}}[\xi]$ and $M \in \mathbb{R}^{\mathtt{w}\times\bullet}[\xi]$ induce, respectively, a kernel and image representation of $\mathfrak{B}$; i.e., $\mathfrak{B} = \ker(R(\frac{d}{d\mathbf{x}})) = \mathrm{im}(M(\frac{d}{d\mathbf{x}}))$. Let $\Phi = \Phi^* \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta,\eta]$ induce a QDF on $\mathfrak{B}$. Then the following conditions are equivalent:*

1. *$\mathfrak{B}$ is lossless with respect to the QDF Q_Φ;*
2. *The QDF induced by Φ is independent of path on $\mathfrak{B}$, i.e., $\int_\Omega Q_\Phi(w)d\mathbf{x}$ is independent of path for all bounded and closed subsets Ω in $\mathbb{R}^n$ with a nonempty interior;*
3. *the QDF corresponding to Φ' is a path integral, where Φ' is given by $\Phi'(\zeta,\eta) := M^T(\zeta)\Phi(\zeta,\eta)M(\eta)$;*
4. *$\Phi'(-\xi,\xi) = 0$;*
5. *there exists a VQDF Q_Ψ, with $\Psi \in (\mathbb{R}^{m\times m}[\zeta,\eta])^n$, where m is the number of columns of M such that*

$$\mathrm{div}\, Q_\Psi(\ell) = Q_{\Phi'}(\ell) = Q_\Phi(w) \tag{8}$$

for all $\ell \in \mathfrak{C}^\infty(\mathbb{R}^n,\mathbb{R}^m)$ and $w = M(\frac{d}{d\mathbf{x}})\ell$.

We focus our attention for a moment on the equivalence of conditions 1 and 5 of the above theorem. It states that $\mathfrak{B}$ is lossless with respect to Q_Φ, i.e., that

$$\int_{\mathbb{R}^n} Q_\Phi(w)d\mathbf{x} = 0 \tag{9}$$

for all $w \in \mathfrak{B}$ of compact support if and only if $\mathfrak{B}$ admits an image representation $w = M(\frac{d}{d\mathbf{x}})\ell$ and there exists some VQDF Ψ such that

$$\mathrm{div}\; Q_\Psi(\ell) = Q_\Phi(w) \tag{10}$$

for all $w \in \mathfrak{B}$ and ℓ such that $w = M(\frac{d}{d\mathbf{x}})\ell$.

The equivalence of the global version of losslessness (9) with the local version (10) is a recurrent theme in the theory of dissipative systems. The local version states that there is a function $Q_\Psi(\ell)(\mathbf{x})$ that plays the role of the amount of supply stored at $\mathbf{x} \in \mathbb{R}^n$. Thus (10) says that for lossless systems, it is possible to define a *storage function* Q_Ψ such that the *conservation equation*

$$\mathrm{div}\; Q_\Psi(\ell) = Q_\Phi(w) \tag{11}$$

is satisfied for all w, ℓ such that $w = M(\frac{d}{d\mathbf{x}})\ell$. Note here that since $\Phi' = \mathrm{div}\;\Psi$, by the Stokes theorem

$$\int_{\partial\Omega}\sum_{i=1}^n (-1)^{i-1}Q_{\Psi_i}(\ell)dx_1\wedge\cdots\wedge\widehat{dx_i}\wedge\cdots\wedge dx_n = \int_\Omega Q_{\Phi'}(\ell)dx_1\wedge\cdots\wedge dx_n$$

(for any $\Omega \subseteq \mathbb{R}^n$ with a reasonable boundary). We can then think of the above as an integral form of the conservation equation (11).

Two important features, both specific to the case when $n > 1$, are worth emphasizing. First is the fact that the storage $Q_\Psi(\ell)$ depends on the latent variable ℓ from the image representation $w = M(\frac{d}{d\mathbf{x}})\ell$. Since $\mathfrak{B} \in \mathfrak{L}^{\mathbb{W}}_{n,\text{cont}}$ may not have an observable image representation, there may not exist a storage function of the form $Q_\Psi(w)$ that depends on the manifest variables $w \in \mathfrak{B}$. Hence the storage in (11) involves "hidden" (i.e., nonobservable) variables. Second, the nonuniqueness of the VQDF Q_Ψ that solves $\operatorname{div} Q_\Psi(\ell) = Q_\Phi(M(\frac{d}{d\mathbf{x}})\ell) = Q_{\Phi'}(\ell)$. Hence, even when the ℓ's have acquired a "physical significance," there will be many possible storage functions. We shall see in the next section that this nonuniqueness is important already in basic physics.

We would like to mention at this point that in many practical examples the independent variables are time and space variables. So, for example, the indexing set would be $\mathbb{R} \times \mathbb{R}^3$. In this case, we will use the notation t, x, y, z to stand for the independent variables (time coordinate and the three space coordinates, respectively), and the partial derivatives with respect to these variables are denoted by $\frac{\partial}{\partial t}, \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}$, respectively. It is important to interpret the storage function Q_Ψ in this context. In the case mentioned above, we denote $\Psi = (\Psi_t, \Psi_x, \Psi_y, \Psi_z)$ and $Q_\Psi = (u, \mathbf{S})$. Here u is the QDF Q_{Ψ_t}, which is the "*internal storage*" and the VQDF $\mathbf{S} := (Q_{\Psi_x}, Q_{\Psi_y}, Q_{\Psi_z})$ is the "*flux*." This interpretation will be useful in the next section. With the above notation, (8) now becomes

$$\frac{\partial}{\partial t} u(\ell) + \nabla \cdot \mathbf{S}(\ell) = Q_\Phi(w),$$

where ∇ is the spatial divergence operator.

7. Maxwell's equations. The prototypical example of a linear shift-invariant differential system is provided by Maxwell's equations in free space:

$$\begin{aligned} \nabla \cdot \mathbf{E} - \frac{\rho}{\epsilon_0} &= 0, \\ \nabla \times \mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} &= 0, \\ c^2 \nabla \times \mathbf{B} - \frac{\partial \mathbf{E}}{\partial t} - \frac{\mathbf{j}}{\epsilon_0} &= 0, \\ \nabla \cdot \mathbf{B} &= 0. \end{aligned} \tag{12}$$

This describes the relation between the electrical field $\mathbf{E} : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}^3$, the magnetic field $\mathbf{B} : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}^3$, the current density $\mathbf{j} : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}^3$, and the charge density $\rho : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}$. In the above equations, the constants c and ϵ_0 stand for the speed of light in vacuum and the electric constant, respectively. Hence (12) defines a system $\mathfrak{B}_{ME} \in \mathfrak{L}^{10}_4$. It is well known that $\mathfrak{B}_{ME}$ can be described in terms of the vector potential $\mathbf{A} : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}^3$ and the scalar potential $\phi : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}$ by

$$\begin{aligned} \mathbf{E} &= -\frac{\partial \mathbf{A}}{\partial t} - \nabla \phi, \\ \rho &= -\epsilon_0 \nabla \cdot \frac{\partial \mathbf{A}}{\partial t} - \epsilon_0 \nabla^2 \phi, \\ \mathbf{B} &= \nabla \times \mathbf{A}, \\ \mathbf{j} &= \epsilon_0 \frac{\partial^2 \mathbf{A}}{\partial t^2} - \epsilon_0 c^2 \nabla^2 \mathbf{A} + \epsilon_0 c^2 \nabla(\nabla \cdot \mathbf{A}) + \epsilon_0 \nabla \frac{\partial \phi}{\partial t}. \end{aligned} \tag{13}$$

It is important to note that (13) is an image representation of $\mathfrak{B}_{ME}$. Hence, by Theorem 2, Maxwell's equations define a controllable system. It is also important to note that (13) is an unobservable image representation of $\mathfrak{B}_{ME}$. In fact, there *do not exist* observable image representations of $\mathfrak{B}_{ME}$.

Strictly speaking, the vector potential $\mathbf{A}$ and the scalar potential ϕ are "free" latent variables (i.e., they are allowed to take on any values in the relevant space of trajectories). Note that we can change $\mathbf{A}$ and ϕ to $\mathbf{A}' = \mathbf{A} + \nabla\psi$ and $\phi' = \phi - \frac{\partial\psi}{\partial t}$ (where ψ is some other arbitrary scalar function) without changing the resulting $\mathbf{E}$, $\mathbf{B}$, ρ, and $\mathbf{j}$. These are called *gauge transformations*. Additional conditions may be imposed on $\mathbf{A}$ and ϕ without changing the fact that the image in (13) remains $\mathfrak{B}_{ME}$. For example, the Lorentz condition

$$\nabla \cdot \mathbf{A} = -\frac{1}{c^2}\frac{\partial \phi}{\partial t} \tag{14}$$

can be imposed on the potentials to obtain symmetry in the representation (13). In this case, the last two terms of the last equation in (13) disappear, thus displaying a symmetry in the equations. Moreover, these new equations then remain invariant under Lorentz transformations of the independent variables. There are other possibilities. The important point is that the gauge transformations and imposition of such conditions like the Lorentz condition do not change the set of $(\mathbf{E}, \mathbf{B}, \mathbf{j}, \rho)$ obtained as solutions to the Maxwell equations. In other words, (13) and (14) together provide a latent variable representation of $\mathfrak{B}_{ME}$. We will not consider such transformations further in this paper.

We are interested in studying the exchange of electrical energy between the environment and the electromagnetic field in free space. This exchange of energy only involves the electrical variables $(\mathbf{E}, \mathbf{j})$. The laws that are described by these variables define, by the elimination theorem, a system $\mathfrak{B}_{\mathrm{E}} \in \mathfrak{L}_4^6$. Consider, therefore, in Maxwell's equations the magnetic field $\mathbf{B}$ and the charge density ρ as latent variables. Then, by eliminating these latent variables, we obtain

$$\begin{aligned} \frac{\partial}{\partial t}\nabla \cdot \mathbf{E} + \frac{1}{\epsilon_0}\nabla \cdot \mathbf{j} &= 0, \\ \frac{\partial^2 \mathbf{E}}{\partial t^2} + c^2 \nabla \times (\nabla \times \mathbf{E}) + \frac{1}{\epsilon_0}\frac{\partial \mathbf{j}}{\partial t} &= 0. \end{aligned} \tag{15}$$

The above equations give a kernel representation for the behavior $\mathfrak{B}_{\mathrm{E}}$ consisting of all trajectories $(\mathbf{E}, \mathbf{j}) \in \mathfrak{C}^\infty(\mathbb{R}^4, \mathbb{R}^6)$ which are compatible with the solutions of Maxwell's equations. Since $\mathfrak{B}_{\mathrm{ME}}$ is controllable, so is $\mathfrak{B}_{\mathrm{E}}$, and so one can obtain an image representation of it.

$$\begin{aligned} \mathbf{E} &= -\nabla\phi - \frac{\partial \mathbf{A}}{\partial t}, \\ \mathbf{j} &= \epsilon_0 \frac{\partial}{\partial t}\nabla\phi + \epsilon_0 \frac{\partial^2 \mathbf{A}}{\partial t^2} + \epsilon_0 c^2 \nabla \times (\nabla \times \mathbf{A}). \end{aligned} \tag{16}$$

Here $\mathbf{A}$ and ϕ are again the vector and scalar potentials, respectively [10].

Consider the QDF $Q_\Phi(\mathbf{E}, \mathbf{j}) = \mathbf{E} \cdot \mathbf{j}$ for all $w \in \mathfrak{B}_{\mathrm{ME}}$. This quantity defines the rate of work done by the field on each unit volume [10].

It is well known that Maxwell's equations define a lossless system. This also follows from Theorem 7. Indeed, by identifying the matrix Φ corresponding to the QDF $Q_\Phi(\mathbf{E}, \mathbf{j}) = \mathbf{E} \cdot \mathbf{j}$ and the M matrix corresponding to the image representation

(13), we can compute $\Phi'(\zeta,\eta) := M^T(\zeta)\Phi(\zeta,\eta)M(\eta)$. It is easily seen that $\Phi'(-\xi,\xi) = 0$. Losslessness follows from Theorem 7. The QDF induced by Φ' is a path integral on the potentials, which in turn implies that Φ is a path integral on the solutions of Maxwell's equations. By Theorem 7, there exists a VQDF, $\Psi \in (\mathbb{R}^{4\times 4}[\zeta,\eta])^4$, such that $\operatorname{div} Q_\Psi(\phi,\mathbf{A}) = Q_\Phi(\mathbf{E},\mathbf{j}) = \mathbf{E}\cdot\mathbf{j}$. By the terminology defined at the end of last section, we can write the VQDF Q_Ψ as $(-u,-\mathbf{S})$ (the negative signs are purely a matter of convention). Then we have

$$\mathbf{E}\cdot\mathbf{j} = \operatorname{div} Q_\Psi(\phi,\mathbf{A}) = -\frac{\partial u(\phi,\mathbf{A})}{\partial t} - \nabla\cdot\mathbf{S}(\phi,\mathbf{A}).$$

On substituting $\mathbf{B} = \nabla\times\mathbf{A}$ and $\mathbf{E} = -\nabla\phi - \frac{\partial\mathbf{A}}{\partial t}$, we obtain

$$\begin{aligned} u &= \frac{\epsilon_0}{2}\mathbf{E}\cdot\mathbf{E} + \frac{\epsilon_0 c^2}{2}\mathbf{B}\cdot\mathbf{B}, \\ \mathbf{S} &= \epsilon_0 c^2 \mathbf{E}\times\mathbf{B}. \end{aligned} \tag{17}$$

This u defines the *energy density* in the field, and $\mathbf{S}$ represents the *energy flux* of the field. The vector $\mathbf{S}$ is known as the "Poynting vector." Thus (8) gives a "conservation law" for Maxwell's equations. It states that the rate at which the field does work on an infinitesimal volume ($Q_\Phi(\mathbf{E},\mathbf{j}) = \mathbf{E}\cdot\mathbf{j}$) is equal to the rate of decrease in the energy density ($-\frac{\partial u}{\partial t}$) and the energy flux ($-\nabla\cdot\mathbf{S}$) that flows into the infinitesimal volume under consideration. Thus (8) states that the total energy is conserved.

We now interpret these results about Maxwell's equations in terms of the theory developed earlier. There are two points that we would like to emphasize.

1. The problem under consideration may be viewed as finding out if the system given by (15) (the behavior $\mathfrak{B}_{\mathrm{E}}$) is lossless with respect to $Q_\Phi(\mathbf{E},\mathbf{j}) = \mathbf{E}\cdot\mathbf{j}$, and if so, finding a storage function for it. Verification of losslessness involves a straightforward calculation. Also, a storage function $(u,\mathbf{S})$ was derived in terms of $\mathbf{E}$ and $\mathbf{B}$ (17). Note that this storage function depends on $\mathbf{E}$ and $\mathbf{B}$. The latter is a latent variable with respect to the electrical quantities $(\mathbf{E},\mathbf{j})$ involved in (15). In fact, $\mathbf{B}$ is not observable from $(\mathbf{E},\mathbf{j})$ in Maxwell's equations. Hence already in this elementary example the storage functions involve hidden variables.
   From Theorem 7 and the example of Maxwell's equations, it is seen that the VQDF acts on some latent variables. These latent variables are related to the latent variables that appear in an image representation of a given controllable behavior. For example, in Maxwell's equations, $\mathbf{B}$ is related to $\mathbf{A}$. One would like the VQDF to act only on the manifest variables. A sufficient condition for the existence of such a VQDF is that the controllable behavior has an observable image representation. In 1D systems, every controllable system has an observable image representation. As a result, in the 1D case, given a QDF induced by Φ which is independent of path on $\mathfrak{B}$, we can actually find a QDF Ψ such that
   $$\frac{d}{dt}Q_\Psi(w) = Q_\Phi(w)$$
   for all $w\in\mathfrak{B}$. In the nD case, a controllable behavior need not necessarily have an observable image representation. So for the nD case, when the QDF induced by Φ is independent of path on $\mathfrak{B}$, it is sufficient for $\mathfrak{B}$ to have

observable potentials for us to find a VQDF Ψ such that

$$\operatorname{div} Q_\Psi(w) = Q_\Phi(w)$$

for all $w \in \mathfrak{B}$.

2. We would also like to make a comment on the nonuniqueness of the VQDF that appears in the conservation equation (8). With reference to Maxwell's equations, we quote from [10], "All we did was to find a *possible* "u" and a *possible* "$\mathbf{S}$." How do we know that juggling the terms around some more we couldn't find another formula for "u" and "$\mathbf{S}$"? ... It's possible. ... There are, in fact, an infinite number of possibilities for u and $\mathbf{S}$, and so far no one has thought of an experiment to tell which one is right!"
We found that this nonuniqueness of the storage function is an intrinsic feature of storage functions for conservative nD systems with $n > 1$. The result in Proposition 5 characterizes the nonuniqueness of the VQDF that goes with a given QDF induced by Φ which is independent of path on all trajectories in $\mathfrak{C}^\infty(\mathbb{R}^n, \mathbb{R}^\ell)$.

8. Supply, storage, and dissipation. In the previous section, we considered QDFs such that $\int Q_\Phi$ is zero when restricted to some behavior $\mathfrak{B}$: the lossless systems. As we have seen, such QDFs define conservation laws. In this section, we consider QDFs where the integral $\int Q_\Phi$ is nonnegative. In the spirit of [23, 26], we refer to these as dissipative systems. We justify the use of this terminology later.

Our plan is as follows. We first introduce the concepts for general controllable behaviors $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}_{n,\mathrm{cont}}$. Subsequently, we analyze the situation $\mathfrak{B} = \mathfrak{C}^\infty(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}})$. We will see that this leads to the problem of factorization of polynomial matrices in several variables. We subsequently return to general controllable behaviors.

DEFINITION 8. *Let $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}_{n,\mathrm{cont}}$ and $\Phi = \Phi^* \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta, \eta]$. Consider the QDF Q_Φ induced by Φ. We call* $\mathfrak{B}$ dissipative with respect to Q_Φ *(briefly Φ-dissipative) if*

$$\int_{\mathbb{R}^n} Q_\Phi(w) d\mathbf{x} \geq 0$$

for all $w \in \mathfrak{B}$ with compact support.

The intuitive interpretation is that $Q_\Phi(w)$ is the rate of supply (Q_Φ is called the *supply rate*) absorbed by the system. Dissipativity hence means that the net supply that is absorbed by the system is nonnegative for any trajectory $w \in \mathfrak{B}$ that is of compact support.

Two related notions are those of storage functions and dissipation rate. As we have already seen in the context of lossless systems, the storage function is in general a function of unobservable latent variables, more specifically of the latent variables that appear in an image representation (thus depending on "potentials"). We incorporate this in the definitions.

DEFINITION 9. *Let $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}_{n,\mathrm{cont}}$, $\Phi = \Phi^* \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta, \eta]$, and $w = M(\frac{d}{d\mathbf{x}})\ell$ be an image representation of $\mathfrak{B}$ with $M \in \mathbb{R}^{\mathtt{w}\times\ell}[\xi]$. Let $\Psi = (\Psi_1, \Psi_2, \ldots, \Psi_n)$ with $\Psi_k = \Psi_k^* \in \mathbb{R}^{\ell\times\ell}[\zeta, \eta]$ for $k = 1, 2, \ldots, n$. The VQDF Q_Ψ is said to be a* storage function *for $\mathfrak{B}$ with respect to Q_Φ if*

$$\operatorname{div} Q_\Psi(\ell) \leq Q_\Phi(w) \tag{18}$$

for all $\ell \in \mathfrak{D}(\mathbb{R}^n, \mathbb{R}^\ell)$ and $w = M(\frac{d}{d\mathbf{x}})\ell$.

$\Delta = \Delta^* \in \mathbb{R}^{\ell\times\ell}[\zeta,\eta]$ *is said to be a* dissipation rate *for* $\mathfrak{B}$ *with respect to* Q_Φ *if*

$$Q_\Delta \geq 0 \text{ and } \int_{\mathbb{R}^n} Q_\Delta(\ell)d\mathbf{x} = \int_{\mathbb{R}^n} Q_\Phi(w)d\mathbf{x}$$

for all $\ell \in \mathfrak{D}(\mathbb{R}^n,\mathbb{R}^\ell)$ *and* $w = M(\frac{d}{d\mathbf{x}})\ell$.

We define $Q_\Delta \geq 0$ if $Q_\Delta(w(\mathbf{x})) \geq 0$ for all $w \in \mathfrak{D}(\mathbb{R}^n,\mathbb{R}^w)$ evaluated at every $\mathbf{x} \in \mathbb{R}^n$. This defines a pointwise positivity condition. Thus $\int_\Omega Q_\Delta(w)d\mathbf{x} \geq 0$ for every $\Omega \subset \mathbb{R}^n$ if $Q_\Delta \geq 0$.

It is easy to see that there is a relation between a storage function for $\mathfrak{B}$ with respect to Q_Φ and a dissipation rate for $\mathfrak{B}$ with respect to Q_Φ, given by

$$Q_\Delta(\ell) = Q_\Phi\left(M\left(\frac{d}{d\mathbf{x}}\right)\ell\right) - \text{div } Q_\Psi(\ell). \tag{19}$$

The definitions of the storage function and the dissipation rate, combined with (19), yield intuitive interpretations. The dissipation rate can be thought of as the rate of supply that is dissipated in the system and the storage function as the rate of supply stored in the system. Intuitively, we could think of the QDF Q_Φ as measuring the power going into the system. In many practical examples, the power is indeed a QDF of some system variables. (For example, $-\mathbf{E}\cdot\mathbf{j}$ is the rate of work done on the system in the case of Maxwell's equations, or, as mentioned earlier, $\mathbf{E}\cdot\mathbf{j}$ is the rate of work done by the field.) Φ-dissipativity would imply that the net power flowing into a system is nonnegative, which in turn implies that the system dissipates energy. Of course, locally the flow of energy could be positive or negative, leading to variations in energy density and fluxes. The energy density and fluxes could be thought of as a storage function for the energy. (Again see the section on Maxwell's equations.) If the system is dissipative, then the rate of change of energy density and fluxes cannot exceed the power delivered into the system. This is captured by the inequality (18) in Definition 9. The excess is precisely what is lost (or dissipated). This interaction between supply, storage, and dissipation is formalized by (19).

When the independent variables are time and space, we can write (19) as

$$\frac{\partial u(\ell)}{\partial t} = Q_\Phi\left(M\left(\frac{d}{d\mathbf{x}}\right)\ell\right) - \nabla\cdot\mathbf{S}(\ell) - Q_\Delta(\ell), \tag{20}$$

where, as before, we use $Q_\Psi = (u,\mathbf{S})$, with u the stored energy and $\mathbf{S}$ the flux. Moreover, $w = M(\frac{d}{d\mathbf{x}})\ell$. Thus (20) states that the change in the stored energy $(\frac{\partial u(\ell)}{\partial t})$ in an infinitesimal volume is exactly equal to the difference between the energy supplied $(Q_\Phi(w))$ into the infinitesimal volume and the energy lost by the infinitesimal volume by means of energy flux flowing out of the volume $(\nabla\cdot\mathbf{S}(\ell))$ and the energy dissipated $(Q_\Delta(\ell))$ within the volume.

The problem we address is the equivalence of (i) dissipativeness of $\mathfrak{B}$ with respect to Q_Φ, (ii) the existence of a storage function, and (iii) the existence of a dissipation rate. Note that this problem also involves the construction of an appropriate image representation. We first consider the case where $\mathfrak{B} = \mathfrak{C}^\infty(\mathbb{R}^n,\mathbb{R}^{\mathbb{w}})$. In this case, the definition of the dissipation rate requires that for all $\ell \in \mathfrak{D}(\mathbb{R}^n,\mathbb{R}^\ell)$

$$\int_{\mathbb{R}^n} Q_\Phi(w)d\mathbf{x} = \int_{\mathbb{R}^n} Q_\Delta(\ell)d\mathbf{x} \tag{21}$$

with $w = M(\frac{d}{d\mathbf{x}})\ell$; $M(\frac{d}{d\mathbf{x}})$ a surjective partial differential operator and $Q_\Delta(\ell) \geq 0$ for all $\ell \in \mathfrak{D}(\mathbb{R}^n,\mathbb{R}^\ell)$. This latter condition is seen to be equivalent to the existence of

a polynomial matrix $D \in \mathbb{R}^{\bullet\times\ell}[\xi]$ such that $\Delta(\zeta,\eta) = D^T(\zeta)D(\eta)$. One direction of the previous claim is trivial. For the other direction, we think of the operator $\Delta(\zeta,\eta)$ as acting on the space of ℓ and its derivatives (the jet space). The operator $\Delta(\zeta,\eta)$ then becomes a symmetric matrix with real entries that acts on this jet space. The condition $Q_\Delta \geq 0$ is a pointwise condition, and so one obtains the matrix $D(\xi)$ in the obvious way. Using Theorem 7, it follows that (21) is equivalent to the factorization equation

$$M^T(-\xi)\Phi(-\xi,\xi)M(\xi) = D^T(-\xi)D(\xi).$$

This equation with $\Phi = \Phi^* \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta,\eta]$ given and $M \in \mathbb{R}^{\mathtt{w}\times\bullet}[\xi]$ and $D \in \mathbb{R}^{\bullet\times\bullet}[\xi]$ unknown is discussed in the next section.

9. Factorization of polynomial matrices. In this section, we discuss the following problem. Let $\Gamma \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\xi]$ be a polynomial matrix in n commuting variables, $\xi = (\xi_1, \xi_2, \ldots, \xi_n)$. Can it be factored as

$$\Gamma(\xi) = F^T(-\xi)F(\xi). \tag{22}$$

We are interested in both the case when $F \in \mathbb{R}^{\bullet\times\mathtt{w}}[\xi]$ is itself a polynomial matrix and the case when $F \in \mathbb{R}^{\bullet\times\mathtt{w}}(\xi)$ is a matrix of rational functions.

Note that $\Gamma^\star = \Gamma$ and $\Gamma(i\omega) \geq 0$ for all $\omega \in \mathbb{R}^n$ are obviously necessary conditions for the existence of a factor $F \in \mathbb{R}^{\bullet\times\mathtt{w}}[\xi]$. The problem is whether these conditions are also sufficient. At this point, it is convenient to discuss the cases when $n = 1$ and $n > 1$ separately.

9.1. The case $n = 1$. In the case when $n = 1$, it is well known that (22) admits a solution $F \in \mathbb{R}^{\bullet\times\mathtt{w}}[\xi]$ if and only if $\Gamma^\star = \Gamma$ and $\Gamma(i\omega) \geq 0$ for all $\omega \in \mathbb{R}$. In fact, there even exist square factors $F \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\xi]$ that are, moreover, Hurwitz (i.e., with the roots of $\det(F)$ in the closed left half of the complex plane) and square factors that are anti-Hurwitz (i.e., with the roots of $\det(F)$ in the closed right half of the complex plane). These factors are called *spectral factors*. Several algorithms exist for obtaining such factorizations [6, 8, 15, 21].

9.2. The case $n > 1$. We start with the scalar case, i.e., when $\Gamma \in \mathbb{R}[\xi]$. So we need to find $F \in \mathbb{R}^{\bullet\times 1}[\xi]$ or $F \in \mathbb{R}^{\bullet\times 1}(\xi)$ such that $\Gamma(\xi) = F^T(-\xi)F(\xi)$. Substituting $i\omega$ for ξ, the above problem reduces to finding F such that

$$\Gamma(i\omega) = F^\star(i\omega)F(i\omega). \tag{23}$$

If $F(i\omega)$ is decomposed into real and imaginary parts as $F(i\omega) = A(\omega) + iB(\omega)$, then (23) becomes $\Gamma(i\omega) = A^2(\omega) + B^2(\omega)$. Thus the problem reduces to the case of finding a sum of "two" squares which add up to a given positive (or nonnegative) polynomial. This problem has a very venerable history. It is Hilbert's 17th problem that he posed at the International Congress of Mathematicians in 1900. It deals with the representation of positive definite functions as sums of squares [18]. This investigation of positive definite functions began in the year 1888 with the following "negative" result of Hilbert: If $f(\xi) \in \mathbb{R}[\xi]$ is a positive definite polynomial in n variables, then f need not be a sum of squares of polynomials in $\mathbb{R}[\xi]$, except in the case when $n = 1$. Several examples of such positive definite polynomials which cannot be expressed as sum of squares of polynomials are available in the literature; for example, the polynomial

$$\xi_1^2\xi_2^2(\xi_1^2 + \xi_2^2 - 1) + 1$$

is not factorizable as a sum of squares of polynomials [4].

Thus the factorization we were looking for in nD systems (in the form stated above) is not solvable for polynomial factors, not even in the scalar case. However, several results on Hilbert's 17th problem allow us to solve the factorization problem (22) with F as a rational function. Indeed,

- if $\Gamma(\xi) \in \mathbb{R}(\xi)$ and $\Gamma \geq 0$ (i.e., $\Gamma(\xi) \geq 0$ for all $\xi \in \mathbb{R}^n$), then there exists some natural number r such that $\Gamma(\xi)$ is the sum of r squares of rational functions in $\mathbb{R}(\xi)$ (shown by Artin [1]);
- there is a sharp upper bound on the number r; it is $r = 2^n$, shown by Pfister [17, 18].

This leads to the following result, which plays a central role in the rest of this paper.

THEOREM 10. *Assume that* $\Gamma \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\xi]$ *satisfies* $\Gamma^\star = \Gamma$ *and* $\Gamma(i\omega) \geq 0$ *for all* $\omega \in \mathbb{R}^n$. *Then there exists an* $F \in \mathbb{R}^{\bullet\times\mathtt{w}}(\xi)$ *such that* $\Gamma(\xi) = F^T(-\xi)F(\xi)$.

Note that even when Γ is a polynomial matrix, the entries of the matrix F are rational functions in n-indeterminates with real coefficients, whereas for the 1D case one can obtain an F with polynomial entries. Several results related to this factorization problem for the two-dimensional (2D) case (with some additional conditions like holomorphicity on certain complex half planes or unit polydiscs) exist in the literature [3, 11, 14, 16]. A different factorization problem involving symmetric nD matrices is shown in [5].

10. Main results. We now return to the problem of existence of a storage function and a dissipation rate for a Φ-dissipative system $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}_{n,\mathrm{cont}}$. For the sake of clarity, we first consider the case of $\mathfrak{B} = \mathfrak{C}^\infty(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}})$ and subsequently the case of a general $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}_{n,\mathrm{cont}}$. We start with a proposition that gives a condition on Φ for $\int Q_\Phi$ to be nonnegative.

PROPOSITION 11. *Let* $\Phi = \Phi^* \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta, \eta]$. *Then* $(\int Q_\Phi \geq 0)$ *if and only if* $(\Phi(-i\omega, i\omega) \geq 0$ for all $\omega \in \mathbb{R}^n)$.

Proposition 11 and the factorizability implied by Theorem 10 readily lead to the following theorem.

THEOREM 12. *Let* $\Phi = \Phi^* \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta, \eta]$. *Then the following conditions are equivalent:*

1. $\int_{\mathbb{R}^n} Q_\Phi(w) d\mathbf{x} \geq 0$ *for all* $w \in \mathfrak{D}(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}})$.
2. *There exists a polynomial matrix* $M \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\xi]$ *such that* $M(\frac{d}{d\mathbf{x}})$ *is surjective and* $\Psi = (\Psi_1, \Psi_2, \ldots, \Psi_n)$ *with* $\Psi_k = \Psi_k^* \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta, \eta]$ *for* $k = 1, 2, \ldots, n$ *such that the VQDF* Q_Ψ *is a storage function, i.e.,*

$$\operatorname{div}\ Q_\Psi(\ell) \leq Q_\Phi(w)$$

   for all $\ell \in \mathfrak{D}(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}})$ *and* $w = M(\frac{d}{d\mathbf{x}})\ell$.
3. *There exists a polynomial matrix* $M \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\xi]$ *such that* $M(\frac{d}{d\mathbf{x}})$ *is surjective and a* $\Delta = \Delta^* \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta, \eta]$ *such that* Q_Δ *is a dissipation rate, i.e.,*

$$Q_\Delta \geq 0 \text{ and } \int_{\mathbb{R}^n} Q_\Delta(\ell) d\mathbf{x} = \int_{\mathbb{R}^n} Q_\Phi(w) d\mathbf{x}$$

   for all $\ell \in \mathfrak{D}(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}})$ *and* $w = M(\frac{d}{d\mathbf{x}})\ell$.
4. *There exists a polynomial matrix* $M \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\xi]$ *such that* $M(\frac{d}{d\mathbf{x}})$ *is surjective, a* $\Psi = (\Psi_1, \Psi_2, \ldots, \Psi_n)$ *with* $\Psi_k = \Psi_k^* \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta, \eta]$ *for* $k = 1, 2, \ldots, n$, *and a*

$\Delta = \Delta^* \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta,\eta]$ *such that*

$$Q_\Delta \geq 0$$
$$\textit{and}$$
$$\operatorname{div} Q_\Psi(\ell) = Q_\Phi(w) - Q_\Delta(\ell) \tag{24}$$

for all $\ell \in \mathfrak{C}^\infty(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}})$ *and* $w = M(\frac{d}{d\mathbf{x}})\ell$. *Note that this states that the VQDF* Q_Ψ *is a storage function and that* Q_Δ *is a dissipation rate.*

It follows from the proof of Theorem 12 that several Δ's may satisfy the same dissipation condition; i.e., several Δ's may be present, all of them satisfying $Q_\Delta \geq 0$ and $\int_{\mathbb{R}^n} Q_\Delta(\ell)d\mathbf{x} = \int_{\mathbb{R}^n} Q_\Phi(w)d\mathbf{x}$ for all $\ell \in \mathfrak{D}(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}})$ and satisfying the equation

$$w = M\left(\frac{d}{d\mathbf{x}}\right)\ell$$

for some $M \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\xi]$. The first part of this nonuniqueness comes from the factorization of the matrix $\Phi(-\xi,\xi)$. Choosing a particular factorization of the matrix $\Phi(-\xi,\xi)$ still leaves us with a choice for Δ depending on the M we choose. Also note that unlike the 1D case, there is no one-to-one correspondence between storage and dissipation functions. Given a supply QDF and an associated dissipation QDF, one can find several VQDFs that satisfy (24). One should also note the unavoidable emergence of latent variables in the dissipation equation (24) for the nD case. In the 1D case, the dissipation equation can be written in terms of manifest variables alone, whereas in the nD case, this is only possible if the latent variables that appear in the dissipation equation are observable.

The case of an arbitrary $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}_{n,\mathrm{cont}}$ is easily reduced to the free case by considering an image representation for $\mathfrak{B}$. This leads to the following theorem, which is the main result of the paper.

THEOREM 13. *Let* $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}_{n,\mathrm{cont}}$ *and* $\Phi = \Phi^* \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta,\eta]$. *The following conditions are equivalent:*

1. $\mathfrak{B}$ *is* Φ*-dissipative; i.e.,* $\int_{\mathbb{R}^n} Q_\Phi(w)d\mathbf{x} \geq 0$ *for all* $w \in \mathfrak{B} \cap \mathfrak{D}(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}})$.
2. *There exists an integer* $\mathtt{l} \in \mathbb{N}$, *a polynomial matrix* $M \in \mathbb{R}^{\mathtt{w}\times\mathtt{l}}[\xi]$ *such that* $M(\frac{d}{d\mathbf{x}})$ *is an image representation of* $\mathfrak{B}$, *a* $\Psi = (\Psi_1, \Psi_2, \ldots, \Psi_n)$ *with* $\Psi_k = \Psi_k^* \in \mathbb{R}^{\mathtt{l}\times\mathtt{l}}[\zeta,\eta]$ *for* $k = 1, 2, \ldots, n$, *and a* $\Delta = \Delta^* \in \mathbb{R}^{\mathtt{l}\times\mathtt{l}}[\zeta,\eta]$ *such that*

$$Q_\Delta \geq 0$$
$$\textit{and}$$
$$\operatorname{div} Q_\Psi(\ell) = Q_\Phi(w) - Q_\Delta(\ell)$$

with $w = M(\frac{d}{d\mathbf{x}})\ell$.

11. Conclusions. In this paper, we dealt with distributed systems described by constant coefficient linear partial differential equations. We started by defining controllability for such systems in terms of patching up of feasible trajectories. We then explained that it is exactly the controllable systems which allow an image representation, i.e., a representation in terms of what in physics is called a potential function. Subsequently, we turned to lossless and dissipative systems.

For lossless systems, we proved the equivalence with the existence of a conservation law involving the storage function. Important features of the storage function are (i) the fact that it depends on latent variables that are in general hidden (i.e.,

nonobservable) and (ii) its nonuniqueness. We have illustrated these features by demonstrating that they are already present in Maxwell's equations.

For dissipative systems, we proved the equivalence with the existence of a storage function and a dissipation rate. The problem of constructing a dissipation rate led to the question of factorizability of certain polynomial matrices in n variables. We reduced this problem to Hilbert's 17th problem, the representation of a nonnegative rational function in n variables as a sum of squares of rational functions.

12. Appendix. We collect the proofs in this appendix.

Proof of Theorem 2. Please refer to [19, Theorem 3]. ∎

Proof of Theorem 4. First we show the equivalence of the first three statements. Then we will show the equivalence of the last two statements. Finally, we link up these two sets of conditions.

(1) ⇔ (3) If two trajectories w_1 and w_2 agree along with all their derivatives on the boundary $\partial\Omega$ of some arbitrary closed bounded subset $\Omega \subset \mathbb{R}^n$ (with nonempty interior), then $w = w_1 - w_2$ can be thought of as a trajectory with its support strictly in the interior of Ω. Thus $w \in \mathfrak{D}(\mathbb{R}^n, \mathbb{R}^{\mathbb{w}})$. Since

$$Q_\Phi(w_1) = Q_\Phi(w_2) + Q_\Phi(w) + 2L_\Phi(w_2, w),$$

we conclude that $\int_\Omega Q_\Phi$ is independent of path for all Ω if and only if

$$\int_\Omega Q_\Phi(w)d\mathbf{x} + 2\int_\Omega L_\Phi(w, v)d\mathbf{x} = 0 \tag{25}$$

for any $w \in \mathfrak{D}(\mathbb{R}^n, \mathbb{R}^{\mathbb{w}})$ with support in Ω, any $v \in \mathfrak{C}^\infty(\mathbb{R}^n, \mathbb{R}^{\mathbb{w}})$, and any closed bounded subset Ω in $\mathbb{R}^n$ with nonempty interior.

In particular, if we choose v in (25) to be zero, we obtain $\int_\Omega Q_\Phi(w)d\mathbf{x} = 0$, which in turn implies $\int Q_\Phi = 0$. Thus (25) yields $\int_\Omega L_\Phi(w, v)d\mathbf{x} = 0$. But

$$L_\Phi(w, v)(\mathbf{x}) = \sum_{\mathbf{k},\mathbf{l}} \left(\frac{d^{\mathbf{k}} w}{d\mathbf{x}^{\mathbf{k}}}(\mathbf{x})\right)^T \Phi_{\mathbf{k},\mathbf{l}} \left(\frac{d^{\mathbf{l}} v}{d\mathbf{x}^{\mathbf{l}}}(\mathbf{x})\right).$$

Now on integrating each of these terms by parts, we obtain

$$\int_\Omega L_\Phi(w, v)d\mathbf{x} = \int_\Omega w \cdot \left[\sum_{\mathbf{k},\mathbf{l}} \Phi_{\mathbf{k},\mathbf{l}} \left((-1)^{\mathbf{k}} \frac{(d)^{\mathbf{k}}}{d\mathbf{x}^{\mathbf{k}}} \left(\frac{d^{\mathbf{l}} v}{d\mathbf{x}^{\mathbf{l}}}\right)\right)\right] = 0$$

for every $w \in \mathfrak{D}(\mathbb{R}^n, \mathbb{R}^{\mathbb{w}})$ with support in the interior of $\Omega \subset \mathbb{R}^n$ and for every such Ω. This can hold if and only if $\sum_{\mathbf{k},\mathbf{l}} \Phi_{\mathbf{k},\mathbf{l}}((-1)^{\mathbf{k}} \frac{(d)^{\mathbf{k}}}{d\mathbf{x}^{\mathbf{k}}}(\frac{d^{\mathbf{l}} v}{d\mathbf{x}^{\mathbf{l}}})) = 0$ or, equivalently, $\Phi(-\xi, \xi) = 0$. A simple reversal of arguments shows that (3) ⇒ (1).

(2) ⇔ (3) If $\Phi(-\xi, \xi) = 0$, then using integration by parts it is clear that $\int Q_\Phi = 0$. Conversely, if $\int Q_\Phi = 0$, then, using integration by parts, we obtain the condition $\sum_{\mathbf{k},\mathbf{l}} \Phi_{\mathbf{k},\mathbf{l}}((-1)^{\mathbf{k}} \frac{(d)^{\mathbf{k}}}{d\mathbf{x}^{\mathbf{k}}}(\frac{d^{\mathbf{l}} w}{d\mathbf{x}^{\mathbf{l}}})) = 0$ for every $w \in \mathfrak{D}(\mathbb{R}^n, \mathbb{R}^{\mathbb{w}})$. This is only possible if $\Phi(-\xi, \xi) = 0$.

(4) ⇔ (5) As mentioned earlier,

$$(\text{div } Q_\Psi)(w) = \frac{\partial}{\partial x_1} Q_{\Psi_1}(w) + \cdots + \frac{\partial}{\partial x_n} Q_{\Psi_n}(w).$$

It is easy to see that $\frac{\partial}{\partial x_1} Q_{\Psi_1}(w) = L_{\Psi_1}(\frac{\partial w}{\partial x_1}, w) + L_{\Psi_1}(w, \frac{\partial w}{\partial x_1}) = Q_{\Psi_1'}(w)$, where $\Psi_1' = (\zeta_1 + \eta_1)\Psi_1$. Now the equivalence of (4) and (5) is obvious.

(4) $\Rightarrow$ (3) This is obvious from substitution.

(3) $\Rightarrow$ (4) Given a $\Phi \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta, \eta]$, which contains polynomials in $2n$ variables, we can rewrite this matrix as a polynomial matrix in the $2n$ new variables given by $\gamma_i = \zeta_i + \eta_i$ and $\gamma_{i+n} = \zeta_i - \eta_i$ for $i = 1, \ldots, n$. Now setting $-\zeta = \eta = \xi$ translates to setting $\gamma_i = 0$ for $i = 1, \ldots, n$. Thus if $\Phi(-\xi, \xi) = 0$, then each term in the matrix can be written as polynomials in γ_i, $i = 1, \ldots, n$, without the free term. By the very definition of these γ_i's and the symmetry of the matrices, we have the result. ∎

Proof of Proposition 5. ($\Leftarrow$): Let

$$\Psi_i = (\zeta_1 + \eta_1)\Delta_{i1} + \cdots + (\zeta_n + \eta_n)\Delta_{in}$$

with $\Delta_{ij} = -\Delta_{ji}$ as given in the statement of the proposition. Let Φ be such that $Q_\Phi = \operatorname{div} Q_\Psi$. Then

$$\begin{aligned}
\Phi &= (\zeta_1 + \eta_1)\Psi_1 + \cdots + (\zeta_n + \eta_n)\Psi_n \\
&= \sum_{i=1}^{n} (\zeta_i + \eta_i)\Psi_i \\
&= \sum_{i=1}^{n} (\zeta_i + \eta_i) \sum_{j=1}^{n} (\zeta_j + \eta_j)\Delta_{ij} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} (\zeta_i\zeta_j + \zeta_i\eta_j + \eta_i\zeta_j + \eta_i\eta_j)\Delta_{ij} \\
&= 0 \quad \text{since } \Delta_{ij} = -\Delta_{ji}.
\end{aligned}$$

($\Rightarrow$): Let $\Psi \in (\mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta, \eta])^n$ define a VQDF with null divergence, that is, div $Q_\Psi = 0$. Clearly, it is enough to just consider any (k, l)th entry of the corresponding QDFs since the same arguments would apply to all other entries. For simplicity, let us denote by p_i the (k, l)th entry of the QDF Ψ_i.

As in the proof of Theorem 4, we will again employ a change of variables. Let $\gamma_i = \zeta_i + \eta_i$ and $\gamma_{i+n} = \zeta_i - \eta_i$ for $i = 1, \ldots, n$. Since Ψ defines a null divergence,

$$\sum_{i=1}^{n} \gamma_i p_i = 0. \tag{26}$$

Setting all γ_i's except γ_j to zero in (26), we conclude that p_j can be written as

$$p_j = \sum_{i=1, i\neq j}^{n} f_{ji}\gamma_i$$

for some arbitrary polynomials f_{ji} in the $2n$ variables γ_i. Now reverting back to ζ_i's and η_i's, this precisely means that $\Psi_j = (\zeta_1 + \eta_1)\Delta_{j1} + \cdots + (\zeta_n + \eta_n)\Delta_{jn}$ with the entries of Δ_{ji}'s being the corresponding f_{ji}'s. Note that f_{ii}'s are all zero; that is, $\Delta_{ii} = 0$ for $i = 1, \ldots, n$.

We can rearrange the terms of Δ_{ij}'s to obtain $\Delta_{ij} = -\Delta_{ji}$. This is in some sense similar to the stepping stone algorithm in operations research. We give an outline of the proof here. For a proof of this, we return to the variables γ_i's. Again, it is enough to consider any (k, l)th entry. As before, let p_i be the (k, l)th entry

of Ψ_i. Let $p_i = \sum_{\mathbf{k}} f_{i,\mathbf{k}} \gamma^{\mathbf{k}}$, where γ is a monomial in terms of γ_i's, $i = 1, \ldots, n$. Thus $\gamma^{\mathbf{k}} = \gamma_1^{k_1} \cdots \gamma_n^{k_n}$, where the multi-index $\mathbf{k} = (k_1, \cdots, k_n)$. Note that $f_{i,\mathbf{k}}$ are polynomials in γ_{i+n}'s, $i = 1, \ldots, n$. Let $\mathbf{k_j} = (k_1, \ldots, k_{j-1}, k_j - 1, k_{j+1}, \ldots, k_n)$, and similarly $\mathbf{k_{ij}}$ is a multi-index defined in the obvious way. Fix a multi-index $\mathbf{k}$ and hence a monomial $\gamma^{\mathbf{k}}$. We shall demonstrate how to rearrange the terms for this particular monomial, and similar operations should be carried out for each such monomial. From the condition $\sum_{i=1}^n \gamma_i p_i = 0$, we can conclude that $\sum_{i=1}^n f_{i,\mathbf{k_i}} = 0$ (since they are the coordinates of the monomial $\gamma^{\mathbf{k}}$). Note that some of these $f_{i,\mathbf{k_i}}$'s may be zero. Set the coordinate of $\gamma^{\mathbf{k_{12}}}$ in the corresponding term ((k,l)th term) in Δ_{12} to be $f_{1,\mathbf{k_1}}$. We force the coordinate of $\gamma^{\mathbf{k_{12}}}$ in the (k,l)th term of Δ_{21} to be $-f_{1,\mathbf{k_1}}$ and the coordinate of $\gamma^{\mathbf{k_{23}}}$ in the (k,l)th term of Δ_{23} to be $f_{2,\mathbf{k_2}} + f_{1,\mathbf{k_1}}$ and so on. We perform the above operation for every multi-index $\mathbf{k}$ that might be involved. (These are finite in number.) Then $p_j = \sum_{i=1}^n \gamma_i g_{ji}$, where g_{ji} are the polynomials obtained by the above procedure and are the (k,l)th terms of the corresponding Δ_{ji}'s. We can easily check the entries of $\Delta_{ij} = -\Delta_{ji}$. Thus we have demonstrated how to construct the required Δ_{ij}'s that satisfy the conditions of the proposition. □

Proof of Theorem 7. (2) ⇒ (1) Following the lines of the proof of Theorem 4, we can conclude that $\int_\Omega Q_\Phi(w) d\mathbf{x}$ is independent of path on $\mathfrak{B}$ if and only if

$$\int_\Omega Q_\Phi(w) d\mathbf{x} + 2 \int_\Omega L_\Phi(w, v) d\mathbf{x} = 0 \tag{27}$$

for any $w \in \mathfrak{D}(\mathbb{R}^n, \mathbb{R}^{\mathbb{w}}) \cap \mathfrak{B}$ with support in Ω, any $v \in \mathfrak{C}^\infty(\mathbb{R}^n, \mathbb{R}^{\mathbb{w}}) \cap \mathfrak{B}$, and any closed bounded set $\Omega \subset \mathbb{R}^n$. Again, since $v = 0$ is a trajectory in $\mathfrak{B}$, we can conclude that $\int_\Omega Q_\Phi(w) d\mathbf{x} = 0$ for any $w \in \mathfrak{D}(\mathbb{R}^n, \mathbb{R}^{\mathbb{w}}) \cap \mathfrak{B}$ with support in Ω. Thus $\mathfrak{B}$ is lossless with respect to the QDF Q_Φ.

(1) ⇒ (4) Since $\int_\Omega Q_\Phi(w) d\mathbf{x} = 0$ for any $w \in \mathfrak{B} \cap \mathfrak{D}(\mathbb{R}^n, \mathbb{R}^{\mathbb{w}})$ with support in Ω, from (27) we can conclude that $\int_\Omega L_\Phi(w, v) d\mathbf{x} = 0$ for all $w \in \mathfrak{D}(\mathbb{R}^n, \mathbb{R}^{\mathbb{w}}) \cap \mathfrak{B}$ with support in Ω and all $v \in \mathfrak{C}^\infty(\mathbb{R}^n, \mathbb{R}^{\mathbb{w}}) \cap \mathfrak{B}$. Using integration by parts, it is easy to see that the integral $\int_\Omega (w^T f) d\mathbf{x}$ equals zero for all $w \in \mathfrak{D}(\mathbb{R}^n, \mathbb{R}^{\mathbb{w}}) \cap \mathfrak{B}$ with support in Ω if f is in the image of the operator $R^T(-\frac{d}{d\mathbf{x}})$. In $L_\Phi(w, v)$, every v we consider is of the form $v = M(\frac{d}{d\mathbf{x}})\ell$, and so we conclude that $\Phi(-\xi, \xi) M(\xi) = R^T(-\xi) Y(\xi)$ for some Y. Now premultiplying by $M^T(-\xi)$, we obtain $\Phi'(-\xi, \xi) = M^T(-\xi)\Phi(-\xi, \xi) M(\xi) = M^T(-\xi) R^T(-\xi) Y(\xi) = 0$ since $R(\xi) M(\xi) = 0$.

(4) ⇒ (2) From [22], we know that, given a controllable behavior $\mathfrak{B}$, there exists some operator $M_1(\xi)$ such that $\mathfrak{B} = \{w : w = M_1(\frac{d}{d\mathbf{x}})\ell\}$ with the additional property that every $w \in \mathfrak{B} \cap \mathfrak{D}(\mathbb{R}^n, \mathbb{R}^{\mathbb{w}})$ can be obtained as the image of some ℓ which is itself compactly supported. It can be shown that if $\Phi'(-\xi, \xi) = M^T(-\xi)\Phi(-\xi, \xi) M(\xi) = 0$ for some image representation of $M(\xi)$, then $M'^T(-\xi)\Phi(-\xi, \xi) M'(\xi) = 0$ for every image representation $M'(\frac{d}{d\mathbf{x}})$ of the controllable behavior $\mathfrak{B}$. In particular, $\Phi_1(\zeta, \eta) = M_1^T(\zeta)\Phi(\zeta, \eta) M_1(\eta)$ has the property that $\Phi_1(-\xi, \xi) = 0$. Using integration by parts, we can now show that $\int_\Omega L_\Phi(w, v) d\mathbf{x} = \int_\Omega L_{\Phi_1}(\ell_1, \ell_2) d\mathbf{x} = 0$ for any $w \in \mathfrak{D}(\mathbb{R}^n, \mathbb{R}^{\mathbb{w}}) \cap \mathfrak{B}$ with support in Ω, $w = M_1(\frac{d}{d\mathbf{x}})\ell_1$ with ℓ_1 having compact support, and $v = M_1(\frac{d}{d\mathbf{x}})\ell_2$. A similar argument also proves that $\int_\Omega Q_\Phi(w) d\mathbf{x} = 0$ for all $w \in \mathfrak{D}(\mathbb{R}^n, \mathbb{R}^{\mathbb{w}}) \cap \mathfrak{B}$ with support in Ω. From (27), we then conclude that the QDF induced by Φ is independent of path on $\mathfrak{B}$.

(3) ⇔ (4) This is already shown in Theorem 4.

(4) ⇔ (5) From Theorem 4, it is clear that there exists $\Psi \in (\mathbb{R}^{m \times m}[\zeta, \eta])^n$ such that $\mathrm{div} Q_\Psi(\ell) = Q_{\Phi'}(\ell)$ for all $\ell \in \mathfrak{C}^\infty(\mathbb{R}^n, \mathbb{R}^m)$. Moreover, since $M(\frac{d}{d\mathbf{x}})$ gives an

image representation of the behavior $\mathfrak{B}$, $Q_\Phi(w) = Q_{\Phi'}(\ell)$ for all w and ℓ related by $w = M(\frac{d}{dx})\ell$. □

Proof of Theorem 10. We will start by considering the scalar case. Let $\Gamma \in \mathbb{R}[\xi]$ with $\Gamma^\star = \Gamma$ and $\Gamma(i\omega) \geq 0$ for all $\omega \in \mathbb{R}^n$. From the above data, we can conclude that the polynomial p defined as $p(\omega) := \Gamma(i\omega)$ is a positive definite (or nonnegative) function. Using the result from Hilbert's 17th problem [1], we can write

$$p(\omega) = \sum_{i=1}^{r} \left(f_i(\omega)\right)^2,$$

where each $f_i \in \mathbb{R}(\omega)$; i.e., f_i's are rational functions in ω. Now consider each of these f_i's. We can write them as

$$f_i(\omega) = f_{i+}(\omega) + f_{i-}(\omega),$$

where

$$f_{i+}(\omega) = \frac{f_i(\omega) + f_i(-\omega)}{2},$$
$$f_{i-}(\omega) = \frac{f_i(\omega) - f_i(-\omega)}{2}.$$

Thus we separate the even and odd parts of the functions f_i. Now

$$\begin{aligned} p(\omega) &= \sum_{i=1}^{r} \left(f_{i+}(\omega) + f_{i-}(\omega)\right)^2 \\ &= \sum_{i=1}^{r} \left[\left(f_{i+}(\omega)\right)^2 + \left(f_{i-}(\omega)\right)^2 + 2f_{i+}(\omega)f_{i-}(\omega)\right], \end{aligned}$$

and

$$\begin{aligned} p(-\omega) &= \sum_{i=1}^{r} \left(f_{i+}(-\omega) + f_{i-}(-\omega)\right)^2 \\ &= \sum_{i=1}^{r} \left[\left(f_{i+}(\omega)\right)^2 + \left(f_{i-}(\omega)\right)^2 - 2f_{i+}(\omega)f_{i-}(\omega)\right]. \end{aligned}$$

Since $p(\omega) = p(-\omega)$ (from the initial assumption on Γ), we can conclude $\sum_{i=1}^{r} f_{i+}(\omega)f_{i-}(\omega) = 0$, and so

$$\begin{aligned} p(\omega) &= \sum_{i=1}^{r} \left[\left(f_{i+}(\omega)\right)^2 + \left(f_{i-}(\omega)\right)^2\right] \\ &= \sum_{i=1}^{r} \left[f_{i+}(\omega) + if_{i-}(\omega)\right]\left[f_{i+}(\omega) - if_{i-}(\omega)\right] \\ &= \sum_{i=1}^{r} \left[f_{i+}(\omega) + if_{i-}(\omega)\right]\left[f_{i+}(-\omega) + if_{i-}(-\omega)\right]. \end{aligned}$$

Thus note that p has now been written down as the sum of r terms, each of which is a product of a function evaluated at ω and $-\omega$. Construct an $r \times 1$ matrix E

whose $(i,1)$th entry is $(f_{i+} + if_{i-})$. Then $\Gamma(i\omega) = p(\omega) = E^T(-\omega)E(\omega)$, which in turn implies that $\Gamma(\xi) = F^T(-\xi)F(\xi)$, where F is obtained from E by the obvious substitution; i.e., the $(i,1)$th entry of F is $(f_{i+} + f_{i-})$.

We now tackle the case where $\Gamma \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\xi]$ with the properties $\Gamma^T(-\xi) = \Gamma(\xi)$ and $\Gamma(i\omega) \geq 0$ for all $\omega \in \mathbb{R}^n$. Let $\Gamma(\xi) = [a_{ij}(\xi)]$; i.e., let the polynomial in the (i,j)th coordinate of Γ be denoted by a_{ij}. Clearly, all diagonal elements a_{ii} are even polynomials in ξ since $\Gamma^T(-\xi) = \Gamma(\xi)$. Let $I = \{i : a_{ii}(\xi) = 0\}$. Consider the submatrix obtained from Γ by choosing elements a_{ij}, where both $i, j \in I$. Let us call this matrix $H(\xi) = [a_{ij}(\xi)]_{i,j\in I}$. Observe that $\Gamma(i\omega)$ (and hence $H(i\omega)$) are Hermetian matrices for any $\omega \in \mathbb{R}^n$. Since the trace of $H(i\omega)$ is zero, we can conclude that $H(i\omega)$ has a negative real eigenvalue (provided $H(i\omega)$ has nonzero entries). In this case, $H(i\omega) \not\geq 0$, and hence $\Gamma(i\omega) \not\geq 0$. However, this contradicts the assumption that $\Gamma(i\omega) \geq 0$. Hence we conclude that $H(\xi) = 0$.

We now construct a 2×2 submatrix of $\Gamma(\xi)$, denoted by $H_1(\xi)$, by considering the four elements $a_{ij}(\xi) \in \Gamma(\xi)$, where $i \in I$ and $j \notin I$ are fixed. Clearly, the determinant of $H_1(\xi)$ is given by $-a_{ij}(\xi)a_{ji}(\xi) = -a_{ij}(\xi)a_{ij}(-\xi)$, and this is an even polynomial in ξ. Substituting $i\omega$ for ξ, we obtain $H_1(i\omega)$—a 2×2 Hermetian matrix with a negative determinant. This implies a negative real eigenvalue, and hence $H_1(i\omega) \not\geq 0$. But since $\Gamma(i\omega) \geq 0$, we conclude that $a_{ij}(\xi) = a_{ji}(\xi) = 0$. From the above discussion, we conclude that the rows and columns corresponding to $i \in I$ have only zero entries.

We now define and prove an algorithmic step. In this step, we take as an input a $\mathtt{w}\times\mathtt{w}$ matrix Γ that has the following properties: (i) $\Gamma^\star = \Gamma$ and (ii) $\Gamma(i\omega) \geq 0$ for all $\omega \in \mathbb{R}^n$. The outputs of this algorithmic step are two matrices, which we call E and Q. E has the same number of columns as Γ. The matrix Q has the same properties as Γ but is a $(\mathtt{w}-1)\times(\mathtt{w}-1)$ matrix. Let

$$\Gamma = \begin{bmatrix} a_{11} & \gamma \\ \gamma^\star & \Gamma_{\mathtt{w}-1} \end{bmatrix}.$$

If $a_{11} \neq 0$, then construct $F_1(\xi)$ such that $a_{11}(\xi) = F_1^T(-\xi)F_1(\xi)$. Define the matrix $E = \begin{bmatrix} F_1 & F_1 a_{11}^{-1}\gamma \end{bmatrix}$. Clearly,

$$\begin{bmatrix} F_1^\star \\ \gamma^\star(a_{11}^{-1})^\star F_1^\star \end{bmatrix} \begin{bmatrix} F_1 & F_1 a_{11}^{-1}\gamma \end{bmatrix} = \begin{bmatrix} a_{11} & \gamma \\ \gamma^\star & \gamma^\star(a_{11}^{-1})^\star\gamma \end{bmatrix}.$$

Note that $(a_{11}^{-1})^\star = a_{11}^{-1}$. On the other hand, if $a_{11} = 0$, from the discussion above we know $\gamma = 0$. In this case, define the matrix $E = \begin{bmatrix} 0_{1\times 1} & 0_{1\times\{\mathtt{w}-1\}} \end{bmatrix}$. Define the $(\mathtt{w}-1)\times(\mathtt{w}-1)$ matrix

$$Q = \begin{cases} \Gamma_{\mathtt{w}-1} - \gamma^\star a_{11}^{-1}\gamma & \text{when} \quad a_{11}, \neq 0 \\ \Gamma_{\mathtt{w}-1} & \text{when} \quad a_{11} = 0. \end{cases}$$

(When $a_{11} \neq 0$, then the above matrix Q is known in the literature as the Schur complement of a_{11} in Γ; see [7, 9].) Clearly, the matrix Q is such that $Q^T(-\xi) = Q(\xi)$. Moreover, we claim that $Q(i\omega) \geq 0$ for all $\omega \in \mathbb{R}^n$. This is clear for the case when $a_{11} = 0$. For the case when $a_{11} \neq 0$, we have $Q = \Gamma_{\mathtt{w}-1} - \gamma^\star a_{11}^{-1}\gamma$. Suppose $Q(i\omega) \not\geq 0$ for some $\omega \in \mathbb{R}^n$. This implies that there exists a vector a such that $a^T Q(i\omega)a < 0$. Consider the vector

$$v = \begin{bmatrix} -(a_{11}^{-1}(i\omega)\gamma(i\omega)a \\ a \end{bmatrix}.$$

Then $v^T\Gamma(i\omega)v = a^TQ(i\omega)a < 0$, and this contradicts $\Gamma(i\omega) \geq 0$ for all $\omega \in \mathbb{R}^n$. Hence we have a $(\mathtt{w}-1)\times(\mathtt{w}-1)$ matrix Q such that $Q^T(-\xi) = Q(\xi)$ and $Q(i\omega) \geq 0$ for all $\omega \in \mathbb{R}^n$.

We now give the algorithm to construct F such that $\Gamma = F^T(-\xi)F(\xi)$.

1. Set $Q_0 = \Gamma$. Set $i = 0$.
2. Invoke the algorithmic step defined above with input Q_i to obtain outputs E and Q. Set $E_{i+1} = E$ and $Q_{i+1} = Q$.
3. If $Q_{i+1} \neq 0$, increment i and go back to step 2.
4. Construct

$$F = \begin{bmatrix} E_1 & \\ 0_{\bullet\times 1} & E_2 \\ 0_{\bullet\times 2} & E_3 \\ \vdots & \vdots \\ 0_{\bullet\times i} & E_i \end{bmatrix}.$$

It is now easy to check that $\Gamma = F^T(-\xi)F(\xi)$. □

Proof of Proposition 11. ($\Leftarrow$) Since $w \in \mathfrak{D}(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}})$, we can take the multidimensional Fourier transform. Let $\hat{w}$ denote the multidimensional Fourier transform of w. Then, using Parseval's theorem, we have

$$\int_{\mathbb{R}^n} Q_\Phi(w)d\mathbf{x} = \frac{1}{(2\pi)^n}\int_{\mathbb{R}^n} \hat{w}(-i\omega)\Phi(-i\omega, i\omega)\hat{w}(i\omega)d\omega, \tag{28}$$

where $\omega \in \mathbb{R}^n$. Then ($\Leftarrow$) is clear.

($\Rightarrow$) Consider a compactly supported function $u(\mathbf{x})$, where the variables can be separated, i.e., $u(\mathbf{x}) = v_1(x_1)v_2(x_2)\cdots v_n(x_n)$. Let $\int_{\mathbb{R}^n} Q_\Phi(u)d\mathbf{x} = E$. Suppose $\Phi(-i\omega, i\omega) \not\geq 0$ for all $\omega \in \mathbb{R}^n$. Let $a \in \mathbb{C}^q$ and $\omega_0 = (\omega_1, \ldots, \omega_n) \in \mathbb{R}^n$ be such that $a^T\Phi(-i\omega_0, i\omega_0)a < 0$. Without loss of generality, we will assume that $\omega_1 \neq 0$. Then we can choose functions $u_N(\mathbf{x})$ given by $u_N(\mathbf{x}) = v_{1,N}(x_1)v_{2,N}(x_2)\cdots v_{n,N}(x_n)$, where

$$v_{1,N}(x_1) = \begin{cases} e^{i<\omega_1, x_1>}a, & |\, x_1 \,| \leq \frac{2\pi N}{\omega_1}, \\ v_1(x_1 + \frac{2\pi N}{\omega_1}), & x_1 < -\frac{2\pi N}{\omega_1}, \\ v_1(x_1 - \frac{2\pi N}{\omega_1}), & x_1 > \frac{2\pi N}{\omega_1}, \end{cases} \tag{29}$$

and for $i = 2, \ldots, n$ and $\omega_i \neq 0$,

$$v_{i,N}(x_i) = \begin{cases} e^{i<\omega_i, x_i>}[1, 1, \ldots, 1]^T, & |\, x_i \,| \leq \frac{2\pi N}{\omega_i}, \\ v_i(x_i + \frac{2\pi N}{\omega_i}), & x_i < -\frac{2\pi N}{\omega_i}, \\ v_i(x_i - \frac{2\pi N}{\omega_i}), & x_i > \frac{2\pi N}{\omega_i}, \end{cases} \tag{30}$$

and for $i = 2, \ldots, n$ with $\omega_i = 0$, we have

$$v_{i,N}(x_i) = v_i(x_i). \tag{31}$$

Then, on evaluating $\int_{\mathbb{R}^n} Q_\Phi(u_N)d\mathbf{x}$, one gets a negative term that depends on N and another term that is independent of N. Then, by choosing N large enough, this integral can be made negative and hence we obtain a contradiction to $\int Q_\Phi \geq 0$. □

Proof of Theorem 12. Note that (4) is a statement that combines (2) and (3). So we first we show the equivalence of the first three statements. Then we show how statement (4) is equivalent.

(1) $\Rightarrow$ (3) Let $\Phi = \Phi^* \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta,\eta]$. By Proposition 11, we know that $\int Q_\Phi \geq 0$ implies that $\Phi(-i\omega, i\omega) \geq 0$ for all $\omega \in \mathbb{R}^n$. Thus $\Phi(-\xi,\xi)$ satisfies all the conditions on Γ in Theorem 10, and hence $\Phi(-\xi,\xi) = F^T(-\xi)F(\xi)$ for some $F \in \mathbb{R}^{\bullet\times\mathtt{w}}(\xi)$.

Let us now consider the matrix $F \in \mathbb{R}^{r\times\mathtt{w}}(\xi)$. (Here we assume that the number of rows of F is some arbitrary (finite) number r.) It is easy to show that $F = DM^{-1}$, where $D \in \mathbb{R}^{r\times\mathtt{w}}[\xi]$ and $M \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\xi]$. Note that both D and M are now matrices with polynomial entries. Moreover, we can choose M such that it is a diagonal matrix (and hence the operator $M(\frac{d}{d\mathbf{x}})$ is surjective). We can now define latent variables ℓ related to the manifest variables w by the equation $w = M(\frac{d}{d\mathbf{x}})\ell$. Now consider $\Delta \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta,\eta]$ given by the matrix $\Delta(\zeta,\eta) = D^T(\zeta)D(\eta)$. Obviously, $Q_\Delta \geq 0$. Now for all $\ell \in \mathfrak{D}(\mathbb{R}^n,\mathbb{R}^{\mathtt{w}})$ we have $M(\frac{d}{d\mathbf{x}})\ell = w \in \mathfrak{D}(\mathbb{R}^n,\mathbb{R}^{\mathtt{w}})$. Hence it follows that

$$\begin{aligned}\int_{\mathbb{R}^n} Q_\Phi(w)d\mathbf{x} &= \int_{\mathbb{R}^n} (w)^T \left(\Phi\left(-\frac{d}{d\mathbf{x}},\frac{d}{d\mathbf{x}}\right)w\right)d\mathbf{x}\\ &= \int_{\mathbb{R}^n} \left(M\left(\frac{d}{d\mathbf{x}}\right)\ell\right)^T \left(\Phi\left(-\frac{d}{d\mathbf{x}},\frac{d}{d\mathbf{x}}\right)M\left(\frac{d}{d\mathbf{x}}\right)\ell\right)d\mathbf{x}\\ &= \int_{\mathbb{R}^n} Q_\Delta(\ell)d\mathbf{x}.\end{aligned}$$

Thus the matrix $\Delta(\zeta,\eta) = D^T(\zeta)D(\eta)$ induces a QDF, which defines a dissipation function for the given Φ.

(3) $\Rightarrow$ (2) Let $\Delta \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta,\eta]$ induce a QDF on the latent variables ℓ such that Q_Δ is a dissipation function associated to the given Φ. Moreover, let $M \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\xi]$ induce a surjective operator, such that $w = M(\frac{d}{d\mathbf{x}})\ell$. Consider the system defined by $w - M(\frac{d}{d\mathbf{x}})\ell = 0$ in the variables (w,ℓ) with a QDF induced by

$$\begin{bmatrix}\Phi & 0\\ 0 & -\Delta\end{bmatrix} \in \mathbb{R}^{(2\mathtt{w})\times(2\mathtt{w})}[\zeta,\eta].$$

Note that the system given by $w - M(\frac{d}{d\mathbf{x}})\ell = 0$ is a controllable system. In fact, this system has an observable image representation given by

$$\begin{bmatrix} w\\ \ell\end{bmatrix} = \begin{bmatrix} M(\frac{d}{d\mathbf{x}})\\ I\end{bmatrix}\ell$$

(for details see [19]). The QDF defined above gives

$$\int_{\mathbb{R}^n} [Q_\Phi(w) - Q_\Delta(\ell)]\, d\mathbf{x} = 0$$

on every compactly supported trajectory in the full behavior involving both w and ℓ, and so, applying Theorem 7, we obtain a QDF $\Phi'(\zeta,\eta) = M^T(\zeta)\Phi(\zeta,\eta)M(\eta) - \Delta(\zeta,\eta)$ which acts on the variables ℓ. Since $\Phi'(-\xi,\xi) = 0$, we can find a corresponding VQDF Q_Ψ acting on the latent variables ℓ such that $Q_\Phi(w) - Q_\Delta(\ell) = \operatorname{div} Q_\Psi(\ell)$. This VQDF can now, in turn, be interpreted as a storage function for the original Φ since $\operatorname{div} Q_\Psi(\ell) \leq Q_\Phi(w)$ for all $\ell \in \mathfrak{D}(\mathbb{R}^n,\mathbb{R}^{\mathtt{w}})$.

(2) $\Rightarrow$ (1) Let Φ admit a storage function. So $\operatorname{div} Q_\Psi(\ell) \leq Q_\Phi(w)$ for all $\ell \in \mathfrak{D}(\mathbb{R}^n,\mathbb{R}^{\mathtt{w}})$ and $w = M(\frac{d}{d\mathbf{x}})\ell$. On integrating this inequality over all of $\mathbb{R}^n$, we get $0 \leq \int_{\mathbb{R}^n} Q_\Phi(w)d\mathbf{x}$. Now consider any $w \in \mathfrak{D}(\mathbb{R}^n,\mathbb{R}^{\mathtt{w}})$. Since $M(\frac{d}{d\mathbf{x}})$ is surjective, there exist some $\ell_0 \in \mathfrak{C}^\infty(\mathbb{R}^n,\mathbb{R}^{\mathtt{w}})$ such that $w = M(\frac{d}{d\mathbf{x}})\ell_0$. Choose a sequence of $\ell_i \in$

$\mathfrak{D}(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}})$ such that $\lim_{i\to\infty} \ell_i = \ell_0$. Let $w_i = M(\frac{d}{d\mathbf{x}})\ell_i$. Clearly, $w_i \in \mathfrak{D}(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}})$. Hence, by continuity, we conclude that $\int_{\mathbb{R}^n} Q_\Phi(w)d\mathbf{x} \geq 0$. Thus $\int Q_\Phi \geq 0$.

(1) $\Rightarrow$ (4) Given a Φ, the steps in (1) $\Rightarrow$ (3) above explain how Δ can be obtained. The proof of (3) $\Rightarrow$ (2) explains how Ψ can be obtained. Observe that the latent variables involved in the construction of Ψ are the latent variables ℓ. These are the latent variables associated with the image representation of the controllable behavior given by $w - M(\frac{d}{d\mathbf{x}})\ell = 0$. Thus we obtain the equation

$$\operatorname{div} Q_\Psi(\ell) = Q_\Phi(w) - Q_\Delta(\ell).$$

Note that the same set of latent variables is associated to storage and dissipation.

(4) $\Rightarrow$ (1) This is obvious from (2) $\Rightarrow$ (1). ∎

Proof of Theorem 13. (1) $\Rightarrow$ (2) First, we convert the problem into a problem on a free behavior by using the image representation of the behavior. Let $M' \in \mathbb{R}^{\mathtt{w}\times\mathtt{l}}[\xi]$ define an operator $M'(\frac{d}{d\mathbf{x}})$ that defines an image representation of the given behavior $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}_{n,\mathrm{cont}}$. Define $\Phi' := M'^T(\zeta)\Phi(\zeta,\eta)M'(\eta)$, and look at the action of the QDF induced by Φ' on the full space of latent variables ℓ' that appear in the image representation of $\mathfrak{B}$ (i.e., $\mathfrak{B} = \{w \mid w = M'(\frac{d}{d\mathbf{x}})\ell'\}$). Clearly, $\int_{\mathbb{R}^n} Q_{\Phi'}(\ell') \geq 0$ for all $\ell' \in \mathfrak{D}(\mathbb{R}^n, \mathbb{R}^\ell)$. By Theorem 12, we can find $\Psi = (\Psi_1, \Psi_2, \ldots, \Psi_n)$ with $\Psi_k = \Psi_k^* \in \mathbb{R}^{\mathtt{l}\times\mathtt{l}}[\zeta,\eta]$ for $k = 1, 2, \ldots, n$ and $\Delta = \Delta^* \in \mathbb{R}^{\mathtt{l}\times\mathtt{l}}[\zeta,\eta]$ such that

$$\operatorname{div} Q_\Psi(\ell) = Q_{\Phi'}(\ell') - Q_\Delta(\ell)$$

with ℓ' and ℓ related by $\ell' = M(\frac{d}{d\mathbf{x}})\ell$ and $M \in \mathbb{R}^{\mathtt{l}\times\mathtt{l}}[\xi]$. Observe that by definition $Q_\Phi(w) = Q_{\Phi'}(\ell')$, where $w = M'(\frac{d}{d\mathbf{x}})\ell'$. Hence we obtain the required result with w and ℓ related by $w = M'(\frac{d}{d\mathbf{x}})M(\frac{d}{d\mathbf{x}})\ell$, which is also an image representation of $\mathfrak{B}$, since $M(\frac{d}{d\mathbf{x}})$ is surjective (Theorem 12).

(2) $\Rightarrow$ (1) Consider any $w \in \mathfrak{B} \cap \mathfrak{D}(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}})$. Then $w = M(\frac{d}{d\mathbf{x}})\ell$ for some $\ell \in \mathfrak{C}^\infty(\mathbb{R}^n, \mathbb{R}^\ell)$. Clearly, if $\ell \in \mathfrak{D}(\mathbb{R}^n, \mathbb{R}^\ell)$, then

$$0 = \int_{\mathbb{R}^n} \operatorname{div} Q_\Psi(\ell)d\mathbf{x} = \int_{\mathbb{R}^n} Q_\Phi(w)d\mathbf{x} - \int_{\mathbb{R}^n} Q_\Delta(\ell)d\mathbf{x}.$$

Thus $\int_{\mathbb{R}^n} Q_\Phi(w)d\mathbf{x} = \int_{\mathbb{R}^n} Q_\Delta(\ell)d\mathbf{x} \geq 0$. If $\ell \notin \mathfrak{D}(\mathbb{R}^n, \mathbb{R}^\ell)$, then one can find a sequence of $\ell_i \in \mathfrak{D}(\mathbb{R}^n, \mathbb{R}^\ell)$ that converges to ℓ. Then, by continuity, we obtain $\int_{\mathbb{R}^n} Q_\Phi(w)d\mathbf{x} \geq 0$ for all $w \in \mathfrak{B} \cap \mathfrak{D}(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}})$. ∎

REFERENCES

[1] E. Artin, *Über die Zerlegung definiter Funktionen in Quadrate*, in Collected Papers, Mathematisches Seminar, Hamburg, 1926, Addison–Wesley, Reading, MA, 1965, pp. 273–288.

[2] J. A. Ball and T. T. Trent, *Unitary colligations, reproducing kernel Hilbert spaces and Nevanlinna-pick interpolation in several variables*, J. Funct. Anal., 157 (1998), pp. 1–61.

[3] S. Basu, *A framework for two-dimensional hyperstability theory based provably convergent adaptive two-dimensional* II*R filtering*, SIAM J. Control Optim., 29 (1991), pp. 1476–1508.

[4] C. Berg, J. P. R. Christensen, and P. Ressel, *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*, Grad. Texts in Math. 100, Springer-Verlag, New York, 1984.

[5] N. K. Bose, *Matrix factorization in a real field*, Linear Algebra Appl., 11 (1975), pp. 21–25.

[6] F. M. Callier, *On polynomial spectral factorization by symmetric extraction*, IEEE Trans. Automat. Control, 30 (1985), pp. 453–464.

[7] D. Carlson, *What are Schur complements, anyway?*, Linear Algebra Appl., 74 (1986), pp. 257–275.

[8] W. A. Coppel, *Linear Systems*, Technical report, Australian National University, Canberra, Australia, 1972.
[9] R. W. Cottle, *Manifestations of the Schur complement*, Linear Algebra Appl., 8 (1974), pp. 189–211.
[10] R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics, Vol.* 2, Addison-Wesley, Reading, MA, 1964.
[11] D. Hilbert, *Über Ternäre Definite Formen*, Acta Math., 32 (1893), pp. 169–197.
[12] D. S. Kalyuzhniy, *Multiparametric dissipative linear stationary scattering systems: Discrete case*, J. Operator Theory, 43 (2000), pp. 427–460.
[13] D. S. Kalyuzhniy, *Multiparemetric dissipative linear stationary dynamical scattering systems: Discrete case* II, Integral Equations Operator Theory, 36 (2000), pp. 107–120.
[14] A. Kummert, *Synthesis of two-dimensional passive m-ports with prescribed scattering matrices*, Circuits Systems Signal Process., 8 (1989), pp. 97–119.
[15] H. Kwakernaak and M. Sebek, *Polynomial J-spectral factorization*, IEEE Trans. Automat. Control, 39 (1994), pp. 315–328.
[16] E. Landau, *Über die Darstellung definiter Funktionen durch Quadrate*, Math. Ann., 62 (1906), p. 272.
[17] A. Pfister, *Zur Darstellung definiter Funktionen als Summe von Quadraten*, Invent. Math., 4 (1967), pp. 229–237.
[18] A. Pfister, *Hilbert's seventeenth problem and related problems on definite forms*, in Mathematical Developments Arising from Hilbert Problems, Proc. Sympos. Pure Math. 28, Felix E. Browder, ed., AMS, Providence, RI, 1974, pp. 483–489.
[19] H. K. Pillai and S. Shankar, *A behavioral approach to control of distributed systems*, SIAM J. Control Optim., 37 (1998), pp. 388–408.
[20] J. W. Polderman and J. C. Willems, *Introduction to Mathematical Systems Theory: A Behavioural Approach*, Springer-Verlag, New York, 1998.
[21] A. C. M. Ran and L. Rodman, *Factorization of matrix polynomials with symmetries*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 845–864.
[22] S. Shankar, *A Duality Theorem for Systems of Partial Differential Equations*, preprint, Chennai Mathematical Institute, Chennai, India.
[23] J. C. Willems, *Dissipative dynamical systems—part* I: *General theory, part* II: *Linear systems with quadratic supply rates*, Arch. Ration. Mech. Anal., 45 (1972), pp. 321–351 and pp. 352–393.
[24] J. C. Willems, *Paradigms and puzzles in the theory of dynamical systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 259–294.
[25] J. C. Willems and H. L. Trentelman, *Synthesis of dissipative systems using quadratic differential forms: Parts* I *and* II, IEEE Trans. Automat. Control, 47 (2002), to appear.
[26] J. C. Willems and H. L. Trentelman, *On quadratic differential forms*, SIAM J. Control Optim., 36 (1998), pp. 1703–1749.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 5, pp. 1431–1454

SECOND ORDER OPTIMALITY CONDITIONS FOR SEMILINEAR ELLIPTIC CONTROL PROBLEMS WITH FINITELY MANY STATE CONSTRAINTS*

EDUARDO CASAS† AND MARIANO MATEOS‡

Abstract. This paper deals with necessary and sufficient optimality conditions for control problems governed by semilinear elliptic partial differential equations with finitely many equality and inequality state constraints. Some recent results on this topic for optimal control problems based upon results for abstract optimization problems are compared with some new results using methods adapted to the control problems. Meanwhile, the Lagrangian formulation is followed to provide the optimality conditions in the first case; the Lagrangian and Hamiltonian functions are used in the second statement. Finally, we prove the equivalence of both formulations.

Key words. necessary and sufficient optimality conditions, control of elliptic equations, state constraints

AMS subject classifications. 49K20, 35J25

PII. S0363012900382011

1. Introduction.

The first goal of this paper is to provide some new second order optimality conditions for control problems of semilinear elliptic partial differential equations with finitely many state constraints. These conditions involve the Lagrangian and the Hamiltonian functions. Therefore, they are not a consequence of some abstract theorems in optimization theory but are proved by using arguments valid only in the framework of control theory. The second goal is to compare these conditions with those obtained recently for the same type of problems by using theorems for abstract optimization in infinite-dimensional spaces.

While there exists a very extensive literature about first order optimality conditions for control problems of partial differential equations, only a few papers are devoted to second order conditions. However, some progress has been made in the last few years. Most of the papers have been devoted to the study of sufficient second order optimality conditions; see Goldberg and Tröltzsch [13], Casas, Tröltzsch, and Unger [9], [10], Raymond and Tröltzsch [20]. Such sufficient optimality conditions are useful for carrying out the numerical analysis of a control problem, for obtaining error estimates in the numerical discretization, and for analyzing the sequential quadratic programming algorithms applied to control problems. However, we also have to study the second order necessary conditions and compare them with the sufficient ones in order to check if there is a reasonable gap between them. This was studied by Casas and Tröltzsch [7], [8] and Casas, Mateos, and Fernández [5] for some control problems. In the last papers, the authors proved the results by using some methods of abstract optimization theory and by stating some new results in this abstract framework. The gap between the established necessary and the sufficient conditions was very small.

*Received by the editors December 4, 2000; accepted for publication (in revised form) June 23, 2001; published electronically January 18, 2002. This research was partially supported by Dirección General de Enseñanza Superior e Investigación Científica (Spain).
http://www.siam.org/journals/sicon/40-5/38201.html

†Departamento de Matemática Aplicada y Ciencias de la Computación, E.T.S.I. Industriales y de Telecomunicación, Universidad de Cantabria, 39005 Santander, Spain (eduardo.casas@unican.es).

‡Departamento de Matemáticas, EUIT Industriales, Universidad de Oviedo. C/Manuel Llaneza, Gijón, Spain (mmateos@orion.ciencias.uniovi.es).

Bonnans and Zidani [2] extended the results for finite-dimensional optimization problems to control problems by assuming that the second derivative with respect to the control of the Lagrangian function is a Legendre form. This is the natural way of doing such an extension, but the inconvenience is that the hypothesis about the Lagrangian function works in only a few cases. In this paper, instead of assuming that the second derivative of the Lagrangian function is a Legendre form, we assume a strict positivity condition on the second derivative of the Hamiltonian function with respect to the control, which is quite close to the necessary relaxed positivity.

The plan of the paper is as follows. In the next section, the control problem is formulated and some derivability results of the functionals are stated. In section 3, we reformulate the control problem as an infinite-dimensional optimization problem with constraints and we apply the second order conditions as deduced in [7] to our particular situation. Finally, in section 4 we deduce necessary and sufficient second order conditions involving the Lagrangian and the Hamiltonian functions and compare them with those established in section 3.

2. The control problem. Let Ω be an open bounded set in $\mathbb{R}^N$ with a boundary Γ of class C^1, and A an elliptic operator of the form

$$Ay = -\sum_{i,j=1}^{N} \partial_{x_j}\left[a_{ij}\partial_{x_i} y\right] + a_0 y,$$

where the coefficients a_{ij} belong to $C(\bar{\Omega})$ and satisfy

$$m\|\xi\|^2 \le \sum_{i,j=1}^{N} a_{ij}(x)\xi_i\xi_j \le M\|\xi\|^2 \ \forall \xi \in \mathbb{R}^N \text{ and } \forall x \in \Omega$$

for some $m, M > 0$ and $a_0 \in L^r(\Omega)$ is not identically zero, with $r \ge Np/(N+p)$ for some $p > N$ fixed, $a_0(x) \ge 0$ in Ω. Let f and L be Carathéodory functions $f : \Omega \times \mathbb{R}^2 \to \mathbb{R}$ and $L : \Omega \times \mathbb{R}^2 \longrightarrow \mathbb{R}$, n_e and n_i be nonnegative integers, and for every $1 \le j \le n_e + n_i$ let us consider a function $F_j : W^{1,p}(\Omega) \longrightarrow \mathbb{R}$.

The control problem is formulated as follows:

$$(\mathbf{P}) \begin{cases} \text{Minimize } J(u) = \int_\Omega L(x, y_u(x), u(x))dx, \\ u_a(x) \le u(x) \le u_b(x) \text{ a.e. } x \in \Omega, \\ F_j(y_u) = 0, \ 1 \le j \le n_e, \\ F_j(y_u) \le 0, \ n_e + 1 \le j \le n_e + n_i, \end{cases}$$

where y_u is the solution of

$$\begin{cases} Ay_u = f(x, y_u, u) & \text{in } \Omega, \\ \partial_{n_A} y_u = g & \text{on } \Gamma, \end{cases} \tag{2.1}$$

$g \in L^{p(1-1/N)}(\Gamma)$ and u_a, $u_b \in L^\infty(\Omega)$, $u_a(x) \le u_b(x)$ for almost every (a.e.) $x \in \Omega$. Let us state the following assumptions on the functional F_j, L, and f.

(A1) f is of class C^2 with respect to the second and third variables,

$$f(\cdot, 0, 0) \in L^{Np/(N+p)}(\Omega), \qquad \frac{\partial f}{\partial y}(x, y, u) \le 0,$$

and for all $M > 0$ there exists a constant $C_{f,M} > 0$ such that

$$\left|\frac{\partial f}{\partial y}(x, y, u)\right| + \left|\frac{\partial f}{\partial u}(x, y, u)\right| + \left|\frac{\partial^2 f}{\partial y^2}(x, y, u)\right| + \left|\frac{\partial^2 f}{\partial y \partial u}(x, y, u)\right| + \left|\frac{\partial^2 f}{\partial u^2}(x, y, u)\right| \le C_{f,M}$$

for a.e. $x \in \Omega$ and $|y|, |u| \le M$. Moreover, given $\rho > 0$ arbitrary, for every $\varepsilon > 0$ there exists $\delta > 0$ such that for almost every point $x \in \Omega$ and $|y_i|, |u_i| \le \rho$, $i = 1, 2$, we have

$$|D^2_{(y,u)} f(x, y_2, u_2) - D^2_{(y,u)} f(x, y_1, u_1)| < \varepsilon \ \text{ if } \ |y_2 - y_1| < \delta, \ |u_2 - u_1| < \delta,$$

where $D^2_{(y,u)} f$ denotes the second derivative of f with respect to (y, u).

(A2) $L : \Omega \times \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}$ is of class C^2 with respect to the second and third variables, $|L(\cdot, 0, 0)| \in L^1(\Omega)$, and for all $M > 0$ there exists a constant $C_M > 0$ and functions $\psi_M \in L^{Np/(N+p)}(\Omega)$ and $\psi^*_M \in L^2(\Omega)$ such that

$$\left|\frac{\partial L}{\partial y}(x, y, u)\right| \le \psi_M(x), \ \left|\frac{\partial L}{\partial u}(x, y, u)\right| \le \psi^*_M(x),$$

and

$$\left|\frac{\partial^2 L}{\partial y^2}(x, y, u)\right| + \left|\frac{\partial^2 L}{\partial y \partial u}(x, y, u)\right| + \left|\frac{\partial^2 L}{\partial u^2}(x, y, u)\right| \le C_M$$

for a.e. $x \in \Omega$ and $|y|, |u| \le M$. Finally, given $\rho > 0$ arbitrary, for every $\varepsilon > 0$ there exists $\delta > 0$ such that for almost every point $x \in \Omega$ and $|y_i|, |u_i| \le \rho$, $i = 1, 2$, we have

$$|D^2_{(y,u)} L(x, y_2, u_2) - D^2_{(y,u)} L(x, y_1, u_1)| < \varepsilon \ \text{ if } \ |y_2 - y_1| < \delta, \ |u_2 - u_1| < \delta,$$

where $D^2_{(y,u)} L$ denotes the second derivative of L with respect to (y, u).

(A3) For every $1 \le j \le n_e + n_i$, F_j is of class C^1 in $W^{1,s}(\Omega)$ and of class C^2 in $W^{1,q}(\Omega)$, where $s \in [1, \frac{N}{N-1})$, $q \in [\max\{s, \frac{2N}{N+2}\}, \frac{2N}{N-2})$, and $q \le p$.

Remark 2.1. The continuity assumption on the coefficients a_{ij} and the C^1 regularity of the boundary of the domain will allow us to consider integral state constraints involving the derivatives of the state. Nevertheless, if the coefficients a_{ij} are only bounded and the boundary Γ is Lipschitz, some results similar to those obtained here can be derived if the constraints do not involve the gradient of the state.

Let us show some examples of state constraints included in the previous formulation.

EXAMPLE 2.2. *Integral constraints on the state.* Given $g_j : \Omega \times \mathbb{R} \longrightarrow \mathbb{R}$, we define $F_j(y) = \int_\Omega g_j(x, y(x))dx$. Assumption **(A3)** is satisfied if we make the following hypotheses: g_j is of class C^2 with respect to the second variable and measurable with respect to the first one, $g_j(\cdot, 0) \in L^1(\Omega)$, and for every $M > 0$ there exist $\psi_M \in L^{Ns/([N+1]s-N)}(\Omega)$, for some $s < N/(N-1)$, and $\psi^*_M \in L^\alpha(\Omega)$, with $\alpha = 1$ if $N < 4$ and $\alpha > N/4$ otherwise, such that for every $y, y_1, y_2 \in [-M, +M]$ and almost every $x \in \Omega$

$$\left|\frac{\partial g_j}{\partial y}(x, y)\right| \le \psi_M(x), \quad \left|\frac{\partial^2 g_j}{\partial y^2}(x, y)\right| \le \psi^*_M(x),$$

$$\forall \varepsilon > 0 \ \exists \delta > 0 \ \text{ such that } \ \left|\frac{\partial^2 g_j}{\partial y^2}(x, y_2) - \frac{\partial^2 g_j}{\partial y^2}(x, y_1)\right| \le \varepsilon \ \text{ if } |y_2 - y_1| < \delta.$$

(A3) holds for $q = \min\{p, 2N/(N-2) - \beta\} > N$ for some $\beta > 0$ small enough.

EXAMPLE 2.3. *Integral constraints on the derivatives of the state.* Given $g_j : \Omega \times \mathbb{R}^N \longrightarrow \mathbb{R}$, we now define $F_j(y) = \int_\Omega g_j(x, \nabla y(x))dx$. Then assumption **(A3)** is

fulfilled if g_j is of class C^2 with respect to the second variable and measurable with respect to the first one, $g_j(\cdot,0) \in L^1(\Omega)$, there exist $C>0$, $r<2p/N$, $\psi \in L^{s'}(\Omega)$ for some $s<N/(N-1)$, and $\psi^* \in L^\alpha(\Omega)$ with $\alpha > N/2$, such that

$$\left|\frac{\partial g_j}{\partial \eta}(x,\eta)\right| \le \psi(x) + C|\eta|^{p(s-1)/s}, \quad \left|\frac{\partial^2 g_j}{\partial \eta^2}(x,\eta)\right| \le \psi^*(x) + C|\eta|^r, \text{ for a.e. } x \in \Omega,$$

and finally, for every $M>0$ and $\varepsilon>0$ there exists $\delta=\delta(\varepsilon,M)>0$ such that

$$\left|\frac{\partial^2 g_j}{\partial \eta^2}(x,\eta_2) - \frac{\partial^2 g_j}{\partial \eta^2}(x,\eta_1)\right| \le \varepsilon \text{ if } |\eta_2-\eta_1|<\delta \text{ and } |\eta_1|,|\eta_2| \le M, \text{ for a.e. } x \in \Omega.$$

Once again, **(A3)** is fulfilled for $q=\min\{p, 2N/(N-2)-\beta\}$ for $\beta>0$ small enough. The reader is referred to [5] for the study of this type of constraints.

The solution of (2.1) must be understood in a variational sense. Let us clarify this point. We define the variational form associated to the operator A in the usual way:

$$a(y,z) = \sum_{i,j=1}^N \int_\Omega a_{ij}(x)\partial_{x_i} y(x) \partial_{x_j} z(x)\,dx + \int_\Omega a_0(x) y(x) z(x)\,dx.$$

Then given $1<r<+\infty$, $\hat f \in (W^{1,r'}(\Omega))'$, and $\hat g \in W^{-\frac{1}{r},r}(\Gamma)$, we say that $y \in W^{1,r}(\Omega)$ is a solution of

$$\left\{ \begin{array}{rcll} Ay & = & \hat f & \text{in } \Omega, \\ \partial_{n_A} y & = & \hat g & \text{on } \Gamma, \end{array} \right. \tag{2.2}$$

if

$$a(y,z) = \langle \hat f, z\rangle_{(W^{1,r'}(\Omega))'\times W^{1,r'}(\Omega)} + \langle \hat g, \gamma z\rangle_{W^{-\frac{1}{r},r}(\Gamma)\times W^{\frac{1}{r},r'}(\Gamma)} \quad \forall z \in W^{1,r'}(\Omega),$$

where $\gamma : W^{1,r'}(\Omega) \to W^{\frac{1}{r},r'}(\Gamma)$ is the trace operator. The following known result deals with the solvability of (2.2); see Mateos [17] for the details, as well as Morrey [19] and Troianiello [21].

LEMMA 2.4. *Let $1<r<+\infty$, $\hat f \in (W^{1,r'}(\Omega))'$, and $\hat g \in W^{-\frac{1}{r},r}(\Gamma)$. Then there exists a unique variational solution $y \in W^{1,r}(\Omega)$ of Neumann's problem* (2.2). *Moreover, the following estimate is satisfied:*

$$\|y\|_{W^{1,r}(\Omega)} \le C\left(\|\hat f\|_{(W^{1,r'}(\Omega))'} + \|\hat g\|_{W^{-\frac{1}{r},r}(\Gamma)}\right), \tag{2.3}$$

where $C>0$ is a constant only depending on r, the dimension N, the operator A, and the domain Ω.

The semilinear case is a consequence of the previous lemma. In particular, $y_u \in W^{1,p}(\Omega)$ is said to be a solution of (2.1) if it satisfies the above variational equation with $\hat f = f(\cdot, y_u, u) \in (W^{1,p'}(\Omega))'$ and $\hat g = g \in L^{p(1-1/N)}(\Gamma) \subset W^{-1/p,p}(\Gamma)$. The next theorem states the existence and uniqueness of the solution of (2.1) as well as the differentiability of the relation between the control u and the associated state y_u.

THEOREM 2.5. *Suppose that* **(A1)** *holds. Then for every $u \in L^\infty(\Omega)$ there exists a unique solution $y_u \in W^{1,p}(\Omega)$ of the state equation* (2.1) *and*

$$\forall M>0\ \exists C_M>0 \text{ such that } \|y_u\|_{W^{1,p}(\Omega)} \le C_M \text{ if } \|u\|_{L^\infty(\Omega)} \le M.$$

The mapping $G : L^\infty(\Omega) \longrightarrow W^{1,p}(\Omega)$, defined by $G(u) = y_u$ is of class C^2 and for all $h, u \in L^\infty(\Omega)$, $z_h = G'(u)h$ is defined as the solution of

$$\left\{ \begin{array}{rcll} Az_h & = & \dfrac{\partial f}{\partial y}(x, y_u, u)z_h + \dfrac{\partial f}{\partial u}(x, y_u, u)h & \text{in } \Omega, \\ \partial_{n_A} z_h & = & 0 & \text{on } \Gamma. \end{array} \right. \tag{2.4}$$

Finally, for every $h_1, h_2 \in L^\infty(\Omega)$, $z_{h_1h_2} = G''(u)h_1h_2$ is the solution of

$$\left\{ \begin{array}{rcl} Az_{h_1h_2} = & \dfrac{\partial f}{\partial y}(x, y_u, u)z_{h_1h_2} + \dfrac{\partial^2 f}{\partial y^2}(x, y_u, u)z_{h_1}z_{h_2} & \\ & + \dfrac{\partial^2 f}{\partial u \partial y}(x, y_u, u)(z_{h_1}h_2 + z_{h_2}h_1) + \dfrac{\partial^2 f}{\partial u^2}(x, y_u, u)h_1h_2 & \text{in } \Omega, \\ \partial_{n_A} z_{h_1h_2} = & 0 \quad \text{on } \Gamma. & \end{array} \right. \tag{2.5}$$

Proof. The proof of the existence, uniqueness, and estimate of the solution of (2.1) is standard. Let us prove the differentiability. For that let us start with a homogeneous boundary condition, $g = 0$. We consider the space

$$V(A) = \Big\{ y \in W^{1,p}(\Omega) : \ Ay \in L^{Np/(N+p)}(\Omega), \ \partial_{n_A} y = 0 \Big\}$$

endowed with the norm

$$\|y\|_{V(A)} = \|y\|_{W^{1,p}(\Omega)} + \|Ay\|_{L^{Np/(N+p)}(\Omega)}.$$

Let us now define the function

$$F : V(A) \times L^\infty(\Omega) \to L^{Np/(N+p)}(\Omega), \quad F(y, u) = Ay - f(\cdot, y, u).$$

Thanks to assumption **(A1)**, F is of class C^2. Moreover, from Lemma 2.4 it follows that

$$\frac{\partial F}{\partial y}(y, u) = A - \frac{\partial f}{\partial y}(\cdot, y, u)$$

is an isomorphism from $V(A)$ to $L^{Np/(N+p)}(\Omega)$. Taking into account that $F(x, y, u) = 0$ if and only if $y = G(u)$, we can apply the implicit function theorem (see, for instance, [3]) to deduce that G is of class C^2 and satisfies $F(G(u), u) = 0$. From this identity, (2.4) and (2.5) follow easily.

If $g \neq 0$, then we can write $G(u) = y_u^0 + y_g = G_0(u) + y_g$, with y_u^0 and y_g solutions of the problems

$$\left\{ \begin{array}{rcll} Ay_g & = & 0 & \text{in } \Omega, \\ \partial_{n_A} y_g & = & g & \text{on } \Gamma, \end{array} \right.$$

$$\left\{ \begin{array}{rcll} Ay_u^0 & = & f^0(x, y_u^0, u) & \text{in } \Omega, \\ \partial_{n_A} y_u^0 & = & 0 & \text{on } \Gamma, \end{array} \right.$$

where $f^0(x, y, u) = f(x, y + y_g(x), u)$. From the previous argument we have that G_0 is of class C^2 and consequently G is C^2 too, with $G' = G_0'$ and $G'' = G_0''$, which concludes the proof. ∎

As a consequence of this theorem we will get the differentiability of the functionals J and $G_j = F_j \circ G$ in the next two theorems.

THEOREM 2.6. *Let us suppose that* **(A1)** *and* **(A2)** *hold. Then the functional* $J : L^\infty(\Omega) \to \mathbb{R}$ *is of class* C^2. *Moreover, for every* $u, h, h_1, h_2 \in L^\infty(\Omega)$,

$$J'(u)h = \int_\Omega \left(\frac{\partial L}{\partial u}(x, y_u, u) + \varphi_{0u} \frac{\partial f}{\partial u}(x, y_u, u) \right) h \, dx \tag{2.6}$$

and

$$\begin{aligned} J''(u)h_1h_2 = \int_\Omega \Bigg[& \frac{\partial^2 L}{\partial y^2}(x, y_u, u) z_1 z_2 + \frac{\partial^2 L}{\partial y \partial u}(x, y_u, u)(z_1 h_2 + z_2 h_1) \\ & + \frac{\partial^2 L}{\partial u^2}(x, y_u, u) h_1 h_2 + \varphi_{0u} \left(\frac{\partial^2 f}{\partial y^2}(x, y_u, u) z_1 z_2 \right. \\ & \left. + \frac{\partial^2 f}{\partial y \partial u}(x, y_u, u)(z_1 h_2 + z_2 h_1) + \frac{\partial^2 f}{\partial u^2}(x, y_u, u) h_1 h_2 \right) \Bigg] \, dx, \end{aligned} \tag{2.7}$$

where $y_u = G(u)$, $\varphi_{0u} \in W^{1,p}(\Omega)$ *is the unique solution of the problem*

$$\begin{cases} A^*\varphi &= \dfrac{\partial f}{\partial y}(x, y_u, u)\varphi + \dfrac{\partial L}{\partial y}(x, y_u, u) & \text{in } \Omega, \\ \partial_{n_{A^*}}\varphi &= 0 & \text{on } \Gamma, \end{cases} \tag{2.8}$$

where A^* *is the adjoint operator of* A *and* $z_i = G'(u)h_i$, $i = 1, 2$.

Proof. Let us consider the function $F_0 : C(\bar\Omega) \times L^\infty(\Omega) \to \mathbb{R}$ defined by

$$F_0(y, u) = \int_\Omega L(x, y(x), u(x)) \, dx.$$

Due to the assumptions on L it is straightforward to prove that F_0 is of class C^2. Now, applying the chain rule to $J(u) = F_0(G(u), u)$ and using Theorem 2.5 and the fact that $W^{1,p}(\Omega) \subset C(\bar\Omega)$ for every $p > N$, we get that J is of class C^2 and

$$J'(u)h = \int_\Omega \left(\frac{\partial L}{\partial y}(x, y_u, u) z_h + \frac{\partial L}{\partial u}(x, y_u, u) h \right) dx.$$

Taking φ_{0u} as the solution of (2.8), we deduce (2.6) from previous identity and (2.4). Let us remark that the assumptions on f and L imply the regularity of φ_{0u}. The second derivative can be deduced in a similar way, making use of Theorem 2.5 once more. ∎

THEOREM 2.7. *Let us suppose that* **(A1)** *and* **(A3)** *hold. Then for each* j, *the functional* $G_j = F_j \circ G : L^\infty(\Omega) \to \mathbb{R}$ *is of class* C^2. *Moreover, for every* $u, h, h_1, h_2 \in L^\infty(\Omega)$,

$$G_j'(u)h = \int_\Omega \varphi_{ju} \frac{\partial f}{\partial u}(x, y_u, u) h \, dx \tag{2.9}$$

and

$$\begin{aligned} G_j''(u)h_1h_2 = F_j''(y_u) z_1 z_2 \\ + \int_\Omega \varphi_{ju} \left(\frac{\partial^2 f}{\partial y^2}(x, y_u, u) z_1 z_2 + \frac{\partial^2 f}{\partial y \partial u}(x, y_u, u)(z_1 h_2 + z_2 h_1) + \frac{\partial^2 f}{\partial u^2}(x, y_u, u) h_1 h_2 \right) dx, \end{aligned} \tag{2.10}$$

where $y_u = G(u)$, $\varphi_{ju} \in W^{1,s'}(\Omega)$ *is the unique solution of the problem*

$$\left\{ \begin{array}{rcll} A^*\varphi_{ju} & = & \dfrac{\partial f}{\partial y}(x, y_u, u)\varphi_{ju} + F_j'(y_u) & \text{in } \Omega, \\ \partial_{n_{A^*}}\varphi_{ju} & = & 0 & \text{on } \Gamma, \end{array} \right. \tag{2.11}$$

and $z_i = G'(u)h_i$, $i = 1, 2$.

The proof of this theorem is very similar to that of Theorem 2.6. Nevertheless we have to make a comment about (2.11). From assumption **(A3)** we have that $F'(\bar{y}) \in (W^{1,s}(\Omega))'$; then the boundary problem (2.11) has a unique solution in $W^{1,s'}(\Omega)$ in the variational sense, analogous to that of (2.2); see Lemma 2.4. Finally recall that $s < N/(N-1)$; then $s' > N$ and therefore $\varphi_{ju} \in W^{1,s'}(\Omega) \subset C(\bar{\Omega})$.

3. First and second order optimality conditions in the Lagrangian form. Let us start this section by reformulating problem **(P)** as follows:

$$(\mathbf{P}) \left\{ \begin{array}{l} \text{Minimize } J(u), \\ u_a(x) \le u(x) \le u_b(x) \text{ for a.e. } x \in \Omega, \\ G_j(u) = 0, \ 1 \le j \le n_e, \\ G_j(u) \le 0, \ n_e + 1 \le j \le n_e + n_i, \end{array} \right.$$

where we are using the functions introduced in the previous section $G_j = F_j \circ G$. We now apply the results obtained in [7]. In order to deduce the first and second order optimality conditions of an optimization problem, it is necessary to make a regularity assumption. This is our first goal. Given $\varepsilon > 0$, we denote the set of *ε-inactive constraints* by

$$\Omega_\varepsilon = \{x \in \Omega : u_a(x) + \varepsilon \le \bar{u}(x) \le u_b(x) - \varepsilon\}.$$

We say that a feasible control $\bar{u}$ is regular if the following assumption is fulfilled:

$$\left\{ \begin{array}{l} \exists \varepsilon_{\bar{u}} > 0 \text{ and } \{\bar{h}_j\}_{j \in I_0} \subset L^\infty(\Omega), \text{ with } \operatorname{supp} \bar{h}_j \subset \Omega_{\varepsilon_{\bar{u}}}, \text{ such that} \\ G_i'(\bar{u})\bar{h}_j = \delta_{ij}, \quad i, j \in I_0, \end{array} \right. \tag{3.1}$$

where

$$I_0 = \{j \le m \,|\, G_j(\bar{u}) = 0\}.$$

I_0 is the set of indices corresponding to active constraints. Associated to **(P)** we define the Lagrangian function

$$\mathcal{L}(u, \lambda) = J(u) + \sum_{j=1}^{n_e+n_i} \lambda_j G_j(u).$$

Obviously (3.1) is equivalent to the independence of the derivatives $\{G_j'(\bar{u})\}_{j \in I_0}$ in $L^1(\Omega_{\varepsilon_{\bar{u}}})$. Under this assumption we can derive the first order necessary conditions for optimality in a qualified form. For the proof the reader is referred to Bonnans and Casas [1] or Clarke [11]; see also Mateos [17].

THEOREM 3.1. *Let us assume that* $\bar{u}$ *is a local solution of* **(P)** *and* (3.1) *holds. Then there exist real numbers* $\{\bar{\lambda}_j\}_{j=1}^{n_e+n_i}$ *such that*

$$\bar{\lambda}_j \ge 0, \quad n_e + 1 \le j \le n_e + n_i, \ \bar{\lambda}_j = 0 \text{ if } G_j(\bar{u}) < 0; \tag{3.2}$$

$$\frac{\partial \mathcal{L}}{\partial u}(\bar{u},\bar{\lambda})(u-\bar{u}) \geq 0 \quad \textit{for all } u_a \leq u \leq u_b. \tag{3.3}$$

Denoting by $\bar{\varphi}_0$ and $\bar{\varphi}_j$ the solutions of (2.8) and (2.11) corresponding to $\bar{u}$ and setting

$$\bar{\varphi} = \bar{\varphi}_0 + \sum_{j=1}^{n_e+n_i} \bar{\lambda}_j \bar{\varphi}_j, \tag{3.4}$$

we deduce from Theorems 2.6 and 2.7 and the definition of $\mathcal{L}$ that

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial u}(\bar{u},\bar{\lambda})h &= \int_\Omega \left(\frac{\partial L}{\partial u}(x,\bar{y},\bar{u}) + \bar{\varphi}_0 \frac{\partial f}{\partial u}(x,\bar{y},\bar{u})\right) h\,dx \;+\; \sum_{j=1}^{n_e+n_i} \bar{\lambda}_j \int_\Omega \bar{\varphi}_j \frac{\partial f}{\partial u}(x,\bar{y},\bar{u}) h\,dx \\ &= \int_\Omega \left(\frac{\partial L}{\partial u}(x,\bar{y},\bar{u}) + \bar{\varphi} \frac{\partial f}{\partial u}(x,\bar{y},\bar{u})\right) h\,dx \\ &= \int_\Omega d(x)h(x)\,dx \quad \forall h \in L^\infty(\Omega), \end{aligned} \tag{3.5}$$

where $\bar{y} = G(\bar{u}) = y_{\bar{u}}$ and

$$d(x) = \frac{\partial L}{\partial u}(x,\bar{y}(x),\bar{u}(x)) + \bar{\varphi}(x)\frac{\partial f}{\partial u}(x,\bar{y}(x),\bar{u}(x)) = \frac{\partial H}{\partial u}(x,\bar{y}(x),\bar{u}(x),\bar{\varphi}(x)), \tag{3.6}$$

$H : \Omega \times \mathbb{R}^3 \longrightarrow \mathbb{R}$ being the Hamiltonian associated to the control problem **(P)**,

$$H(x,y,u,\varphi) = L(x,y,u) + \varphi f(x,y,u).$$

From (3.3) we deduce that

$$d(x) = \begin{cases} 0 & \text{for a.e. } x \in \Omega, \text{ where } u_a(x) < \bar{u}(x) < u_b(x), \\ \geq 0 & \text{for a.e. } x \in \Omega, \text{ where } \bar{u}(x) = u_a(x), \\ \leq 0 & \text{for a.e. } x \in \Omega, \text{ where } \bar{u}(x) = u_b(x). \end{cases} \tag{3.7}$$

Remark 3.2. From (3.3), (3.7), and assumption (3.1) we get

$$\int_\Omega \left(\frac{\partial L}{\partial u}(x,\bar{y}(x),\bar{u}(x)) + \bar{\varphi}_0(x)\frac{\partial f}{\partial u}(x,\bar{y}(x),\bar{u}(x))\right)\bar{h}_j(x)\,dx + \bar{\lambda}_j = \frac{\partial \mathcal{L}}{\partial u}(\bar{u},\bar{\lambda})\bar{h}_j = 0,$$

which implies the uniqueness of the Lagrange multipliers provided in Theorem 3.1.

Associated with d we set

$$\Omega^0 = \{x \in \Omega : |d(x)| > 0\}. \tag{3.8}$$

Given $\{\bar{\lambda}_j\}_{j=1}^{n_e+n_i}$ by Theorem 3.1 we define the *cone of critical directions*

$$C_{\bar{u}}^0 = \{h \in L^\infty(\Omega) \text{ satisfying (3.10) and } h(x) = 0 \text{ for a.e. } x \in \Omega^0\}, \tag{3.9}$$

with

$$\begin{cases} G_j'(\bar{u})h = 0 \text{ if } (j \leq n_e) \text{ or } (j > n_e,\ G_j(\bar{u}) = 0, \text{ and } \bar{\lambda}_j > 0); \\ G_j'(\bar{u})h \leq 0 \text{ if } j > n_e,\ G_j(\bar{u}) = 0, \text{ and } \bar{\lambda}_j = 0; \\ h(x) = \begin{cases} \geq 0 & \text{if } \bar{u}(x) = u_a(x); \\ \leq 0 & \text{if } \bar{u}(x) = u_b(x). \end{cases} \end{cases} \tag{3.10}$$

Now we are ready to state the second order necessary optimality conditions.

THEOREM 3.3. *Let us assume that $\bar{u}$ is a local solution of* **(P)**, (3.1) *holds, and* $\{\bar{\lambda}_j\}_{j=1}^{n_e+n_i}$ *are the Lagrange multipliers satisfying* (3.2) *and* (3.3). *Then the following inequality is satisfied:*

$$\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u},\bar{\lambda})h^2 \geq 0 \quad \forall h \in C^0_{\bar{u}}. \tag{3.11}$$

This theorem follows from Theorem 2.2 of [7]. Indeed it is enough to check the assumptions **(A1)** and **(A2)** of such a paper. **(A1)** says that $J'(\bar{u})$ and $G_j'(\bar{u})$ must be continuous functionals on $L^2(\Omega)$, which is an immediate consequence of Theorems 2.6 and 2.7. Assumption **(A2)** of [7] says that

$$\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u},\bar{\lambda})h_k^2 \longrightarrow \frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u},\bar{\lambda})h^2$$

whenever $\{h_k\}_{k=1}^{\infty}$ is bounded in $L^\infty(\Omega)$ and $h_k(x) \to h(x)$ a.e. in Ω. Taking into account that

$$\begin{aligned}\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u},\bar{\lambda})h^2 = &\sum_{j=1}^{n_e+n_i} \bar{\lambda}_j F_j''(\bar{y})z_h^2 \\ &+ \int_\Omega \left(\frac{\partial^2 L}{\partial y^2}(x,\bar{y},\bar{u}) + \bar{\varphi}\frac{\partial^2 f}{\partial y^2}(x,\bar{y},\bar{u})\right) z_h^2\,dx \\ &+ \int_\Omega \left(\frac{\partial^2 L}{\partial y \partial u}(x,\bar{y},\bar{u}) + \bar{\varphi}\frac{\partial^2 f}{\partial y \partial u}(x,\bar{y},\bar{u})\right) z_h h\,dx \\ &+ \int_\Omega \left(\frac{\partial^2 L}{\partial u^2}(x,\bar{y},\bar{u}) + \bar{\varphi}\frac{\partial^2 f}{\partial u^2}(x,\bar{y},\bar{u})\right) h^2\,dx,\end{aligned} \tag{3.12}$$

where z_h is the solution of (2.4) corresponding to the pair $(\bar{y},\bar{u})$, the desired convergence property follows from the boundedness of the second derivatives of L and f along with the convergence $z_{h_k} \to z_h$ in $W^{1,q}(\Omega) \subset L^2(\Omega)$ and our assumption **(A3)**.

In order to obtain the sufficient second order optimality conditions for problem **(P)**, we need to check some additional properties of the first and second derivatives of J and G_j. Let us take a ball in $L^\infty(\Omega)$, $B_\rho(\bar{u})$. From Theorem 2.5, we deduce the existence of a constant $C_\rho > 0$ such that $\{y_u\}_{u\in B_\rho(\bar{u})}$ is uniformly bounded by C_ρ in the $W^{1,p}(\Omega)$ norm and therefore in the $L^\infty(\Omega)$ norm too. This implies the uniform boundedness of the derivatives of f at every point (y_u,u), for $u \in B_\rho(\bar{u})$, as well as the boundedness of the second derivatives of L and the domination of the first derivatives by some functions $\psi_\rho \in L^{Np/(N+p)}(\Omega)$ and $\psi_\rho^* \in L^2(\Omega)$. Then from Lemma 2.4 we deduce that $\{\varphi_{ju}\}_{u\in B_\rho(\bar{u})}$ are bounded in $W^{1,p}(\Omega) \subset L^\infty(\Omega)$ for $j=0$ and $W^{1,s'}(\Omega) \subset L^\infty(\Omega)$ for $1 \leq j \leq n_e+n_i$, respectively. Finally, using Lemma 2.4 once more, we get that

$$\|z_h\|_{W^{1,q}(\Omega)} \leq C\|h\|_{L^2(\Omega)},$$

which follows from the imbedding $L^2(\Omega) \subset (W^{1,q'}(\Omega))'$ due to the fact $q < 2N/(N-2)$. Collecting all these things, we get the existence of constants $M_{j,1}, M_{j,2} > 0$, with

$0 \le j \le n_e + n_i$, such that for every $u \in B_\rho(\bar u)$ and all $h, h_1, h_2 \in L^\infty(\Omega)$ we have

$$\begin{cases} |J'(u)h| \le M_{0,1}\|h\|_{L^2(\Omega)}, \quad |G_j'(u)h| \le M_{j,1}\|h\|_{L^2(\Omega)}, \\ |J''(u)h_1h_2| \le M_{0,2}\|h_1\|_{L^2(\Omega)}\|h_2\|_{L^2(\Omega)}, \\ |G_j''(u)h_1h_2| \le M_{j,2}\|h_1\|_{L^2(\Omega)}\|h_2\|_{L^2(\Omega)}. \end{cases} \tag{3.13}$$

We have to check a last condition, which is established in the following lemma.

LEMMA 3.4. *For every $\delta > 0$ there exists $\varepsilon \in (0, \rho)$ such that for every $h \in L^\infty(\Omega)$ and $\|u - \bar u\|_\infty < \varepsilon$ the following inequality is fulfilled:*

$$\left| \left[\frac{\partial^2 \mathcal{L}}{\partial u^2}(u, \bar\lambda) - \frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar u, \bar\lambda) \right] h^2 \right| \le \delta \, \|h\|^2_{L^2(\Omega)}. \tag{3.14}$$

Proof. Let us take $h \in L^\infty(\Omega)$ and $\delta > 0$. We are going to check that

$$\begin{aligned} &\left| \left[\frac{\partial^2 \mathcal{L}}{\partial u^2}(v, \bar\lambda) - \frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar u, \bar\lambda) \right] h^2 \right| \\ &\le \int_\Omega \left| \frac{\partial^2 L}{\partial u^2}(x, y_v, v) + \varphi_v \frac{\partial^2 f}{\partial u^2}(x, y_v, v) - \frac{\partial^2 L}{\partial u^2}(x, \bar y, \bar u) - \bar\varphi \frac{\partial^2 f}{\partial u^2}(x, \bar y, \bar u) \right| h^2 \, dx \\ &\quad + \int_\Omega \left| \left(\frac{\partial^2 L}{\partial y \partial u}(x, y_v, v) + \varphi_v \frac{\partial^2 f}{\partial y \partial u}(x, y_v, v) \right) z_h \right. \\ &\quad \left. - \left(\frac{\partial^2 L}{\partial y \partial u}(x, \bar y, \bar u) + \bar\varphi \frac{\partial^2 f}{\partial y \partial u}(x, \bar y, \bar u) \right) \bar z_h \right| |h| \, dx \\ &\quad + \int_\Omega \left| \left(\frac{\partial^2 L}{\partial y^2}(x, y_v, v) + \varphi_v \frac{\partial^2 f}{\partial y^2}(x, y_v, v) \right) z_h^2 \right. \\ &\quad \left. - \left(\frac{\partial^2 L}{\partial y^2}(x, \bar y, \bar u) + \bar\varphi \frac{\partial^2 f}{\partial y^2}(x, \bar y, \bar u) \right) \bar z_h^2 \right| dx \\ &\quad + \sum_{j=1}^{n_i + ni} |\bar\lambda_j| \left| F_j''(y_v) z_h^2 - F_j''(\bar y) \bar z_h^2 \right| \le \delta \|h\|^2_{L^2(\Omega)}, \end{aligned} \tag{3.15}$$

supposing $\|v - \bar u\|_{L^\infty(\Omega)} < \varepsilon$ with ε small enough, where

$$\begin{cases} Az_h &= \dfrac{\partial f}{\partial y}(x, y_v, v) z_h + \dfrac{\partial f}{\partial u}(x, y_v, v) h & \text{in } \Omega, \\ \partial_{n_A} z_h &= 0 & \text{on } \Gamma, \end{cases}$$

and

$$\begin{cases} A\bar z_h &= \dfrac{\partial f}{\partial y}(x, \bar y, \bar u) \bar z_h + \dfrac{\partial f}{\partial u}(x, \bar y, \bar u) h & \text{in } \Omega, \\ \partial_{n_A} \bar z_h &= 0 & \text{on } \Gamma. \end{cases}$$

We discuss every term in a separate way. The inequality

$$\left\| \frac{\partial^2 L}{\partial u^2}(x, y_v, v) + \varphi_v \frac{\partial^2 f}{\partial u^2}(x, y_v, v) - \frac{\partial^2 L}{\partial u^2}(x, \bar y, \bar u) - \bar\varphi \frac{\partial^2 f}{\partial u^2}(x, \bar y, \bar u) \right\|_{L^\infty(\Omega)} < \frac{\delta}{4}$$

is a direct consequence of the continuity $v \in L^\infty(\Omega) \longrightarrow \varphi_v \in W^{1,\min\{s',p\}}(\Omega) \subset C(\bar\Omega)$ (see Lemma 2.4 and Theorems 2.6 and 2.7) and the continuity properties of the second derivatives of f and L assumed in **(A1)** and **(A2)**, as well as assumption **(A3)**.

Let us study the second term of (3.15). Hölder's inequality leads us to

$$\begin{aligned}
&\int_\Omega \left| \left(\frac{\partial^2 L}{\partial y \partial u}(x, y_v, v) + \varphi_v \frac{\partial^2 f}{\partial y \partial u}(x, y_v, v) \right) z_h \right. \\
&\qquad \left. - \left(\frac{\partial^2 L}{\partial y \partial u}(x, \bar y, \bar u) + \bar\varphi \frac{\partial^2 f}{\partial y \partial u}(x, \bar y, \bar u) \right) \bar z_h \right| |h| \, dx \\
&\le \|h\|_{L^2(\Omega)} \left(\left\| \frac{\partial^2 L}{\partial y \partial u}(x, y_v, v) - \frac{\partial^2 L}{\partial y \partial u}(x, \bar y, \bar u) \right\|_{L^\infty(\Omega)} \|z_h\|_{L^2(\Omega)} \right. \\
&\qquad + \left\| \frac{\partial^2 L}{\partial y \partial u}(x, \bar y, \bar u) \right\|_{L^\infty(\Omega)} \|z_h - \bar z_h\|_{L^2(\Omega)} \\
&\qquad + \left\| \varphi_v \frac{\partial^2 f}{\partial y \partial u}(x, y_v, v) - \bar\varphi \frac{\partial^2 f}{\partial y \partial u}(x, \bar y, \bar u) \right\|_{L^\infty(\Omega)} \|z_h\|_{L^2(\Omega)} \\
&\qquad \left. + \left\| \bar\varphi \frac{\partial^2 f}{\partial y \partial u}(x, \bar y, \bar u) \right\|_{L^\infty(\Omega)} \|z_h - \bar z_h\|_{L^2(\Omega)} \right) < \frac{\delta}{4} \|h\|^2_{L^2(\Omega)},
\end{aligned}$$

the last inequality being a consequence of **(A1)** and **(A2)** along with the estimates

$$\|z_h\|_{L^2(\Omega)} + \|\bar z_h\|_{L^2(\Omega)} \le C_1 \left(\|z_h\|_{W^{1,q}(\Omega)} + \|\bar z_h\|_{W^{1,q}(\Omega)} \right) \le C_2 \|h\|_{L^2(\Omega)} \tag{3.16}$$

and

$$\|z_h - \bar z_h\|_{L^2(\Omega)} \le C_1 \|z_h - \bar z_h\|_{W^{1,q}(\Omega)} \le O(\varepsilon) \|h\|_{L^2(\Omega)}, \tag{3.17}$$

with $O(\varepsilon) \to 0$ when $\varepsilon \to 0$. Let us notice that (3.16) follows from the inequalities $2N/(N+2) \le q < 2N/(N-2)$, Sobolev imbeddings, and Lemma 2.4.

Analogously we have

$$\begin{aligned}
&\int_\Omega \left| \left(\frac{\partial^2 L}{\partial y^2}(x, y_v, v) + \varphi_v \frac{\partial^2 f}{\partial y^2}(x, y_v, v) \right) z_h^2 - \left(\frac{\partial^2 L}{\partial y^2}(x, \bar y, \bar u) + \bar\varphi \frac{\partial^2 f}{\partial y^2}(x, \bar y, \bar u) \right) \bar z_h^2 \right| dx \\
&\le \left\| \frac{\partial^2 L}{\partial y^2}(x, y_v, v) - \frac{\partial^2 L}{\partial y^2}(x, \bar y, \bar u) \right\|_{L^\infty(\Omega)} \|z_h\|^2_{L^2(\Omega)} \\
&\quad + \left\| \frac{\partial^2 L}{\partial y^2}(x, \bar y, \bar u) \right\|_{L^\infty(\Omega)} \|z_h - \bar z_h\|_{L^2(\Omega)} \|z_h + \bar z_h\|_{L^2(\Omega)} \\
&\quad + \left\| \varphi_v \frac{\partial^2 f}{\partial y^2}(x, y_v, v) - \bar\varphi \frac{\partial^2 f}{\partial y^2}(x, \bar y, \bar u) \right\|_{L^\infty(\Omega)} \|z_h\|^2_{L^2(\Omega)} \\
&\quad + \left\| \bar\varphi \frac{\partial^2 f}{\partial y^2}(x, \bar y, \bar u) \right\|_{L^\infty(\Omega)} \|z_h - \bar z_h\|_{L^2(\Omega)} \|z_h + \bar z_h\|_{L^2(\Omega)} < \frac{\delta}{4} \|h\|^2_{L^2(\Omega)},
\end{aligned}$$

thanks again to **(A1)**, **(A2)**, (3.16), and (3.17).

For the fourth term of (3.15) it is enough to take into account assumption **(A3)**, and once more (3.16) and (3.17), and to use the inequality

$$\begin{aligned}
|F_j''(y_v) z_h^2 - F_j''(\bar y) \bar z_h^2| &= |F_j''(y_v)(z_h^2 - \bar z_h^2) + (F_j''(y_v) - F_j''(\bar y)) \bar z_h^2| \\
&\le |F_j''(y_v)(z_h + \bar z_h)(z_h - \bar z_h)| + |(F_j''(y_v) - F_j''(\bar y)) \bar z_h^2|. \qquad \square
\end{aligned}$$

Before writing the sufficient optimality conditions, we have to fix some notation. Analogously to (3.8) and (3.9), we define for every $\tau > 0$

$$\Omega^\tau = \{x \in \Omega : |d(x)| > \tau\} \tag{3.18}$$

and

$$C_{\bar u}^\tau = \{h \in L^\infty(\Omega) \text{ satisfying (3.10) and } h(x) = 0 \text{ for a.e. } x \in \Omega^\tau\}. \tag{3.19}$$

The next theorem provides the second order sufficient optimality conditions of problem **(P)**.

THEOREM 3.5. *Let $\bar u$ be a feasible point for problem* **(P)** *satisfying* (3.2) *and* (3.3) *and let us suppose that assumption* (3.1) *holds. Let us also assume that*

$$\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar u, \bar\lambda)h^2 \ge \delta \|h\|^2_{L^2(\Omega)} \quad \forall h \in C_{\bar u}^\tau \tag{3.20}$$

for some $\delta > 0$ and $\tau > 0$ given. Then there exist $\varepsilon > 0$ and $\alpha > 0$ such that $J(\bar u) + \alpha\|u - \bar u\|^2_{L^2(\Omega)} \le J(u)$ for every feasible point u for **(P)**, *with $\|u - \bar u\|_{L^\infty(\Omega)} < \varepsilon$.*

Relations (3.13) and (3.14) prove that the hypotheses of Corollary 3.3 of [7] are fulfilled, which leads straightforwardly to the above theorem. In that paper it is also proved that we can not relax the sufficient condition by taking $\tau = 0$; see also Dunn [12].

The last two theorems concerning the necessary and sufficient second order optimality conditions involve two norms: those of $L^2(\Omega)$ and $L^\infty(\Omega)$. This is motivated by the so-called two norms discrepancy; see, for instance; A. Ioffe [15] and H. Maurer [18]. In particular, the cones $C_{\bar u}^\tau$, for $\tau \ge 0$, as defined in (3.8) and (3.19), are subsets of $L^\infty(\Omega)$, but only the $L^2(\Omega)$-norm of the elements of $C_{\bar u}^\tau$ is involved in the optimality conditions (3.11) and (3.20). Now there is a natural question. Let us define for each $\tau \ge 0$

$$C_{\bar u, L^2(\Omega)}^\tau = \{h \in L^2(\Omega) \text{ satisfying (3.10) and } h(x) = 0 \text{ for a.e. } x \in \Omega^\tau\}. \tag{3.21}$$

Can we replace $C_{\bar u}^\tau$ by $C_{\bar u, L^2(\Omega)}^\tau$ in Theorems 3.3 and 3.5? The next proposition provides a positive answer.

PROPOSITION 3.6. *Let us assume that* (3.1) *holds. Then $C_{\bar u, L^2(\Omega)}^\tau = \bar C_{\bar u}^\tau$, where $\bar C_{\bar u}^\tau$ denotes the closure of $C_{\bar u}^\tau$ in $L^2(\Omega)$.*

Proof. Since $C_{\bar u, L^2(\Omega)}^\tau$ is closed in $L^2(\Omega)$, we obviously have $\bar C_{\bar u}^\tau \subset C_{\bar u, L^2(\Omega)}^\tau$. Let us prove the reverse inclusion. Let $h \in C_{\bar u, L^2(\Omega)}^\tau$. We are going to obtain a sequence $\{h_k\}_{k=1}^\infty \subset C_{\bar u}^\tau$ such that $h_k \to h$ in $L^2(\Omega)$. Let us take

$$\hat h_k = \begin{cases} +k & \text{if } h(x) > +k, \\ h(x) & \text{if } |h(x)| \le k, \\ -k & \text{if } h(x) < -k. \end{cases}$$

For every $j \in I_0$ let us set

$$\alpha_{kj} = G_j'(\bar u)\hat h_k - G_j'(\bar u)h.$$

It is clear that $\hat h_k \to h$ in $L^2(\Omega)$ and $\alpha_{kj} \to 0$ for every $j \in I_0$. Finally we define

$$h_k = \hat h_k - \sum_{j \in I_0} \alpha_{kj} \bar h_j,$$

where $\bar{h}_j$ is given in (3.1). It is obvious that $h_k \to h$ in $L^2(\Omega)$ and $\{h_k\}_{k=1}^{\infty} \subset L^\infty(\Omega)$. Let us prove that $h_k \in C^\tau_{\bar{u}}$ for every k. First of all, $h_k(x) = 0$ for almost every $x \in \Omega^\tau$. Indeed, since $h \in C^\tau_{\bar{u},L^2(\Omega)}$, then $h(x) = 0$ for almost every $x \in \Omega^\tau$; consequently $\hat{h}_k$ keeps the same property. On the other hand, the support of $\bar{h}_j$ is in $\Omega_{\varepsilon_{\bar{u}}}$, and $d(x) = 0$ for almost every $x \in \Omega_{\varepsilon_{\bar{u}}}$; therefore $\bar{h}_j(x) = 0$ for almost all $x \in \Omega^\tau$. Hence $h_k(x)$ also vanishes almost everywhere in Ω^τ. Moreover, since h and $\hat{h}_k$ have the same sign, it follows that $h_k(x) = \hat{h}_k(x) \geq 0$ if $\bar{u}(x) = u_a(x)$. Analogously, if $\bar{u}(x) = u_b(x)$, then $h_k(x) = \hat{h}_k(x) \leq 0$. Finally, let us fix $j \in I_0$

$$G'_j(\bar{u})h_k = G'_j(\bar{u})\hat{h}_k - \sum_{i\in I_0} \alpha_{ki} G'_j(\bar{u})\bar{h}_i = G'_j(\bar{u})\hat{h}_k - \alpha_{kj} = G'_j(\bar{u})h.$$

Using the fact that h is in the cone of critical directions of $L^2(\Omega)$, we deduce that h_k satisfies in the same way as h the conditions on the derivatives of G_j, for every j, which proves that $h_k \in C^\tau_{\bar{u}}$. □

Remark 3.7. As a consequence of the previous proposition and the fact that $\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\lambda})$ is a bilinear and continuous form in $L^2(\Omega)$, we get the following equivalences:

$$\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\lambda})h^2 \geq 0 \ \ \forall h \in C^0_{\bar{u}} \iff \frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\lambda})h^2 \geq 0 \ \ \forall h \in C^0_{\bar{u},L^2(\Omega)}$$

and

$$\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\lambda})h^2 \geq \delta\|h\|^2_{L^2(\Omega)} \ \ \forall h \in C^\tau_{\bar{u}} \iff \frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\lambda})h^2 \geq \delta\|h\|^2_{L^2(\Omega)} \ \ \forall h \in C^\tau_{\bar{u},L^2(\Omega)}.$$

4. First and second order optimality conditions involving the Hamiltonian. As in the previous section, we denote with $H : \Omega \times \mathbb{R}^3 \longrightarrow \mathbb{R}$ the Hamiltonian associated to the control problem **(P)**:

$$H(x, y, u, \varphi) = L(x, y, u) + \varphi f(x, y, u).$$

Pontryagin's principle for **(P)** is formulated in terms of H in the next proposition.

PROPOSITION 4.1. *Let $\bar{u}$ be a solution of* (**P**). *Suppose that the assumptions* **(A1)**–**(A3)** *and* (3.1) *hold. Then there exist real numbers $\bar{\lambda}_j$, $j = 1, \ldots, n_i + n_e$, and functions $\bar{y} \in W^{1,p}(\Omega)$, $\bar{\varphi} \in W^{1,\min\{s',p\}}(\Omega)$ such that*

$$\bar{\lambda}_j \geq 0, \quad n_e + 1 \leq j \leq n_e + n_i, \quad \bar{\lambda}_j F_j(\bar{y}) = 0, \tag{4.1}$$

$$\begin{cases} A\bar{y} = f(x, \bar{y}(x), \bar{u}(x)) & \text{in } \Omega, \\ \partial_{n_A}\bar{y} = 0 & \text{on } \Gamma, \end{cases} \tag{4.2}$$

$$\begin{cases} A^*\bar{\varphi} = \dfrac{\partial f}{\partial y}(x, \bar{y}, \bar{u})\bar{\varphi} + \dfrac{\partial L}{\partial y}(x, \bar{y}, \bar{u}) + \displaystyle\sum_{j=1}^{n_e+n_i} \bar{\lambda}_j F'_j(\bar{y}) & \text{in } \Omega, \\ \partial_{n_{A^*}}\bar{\varphi} = 0 & \text{on } \Gamma, \end{cases} \tag{4.3}$$

and for a.e. $x \in \Omega$

$$H(x, \bar{y}(x), \bar{u}(x), \bar{\varphi}(x)) = \min_{k\in[u_a(x),u_b(x)]} H(x, \bar{y}(x), k, \bar{\varphi}(x)). \tag{4.4}$$

Proof. Let us define $H_\nu : \Omega \times \mathbb{R}^3 \longrightarrow \mathbb{R}$ by

$$H_\nu(x,y,u,\varphi) = \nu L(x,y,u) + \varphi f(x,y,u).$$

It is known (see Casas [4], Casas, Raymond, and Zidani [6], Li and Yong [16], or Mateos [17]) that there exist $\bar\nu \ge 0$, $\bar\lambda = (\bar\lambda_j)_{1\le j\le n_i+n_e}$, and functions $\bar y \in W^{1,p}(\Omega)$, $\bar\varphi \in W^{1,\min\{s',p\}}(\Omega)$ such that $(\bar\nu,\bar\lambda) \ne 0$, (4.1) and (4.2) hold, and

$$\text{(4.5)}\quad \begin{cases} A^*\bar\varphi &= \dfrac{\partial f}{\partial y}(x,\bar y,\bar u)\bar\varphi + \bar\nu\dfrac{\partial L}{\partial y}(x,\bar y,\bar u) + \displaystyle\sum_{j=1}^{n_e+n_i}\bar\lambda_j F_j'(\bar y) & \text{in } \Omega,\\ \partial_{n_{A^*}}\bar\varphi &= 0 & \text{on } \Gamma, \end{cases}$$

and

$$\text{(4.6)}\quad H_{\bar\nu}(x,\bar y(x),\bar u(x),\bar\varphi(x)) = \min_{k\in[u_a(x),u_b(x)]} H_{\bar\nu}(x,\bar y(x),k,\bar\varphi(x)) \ \text{ for a.e. } x\in\Omega.$$

In the case $\bar\nu > 0$, we can rename $\bar\lambda/\bar\nu$ by $\bar\lambda$ and obtain (4.1)–(4.5). So it is enough to prove that $\bar\nu \ne 0$. Let us argue by contradiction and let us suppose that $\bar\nu = 0$. Since $H_{\bar\nu}$ is C^1 with respect to $(y,u) \in \mathbb{R}\times\mathbb{R}$, we deduce from (4.6) and Theorem 2.7 that for every $u_a \le u \le u_b$

$$\sum_{j=1}^{n_i+n_e}\bar\lambda_j G_j'(\bar u)(u-\bar u) = \int_\Omega \frac{\partial H_{\bar\nu}}{\partial u}(x,\bar y(x),\bar u(x),\bar\varphi(x))(u(x)-\bar u(x))\,dx \ge 0.$$

Let us take $\bar h_j$ as defined in assumption (3.1) and $|\rho| < \varepsilon$ small enough such that $u_a \le u = \bar u + \rho\bar h_j \le u_b$; then

$$\rho\bar\lambda_j = \sum_{i=1}^{n_i+n_e}\bar\lambda_i G_i'(\bar u)(u-\bar u) \ge 0.$$

By taking ρ positive and negative, respectively, we get that $\bar\lambda_j = 0$ for every $j \in I_0$. So we have the contradiction with the fact that $(\bar\nu,\bar\lambda) \ne 0$. ∎

Let us notice that

$$\text{(4.7)}\quad \begin{cases} d(x) = \dfrac{\partial H}{\partial u}(x,\bar y(x),\bar u(x),\bar\varphi(x)),\\[2ex] \dfrac{\partial\mathcal{L}}{\partial u}(\bar u,\bar\lambda)h = \displaystyle\int_\Omega \frac{\partial H}{\partial u}(x,\bar y(x),\bar u(x),\bar\varphi(x))h(x)\,dx. \end{cases}$$

As an immediate consequence of Pontryagin's principle, Theorem 3.1, and Remark 3.7, we obtain the necessary first and second order optimality conditions as follows.

COROLLARY 4.2. *Suppose that $\bar u$ is a local solution for problem* (**P**). *Suppose also that assumptions* **(A1)**–**(A3)** *and the regularity assumption* (3.1) *hold. Then there exist real numbers $\bar\lambda_j$, $j = 1,\dots,n_i+n_e$, and functions $\bar y \in W^{1,p}(\Omega)$, $\bar\varphi \in W^{1,\min\{s',p\}}(\Omega)$ such that* (4.1)–(4.3) *hold as well as the following relations:*

$$\text{(4.8)}\quad \frac{\partial H}{\partial u}(x,\bar y(x),\bar u(x),\bar\varphi(x))(k-\bar u(x)) \ge 0 \ \textit{for all } u_a(x)\le k\le u_b(x), \ \textit{for a.e. } x\in\Omega,$$

$$\text{(4.9)}\quad \frac{\partial^2\mathcal{L}}{\partial u^2}(\bar u,\bar\lambda)h^2 \ge 0 \ \textit{for all } h\in C^0_{\bar u,L^2(\Omega)},$$

and

$$\frac{\partial^2 H}{\partial u^2}(x,\bar{y}(x),\bar{u}(x),\bar{\varphi}(x)) \geq 0 \text{ for a.e. } x \in \Omega \setminus \Omega^0. \tag{4.10}$$

Let us notice that

$$\frac{\partial H}{\partial u}(x,\bar{y}(x),\bar{u}(x),\bar{\varphi}(x)) = d(x) = 0, \quad x \in \Omega \setminus \Omega^0.$$

Then it is enough to use elementary calculus to deduce (4.10) from (4.4) and the above equality.

In finite dimension, the first order optimality conditions and the strict positivity of the second derivative of the Lagrangian with respect to u on $C^0_{\bar{u}}$ are sufficient conditions for a local minimum. The argument of the proof uses in an essential way the compactness of the balls in finite dimension. To extend this argumentation to infinite-dimensional optimization problems, Bonnans and Zidani [2] made the assumption that the second derivative of the Lagrangian with respect to u was a Legendre form. Let us recall that a quadratic form Q on a Hilbert space X is said to be a Legendre form if it is weakly lower semicontinuous, and for every sequence $\{x_k\} \subset X$ that converges weakly $x_k \rightharpoonup x$ and such that $Q(x_k) \to Q(x)$, we have that $x_k \to x$ strongly. Unfortunately this assumption is not fulfilled, in general, in the context of control theory. We follow a different approach to achieve the same result. Along with the strict positivity of the second derivative of the Lagrangian, we assume that the second derivative of the Hamiltonian with respect to u is strictly positive on $\Omega \setminus \Omega^\tau$, for $\tau > 0$, which is not far from the necessary condition provided in (4.10). More precisely, we have the following result.

THEOREM 4.3. *Let $\bar{u}$ be an admissible control for problem* (**P**) *satisfying* **(A1)**–**(A3)**, *the regularity assumption* (3.1), *and* (4.1)–(4.4) *for some $\bar{\lambda}_j$, $j = 1, \dots, n_i + n_e$. Let us suppose also that there exist $\omega > 0$ and $\tau > 0$ such that*

$$\begin{cases} \dfrac{\partial^2 H}{\partial u^2}(x,\bar{y}(x),\bar{u}(x),\bar{\varphi}(x)) \geq \omega \text{ for a.e. } x \in \Omega \setminus \Omega^\tau, \\ \\ \dfrac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u},\bar{\lambda})h^2 > 0 \text{ for all } h \in C^0_{\bar{u},L^2(\Omega)} \setminus \{0\}. \end{cases} \tag{4.11}$$

Then there exist $\varepsilon > 0$ and $\alpha > 0$ such that $J(\bar{u}) + \alpha\|u - \bar{u}\|^2_{L^2(\Omega)} \leq J(u)$ for all admissible control u with $\|u - \bar{u}\|_{L^\infty(\Omega)} \leq \varepsilon$.

Proof. We will argue by contradiction. The proof is divided into five steps.

(i) *Definition of a sequence $\{h_k\}$ of the unit sphere of $L^2(\Omega)$ converging weakly to h.* Let us suppose that the theorem is false. Then there exists a sequence $\{u_k\}$ of admissible controls with $u_k \to \bar{u}$ in $L^\infty(\Omega)$ such that

$$J(\bar{u}) + \frac{1}{k}\|u_k - \bar{u}\|^2_{L^2(\Omega)} > J(u_k). \tag{4.12}$$

Let us set $\delta_k = \|u_k - \bar{u}\|_{L^2(\Omega)}$ and

$$h_k = \frac{u_k - \bar{u}}{\delta_k}.$$

Since $\|h_k\|_{L^2(\Omega)} = 1$ for every k, there exists a subsequence of $\{h_k\}$, which will be denoted in the same way, and $h \in L^2(\Omega)$ such that $h_k \rightharpoonup h$ weakly in $L^2(\Omega)$.

(ii) $\frac{\partial \mathcal{L}}{\partial u}(\bar{u},\bar{\lambda})h = 0$. Let us denote $y_k = y_{u_k}$. Since u_k is admissible, we have that

$$F_j(y_k) = 0 \ \text{ if } \ 1 \le j \le n_e$$

and

$$F_j(y_k) \le 0 \ \text{ if } \ n_e + 1 \le j \le n_e + n_i.$$

Since $\bar{\lambda}_j \ge 0$ if $n_e + 1 \le j \le n_e + n_i$, we have that

$$\bar{\lambda}_j F_j(y_k) \le 0 \ \text{ for } \ 1 \le j \le n_e + n_i.$$

On the other hand $\bar{\lambda}_j F_j(\bar{y}) = 0$. Hence (4.12) implies

$$\mathcal{L}(\bar{u},\bar{\lambda}) + \frac{1}{k}\|u_k - \bar{u}\|^2_{L^2(\Omega)} > \mathcal{L}(u_k,\bar{\lambda}). \tag{4.13}$$

Moreover, h satisfies the sign condition in (3.10), because every h_k satisfies it, and the set of functions that satisfy the sign condition in (3.10) is convex and closed in $L^2(\Omega)$, and therefore weakly closed. Furthermore

$$\mathcal{L}(u_k,\bar{\lambda}) = \mathcal{L}(\bar{u},\bar{\lambda}) + \delta_k \frac{\partial \mathcal{L}}{\partial u}(v_k,\bar{\lambda})h_k,$$

where v_k is an intermediate point between $\bar{u}$ and u_k. Using (4.13) and that $\delta_k > 0$, we have that

$$\frac{\partial \mathcal{L}}{\partial u}(v_k,\bar{\lambda})h_k < \frac{1}{k\delta_k}\|u_k - \bar{u}\|^2_{L^2(\Omega)} = \frac{1}{k}\|u_k - \bar{u}\|_{L^2(\Omega)}.$$

This expression can be written as follows:

$$\int_\Omega \left(\frac{\partial L}{\partial u}(x, y_{v_k}, v_k) + \varphi_{v_k} \frac{\partial f}{\partial u}(x, y_{v_k}, v_k) \right) h_k dx < \frac{1}{k}\|u_k - \bar{u}\|_{L^2(\Omega)}, \tag{4.14}$$

where y_{v_k} and φ_{v_k} are, respectively, the state and adjoint state associated to v_k. The conditions imposed on F_j and the uniform convergence $v_k \to \bar{u}$ imply the convergences $y_{v_k} \to \bar{y}$ uniformly and $\varphi_{v_k} \to \bar{\varphi}$ in $L^2(\Omega)$. Using **(A1)**, **(A2)**, and the weak convergence $h_k \rightharpoonup h$ in $L^2(\Omega)$, we can pass to the limit in (4.14) and obtain

$$\frac{\partial \mathcal{L}}{\partial u}(\bar{u},\bar{\lambda})h \le 0. \tag{4.15}$$

On the other hand, from (4.7), (4.8), and $h_k = (u_k - \bar{u})/\delta_k$, with $\delta_k > 0$ and $u_a \le u_k \le u_b$, we get

$$\frac{\partial \mathcal{L}}{\partial u}(\bar{u},\bar{\lambda})h_k \ge 0.$$

Taking the limit we obtain

$$\frac{\partial \mathcal{L}}{\partial u}(\bar{u},\bar{\lambda})h \ge 0. \tag{4.16}$$

So (4.15) and (4.16) lead to

$$\frac{\partial \mathcal{L}}{\partial u}(\bar{u},\bar{\lambda})h = 0. \tag{4.17}$$

(iii) $h \in C^0_{\bar{u},L^2(\Omega)}$. First we check that

$$F_j'(\bar{y})z_h = 0 \text{ if } \begin{cases} j \le n_e \\ \text{or} \\ j > n_e,\ F_j(\bar{y}) = 0,\ \bar{\lambda}_j > 0, \end{cases}$$

and

$$F_j'(\bar{y})z_h \le 0 \text{ if } j > n_e,\ F_j(\bar{y}) = 0,\ \bar{\lambda}_j = 0.$$

If $j \le n_e$, then $F_j(y_k) = F_j(y_{\bar{u}+\delta_k h_k}) = 0$ and $F_j(\bar{y}) = 0$. Therefore

$$0 = \frac{F_j(y_{\bar{u}+\delta_k h_k}) - F_j(\bar{y})}{\delta_k},$$

and taking the limit we obtain with the help of assumption **(A3)**

$$F_j'(\bar{y})z_h = 0.$$

If $j > n_e$ and $F_j(\bar{y}) = 0$, we have that $F_j(y_k) = F_j(y_{\bar{u}+\delta_k h_k}) \le 0$. So

$$0 \ge \frac{F_j(y_{\bar{u}+\delta_k h_k}) - F_j(\bar{y})}{\delta_k},$$

and once again taking the limit as before we deduce

$$F_j'(\bar{y})z_h \le 0.$$

Let us see what happens when $\bar{\lambda}_j > 0$. Taking into account (4.12) and that $\delta_k = \|u_k - \bar{u}\|_{L^2(\Omega)}$, we get

$$\frac{\delta_k}{k} \ge \frac{J(u_k) - J(\bar{u})}{\delta_k} = \frac{J(\bar{u} + \delta_k h_k) - J(\bar{u})}{\delta_k}.$$

Since $\delta_k \to 0$, by passing to the limit in this expression, it follows that

$$0 \ge J'(\bar{u})h.$$

Using (4.17) and the expression for the derivative of the Lagrangian, we now have that

$$0 = J'(\bar{u})h + \sum_{j=1}^{n_e+n_i} \bar{\lambda}_j F_j'(\bar{y})z_h.$$

Taking into account that if $j \le n_e$, then we have already proved the equalities $F_j'(\bar{y})z_h = 0$, and that if $F_j(\bar{y}) < 0$, then $\bar{\lambda}_j = 0$, we have that

$$0 = J'(\bar{u})h + \sum_{j \in I_1} \bar{\lambda}_j F_j'(\bar{y})z_h,$$

where

$$I_1 = \{j : n_e < j \le n_e + n_i;\ F_j(\bar{y}) = 0;\ \bar{\lambda}_j > 0\}.$$

So

$$0 \le -J'(\bar u)h = \sum_{j\in I_1} \bar\lambda_j F_j'(\bar y) z_h \le 0.$$

Thus, if $j \in I_1$, then necessarily $F_j'(\bar y)z_h = 0$. To conclude the proof of the inclusion $h \in C^0_{\bar u, L^2(\Omega)}$ it remains to check that $h(x) = 0$ for a.e. $x \in \Omega^0$. As signaled above, h satisfies the sign condition; then we have that $d(x)h(x) \ge 0$ for a.e. $x \in \Omega$; recall (3.7). Therefore

$$\int_\Omega |d(x)h(x)|\,dx = \int_\Omega d(x)h(x)\,dx = \frac{\partial \mathcal{L}}{\partial u}(\bar u, \bar\lambda)h = 0,$$

which implies $h(x) = 0$ in a.e. Ω^0 and $h \in C^0_{\bar u, L^2(\Omega)}$.

(iv) $h = 0$. Due to the assumption of the theorem, we have that

$$\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar u, \bar\lambda)h^2 > 0 \quad \text{if} \quad h \ne 0. \tag{4.18}$$

Let us prove that the reverse inequality is satisfied, which will lead to the identity $h = 0$. By applying the mean value theorem we get

$$\mathcal{L}(u_k, \bar\lambda) = \mathcal{L}(\bar u, \bar\lambda) + \delta_k \frac{\partial \mathcal{L}}{\partial u}(\bar u, \bar\lambda)h_k + \frac{\delta_k^2}{2}\frac{\partial^2 \mathcal{L}}{\partial u^2}(w_k, \bar\lambda)h_k^2, \tag{4.19}$$

where w_k is an intermediate point between u_k and $\bar u$. In order to simplify the expression of the derivatives of $\mathcal{L}$, let us introduce some notation:

$$\bar H_u(x) = \frac{\partial H}{\partial u}(x, \bar y(x), \bar u(x), \bar\varphi(x)),$$

$$\bar H_{uu}(x) = \frac{\partial^2 H}{\partial u^2}(x, \bar y(x), \bar u(x), \bar\varphi(x)).$$

Analogously we define $\bar H_{uy}$ or $\bar H_{yy}$. Inserting this notation into the expressions of the derivatives of $\mathcal{L}$ given in (3.5) and (3.12), we get

$$\delta_k \frac{\partial \mathcal{L}}{\partial u}(\bar u, \bar\lambda)h_k + \frac{\delta_k^2}{2}\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar u, \bar\lambda)h_k^2 = \delta_k \int_\Omega \bar H_u(x)h_k(x)\,dx + \frac{\delta_k^2}{2}\int_\Omega \bar H_{uu}(x)h_k^2(x)\,dx$$

$$+ \frac{\delta_k^2}{2}\left[\int_\Omega \bar H_{yy}(x)z_{h_k}^2(x)\,dx + 2\int_\Omega \bar H_{yu}(x)h_k(x)z_{h_k}(x)\,dx + \sum_{j=1}^{n_e+n_i} \bar\lambda_j F_j''(\bar y)z_{h_k}^2\right].$$

Taking into account that $\bar H_u(x) = d(x) = 0$ in $\Omega \setminus \Omega^0$,

$$A_k = \delta_k \int_\Omega \bar H_u(x)h_k(x)\,dx + \frac{\delta_k^2}{2}\int_\Omega \bar H_{uu}(x)h_k^2(x)\,dx = \delta_k \int_{\Omega^0\setminus\Omega^\tau} \bar H_u(x)h_k(x)\,dx$$

$$+ \delta_k \int_{\Omega^\tau} \bar H_u(x)h_k(x)\,dx + \frac{\delta_k^2}{2}\int_{\Omega^\tau} \bar H_{uu}(x)h_k^2(x)\,dx + \frac{\delta_k^2}{2}\int_{\Omega\setminus\Omega^\tau} \bar H_{uu}(x)h_k^2(x)\,dx.$$

Using now that $\bar{H}_u(x)h_k(x) \geq 0$ for a.e. $x \in \Omega$ and $\bar{H}_u(x) \geq \tau$ for a.e. $x \in \Omega^\tau$, we have that

$$A_k \geq \delta_k \tau \int_{\Omega^\tau} |h_k(x)|\, dx + \frac{\delta_k^2}{2} \int_{\Omega^\tau} \bar{H}_{uu}(x) h_k^2(x)\, dx + \frac{\delta_k^2}{2} \int_{\Omega \setminus \Omega^\tau} \bar{H}_{uu}(x) h_k^2(x)\, dx.$$

Since $\|\delta_k h_k\|_{L^\infty(\Omega)} = \|u_k - \bar{u}\|_{L^\infty(\Omega)} < \varepsilon$, then for a.e. $x \in \Omega$, $\delta_k |h_k(x)| \leq \varepsilon$. Therefore

$$\frac{\delta_k^2 h_k^2(x)}{\varepsilon} \leq \delta_k |h_k(x)|.$$

Hence

$$A_k \geq \frac{\delta_k^2}{2} \int_{\Omega^\tau} \left(\frac{2\tau}{\varepsilon} + \bar{H}_{uu}(x) \right) h_k^2(x)\, dx + \frac{\delta_k^2}{2} \int_{\Omega \setminus \Omega^\tau} \bar{H}_{uu}(x) h_k^2(x)\, dx.$$

Now, from (4.13), (4.19) and taking into account the previous considerations, we have

$$\begin{aligned} \frac{\delta_k^2}{k} &> \delta_k \frac{\partial \mathcal{L}}{\partial u}(\bar{u}, \bar{\lambda}) h_k + \frac{\delta_k^2}{2} \frac{\partial^2 \mathcal{L}}{\partial u^2}(w_k, \bar{\lambda}) h_k^2 \\ &= \delta_k \frac{\partial \mathcal{L}}{\partial u}(\bar{u}, \bar{\lambda}) h_k + \frac{\delta_k^2}{2} \frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\lambda}) h_k^2 + \frac{\delta_k^2}{2} \left[\frac{\partial^2 \mathcal{L}}{\partial u^2}(w_k, \bar{\lambda}) h_k^2 - \frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\lambda}) h_k^2 \right] \\ &\geq \frac{\delta_k^2}{2} \int_{\Omega^\tau} \left(\frac{2\tau}{\varepsilon} + \bar{H}_{uu}(x) \right) h_k^2(x)\, dx + \frac{\delta_k^2}{2} \int_{\Omega \setminus \Omega^\tau} \bar{H}_{uu}(x) h_k^2(x) dx \\ &+ \frac{\delta_k^2}{2} \left[\int_\Omega \bar{H}_{yy}(x) z_{h_k}^2(x)\, dx + 2 \int_\Omega \bar{H}_{yu}(x) h_k(x) z_{h_k}(x)\, dx + \sum_{j=1}^{n_e+n_i} \bar{\lambda}_j F''(\bar{y}) z_{h_k}^2 \right] \\ &+ \frac{\delta_k^2}{2} \left[\frac{\partial^2 \mathcal{L}}{\partial u^2}(w_k, \bar{\lambda}) h_k^2 - \frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\lambda}) h_k^2 \right]. \end{aligned} \tag{4.20}$$

Taking into account the assumptions made on the second derivatives of the functions, there exists a constant $C_H > 0$ such that $\bar{H}_{uu}(x) \geq -C_H$ for a.e. $x \in \Omega$. So, taking ε small enough, we have that

$$\frac{2\tau}{\varepsilon} + \bar{H}_{uu}(x) \geq \frac{2\tau}{\varepsilon} - C_H > 0 \text{ for a.e. } x \in \Omega.$$

Thus

$$\liminf_{k \to \infty} \int_{\Omega^\tau} \left(\frac{2\tau}{\varepsilon} + \bar{H}_{uu}(x) \right) h_k^2(x) dx \geq \int_{\Omega^\tau} \left(\frac{2\tau}{\varepsilon} + \bar{H}_{uu}(x) \right) h^2(x) dx.$$

Moreover, in $\Omega \setminus \Omega^\tau$, $\bar{H}_{uu}(x) > \omega > 0$, and then

$$\liminf_{k \to \infty} \int_{\Omega \setminus \Omega^\tau} \bar{H}_{uu}(x) h_k^2(x)\, dx \geq \int_{\Omega \setminus \Omega^\tau} \bar{H}_{uu}(x) h^2(x)\, dx.$$

Now dividing (4.20) by $\delta_k^2/2$ and using (3.14) and assumption **(A3)**, we can take the lower limit of the resulting expression and obtain

$$\begin{aligned} 0 &\geq \int_{\Omega^\tau} \left(\frac{2\tau}{\varepsilon} + \bar{H}_{uu}(x) \right) h^2(x)\, dx + \int_{\Omega \setminus \Omega^\tau} \bar{H}_{uu}(x) h^2(x)\, dx \\ &+ \int_\Omega \bar{H}_{yy}(x) z_h^2(x)\, dx + 2 \int_\Omega \bar{H}_{yu}(x) h(x) z_h(x)\, dx + \sum_{j=1}^{n_e+n_i} \bar{\lambda}_j F_j''(\bar{y}) z_h^2 \\ &= \frac{2\tau}{\varepsilon} \int_\Omega h^2(x)\, dx + \frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\lambda}) h^2. \end{aligned}$$

Combining this inequality with (4.18) we deduce that $h = 0$.

(v) $h_k \to 0$ *strongly in* $L^2(\Omega)$. We have that $h_k \rightharpoonup h = 0$ weakly in $L^2(\Omega)$, and consequently $z_{h_k} \to 0$ strongly in $W^{1,q}(\Omega)$. Therefore again dividing (4.20) by $\delta_k^2/2$ and using (3.14) we get

$$\begin{aligned}
&\min\left\{\omega, \frac{2\tau}{\varepsilon} - C_H\right\} \limsup_{k\to\infty} \int_\Omega h_k^2(x)dx \\
&\leq \limsup_{k\to\infty} \left\{ \int_{\Omega^\tau} \left(\frac{2\tau}{\varepsilon} + \bar{H}_{uu}(x)\right) h_k^2(x)\,dx + \int_{\Omega\setminus\Omega^\tau} \bar{H}_{uu}(x) h_k^2(x)dx \right\} \\
&\leq \limsup_{k\to\infty} \left\{ \frac{1}{k} - \left[\int_\Omega \bar{H}_{yy}(x) z_{h_k}^2(x)\,dx + 2\int_\Omega \bar{H}_{yu}(x) h_k(x) z_{h_k}(x)\,dx + \sum_{j=1}^{n_e+n_i} \bar{\lambda}_j F''(\bar{y}) z_{h_k}^2 \right] \right. \\
&\qquad \left. - \left[\frac{\partial^2 \mathcal{L}}{\partial u^2}(w_k, \bar{\lambda}) h_k^2 - \frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\lambda}) h_k^2 \right] \right\} = 0.
\end{aligned}$$

Hence

$$\lim_{k\to\infty} \|h_k\|_{L^2(\Omega)} = 0.$$

But $\|h_k\|_{L^2(\Omega)} = 1$ for every k. So we have achieved the contradiction. ∎

The next theorem shows the equivalence of (3.20) and (4.11).

THEOREM 4.4. *Let $\bar{u}$ be an admissible control for problem* (**P**) *that satisfies* **(A1)**–**(A3)**, *the regularity assumption* (3.1), *and* (4.1)–(4.4). *Then the following two statements are equivalent:*

(1) *There exist $\delta > 0$ and $\tau' > 0$ such that*

$$\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\lambda}) h^2 \geq \delta \|h\|_{L^2(\Omega)}^2 \text{ for all } h \in C_{\bar{u}, L^2(\Omega)}^{\tau'}. \tag{4.21}$$

(2) *There exist $\omega > 0$ and $\tau > 0$ such that*

$$\begin{cases} \dfrac{\partial^2 H}{\partial u^2}(x, \bar{y}(x), \bar{u}(x), \bar{\varphi}(x)) \geq \omega \text{ for a.e. } x \in \Omega \setminus \Omega^\tau, \\ \dfrac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\lambda}) h^2 > 0 \text{ for all } h \in C_{\bar{u}, L^2(\Omega)}^0 \setminus \{0\}. \end{cases} \tag{4.22}$$

Proof. (1) $\Longrightarrow$ (2). Since $C_{\bar{u}, L^2(\Omega)}^0 \subset C_{\bar{u}, L^2(\Omega)}^{\tau'}$, the second inequality of (4.22) is an obvious consequence of (4.21). Let us prove the existence of ω and τ satisfying the first inequality of (4.22). Let us take $\alpha > 0$ and $\varepsilon > 0$, as in Theorem 3.5, and consider the problem

$$(P_\alpha) \begin{cases} \text{Minimize } J_\alpha(u) = J(u) - \frac{\alpha}{2}\|u - \bar{u}\|_{L^2(\Omega)}^2, \\ u_a(x) \leq u(x) \leq u_b(x) \text{ for a.e. } x \in \Omega, \\ G_j(u) = 0,\ 1 \leq j \leq n_e, \\ G_j(u) \leq 0,\ n_e + 1 \leq j \leq n_e + n_i. \end{cases}$$

Then for any feasible point u of this problem, with $\|u - \bar{u}\|_\infty < \varepsilon$ and $u \neq \bar{u}$, we have

$$J_\alpha(\bar{u}) = J(\bar{u}) \leq J(u) - \alpha\|u - \bar{u}\|_{L^2(\Omega)}^2 < J(u) - \frac{\alpha}{2}\|u - \bar{u}\|_{L^2(\Omega)}^2 = J_\alpha(u).$$

Then $\bar{u}$ is the unique solution of (P_α) in the $L^\infty(\Omega)$-ball $B_\varepsilon(\bar{u})$. The Hamiltonian for problem (P_α) is

$$H^\alpha(x,y,u,\varphi) = H(x,y,u,\varphi) - \frac{\alpha}{2}(u-\bar{u}(x))^2.$$

Therefore we can apply Corollary 4.2 to (P_α) and deduce, with the notation of the proof of Theorem 4.3, that

$$\bar{H}_{uu}(x) - \alpha = \bar{H}^\alpha_{uu}(x) \geq 0 \quad \text{for a.e. } x \in \Omega \setminus \Omega^0,$$

which implies

$$\bar{H}_{uu}(x) \geq \alpha > 0 \quad \text{for a.e. } x \in \Omega \setminus \Omega^0. \tag{4.23}$$

In the case in which the bound constraints on the control are not active, i.e., $u_a(x) < \bar{u}(x) < u_b(x)$ a.e., then the Lebesgue measure of Ω^0 is zero; hence (4.23) implies the first inequality of (4.22). Let us analyze the case where Ω^0 has a strictly positive Lebesgue measure. We will proceed by contradiction and we assume that there exist no $\omega > 0$ and $\tau > 0$ such that (4.22) is satisfied. Then we define for every $k \geq 1$

$$\hat{h}_k(x) = \begin{cases} +1 & \text{if } |d(x)| \leq 1/k,\ \bar{H}_{uu}(x) < 1/k, \text{ and } \bar{u}(x) = u_a(x), \\ -1 & \text{if } |d(x)| \leq 1/k,\ \bar{H}_{uu}(x) < 1/k, \text{ and } \bar{u}(x) = u_b(x), \\ 0 & \text{otherwise.} \end{cases}$$

Since (4.22) is not satisfied for $\omega = 1/k$ and $\tau = 1/k$, with arbitrarily large k, and the measure of Ω^0 is not zero, we have that $\hat{h}_k \neq 0$. Then we define $\tilde{h}_k = \hat{h}_k / \|\hat{h}_k\|_{L^2(\Omega)}$. Let us prove that $\tilde{h}_k \rightharpoonup 0$ weakly in $L^2(\Omega)$. From (4.23) we deduce that the set

$$B = \{x \in \Omega : |d(x)| = 0 \ \text{ and } \ \bar{H}_{uu}(x) \leq 0\}$$

has zero Lebesgue measure. Therefore

$$\bigcap_{k=1}^{\infty} \mathrm{supp}\{\tilde{h}_k\} \subset B \Rightarrow \mathrm{measure}\left(\bigcap_{k=1}^{\infty} \mathrm{supp}\{\tilde{h}_k\}\right) \leq \mathrm{measure}(B) = 0.$$

Taking into account that $\mathrm{supp}\{\tilde{h}_k\} \subset \mathrm{supp}\{\tilde{h}_{k'}\}$ for every $k > k'$, we deduce that $\tilde{h}_k(x) \to 0$ pointwise a.e. in Ω. On the other hand, $\{\tilde{h}_k\}_{k=1}^\infty$ is bounded in $L^2(\Omega)$; consequently $\tilde{h}_k \rightharpoonup 0$ weakly in $L^2(\Omega)$; see Hewitt and Stromberg [14, p. 207]. Furthermore for $\tau' > 1/k$ we have that $\tilde{h}_k(x) = 0$ for every $x \in \Omega^{\tau'}$ and $\tilde{h}_k$ satisfies the sign condition of (3.10). Let us define a new function $h_k \in C^{\tau'}_{\bar{u},L^2(\Omega)}$ close to $\tilde{h}_k$. Using the functions $\{\bar{h}_j\}_{j\in I_0}$ introduced in (3.1), we set

$$h_k = \tilde{h}_k - \sum_{j\in I_0} \alpha_{kj}\bar{h}_j, \quad \text{with } \ \alpha_{kj} = G'_j(\bar{u})\tilde{h}_k.$$

As in the proof of Proposition 3.6, we deduce that $h_k \in C^{\tau'}_{\bar{u},L^2(\Omega)}$ for every $k > 1/\tau'$. Moreover, since $\tilde{h}_k \rightharpoonup 0$ weakly in $L^2(\Omega)$, we deduce that $\alpha_{kj} \to 0$, and therefore $h_k \rightharpoonup 0$ weakly in $L^2(\Omega)$. On the other hand, since $\mathrm{supp}\{\tilde{h}_k\}$ is included in the set

of points of Ω where the bound constraints on the control are active, which has an empty intersection with the support of each $\bar{h}_j$ $(j \in I_0)$,

$$\|h_k\|_{L^2(\Omega)} = \left\{ \int_{\operatorname{supp}\{\tilde{h}_k\}} h_k^2 \, dx + \int_{\Omega \setminus \operatorname{supp}\{\tilde{h}_k\}} h_k^2 \, dx \right\}^{\frac{1}{2}}$$

$$= \left\{ \int_{\operatorname{supp}\{\tilde{h}_k\}} \tilde{h}_k^2 \, dx + \int_{\Omega \setminus \operatorname{supp}\{\tilde{h}_k\}} \left(\sum_{j \in I_0} \alpha_{kj} \bar{h}_j \right)^2 dx \right\}^{\frac{1}{2}}$$

$$= \left\{ \|\tilde{h}_k\|^2_{L^2(\Omega)} + \| \sum_{j \in I_0} \alpha_{kj} \bar{h}_j \|^2_{L^2(\Omega)} \right\}^{\frac{1}{2}} \geq \left\{ 1 - 2 \sum_{j \in I_0} \alpha_{kj}^2 \|\bar{h}_j\|^2_{L^2(\Omega)} \right\}^{\frac{1}{2}} \overset{k \to \infty}{\longrightarrow} 1.$$

From this relation and (4.21) with $h = h_k$, we get

$$\delta \leq \delta \liminf_{k \to \infty} \|h_k\|^2_{L^2(\Omega)} \leq \liminf_{k \to \infty} \frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\lambda}) h_k^2. \tag{4.24}$$

On the other hand, the weak convergence $h_k \rightharpoonup 0$ in $L^2(\Omega)$ implies the strong convergence in $(W^{1,q'}(\Omega))'$, and thanks to Lemma 2.4 we deduce $z_{h_k} \to 0$ in $W^{1,q}(\Omega) \subset L^2(\Omega)$ strongly. Writing the second derivative of the Lagrangian in terms of the derivatives of the Hamiltonian, as was done in the proof of Theorem 4.3, and taking into account that $\bar{H}_{uu}(x) < 1/k$ in the support of $\tilde{h}_k$ and $\alpha_{kj} \to 0$, we get

$$\liminf_{k \to \infty} \frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\lambda}) h_k^2 \leq \limsup_{k \to \infty} \int_\Omega \bar{H}_{uu}(x) h_k^2(x) \, dx + \limsup_{k \to \infty} \int_\Omega \bar{H}_{yy}(x) z_{h_k}^2(x) \, dx$$

$$+ 2 \limsup_{k \to \infty} \int_\Omega \bar{H}_{yu}(x) h_k(x) z_{h_k}(x) \, dx + \limsup_{k \to \infty} \left(\sum_{j=1}^{n_e + n_i} \bar{\lambda}_j F_j''(\bar{y}) z_{h_k}^2 \right)$$

$$\leq \limsup_{k \to \infty} \int_{\operatorname{supp}\{\tilde{h}_k\}} \bar{H}_{uu}(x) h_k^2(x) \, dx + \limsup_{k \to \infty} \int_{\Omega \setminus \operatorname{supp}\{\tilde{h}_k\}} \bar{H}_{uu}(x) h_k^2(x) \, dx$$

$$\leq \limsup_{k \to \infty} \frac{1}{k} \int_{\operatorname{supp}\{\tilde{h}_k\}} \tilde{h}_k^2(x) \, dx + \limsup_{k \to \infty} \int_{\Omega \setminus \operatorname{supp}\{\tilde{h}_k\}} \bar{H}_{uu}(x) \left[\sum_{j \in I_0} \alpha_{kj} \bar{h}_j(x) \right]^2 dx$$

$$= \limsup_{k \to \infty} \frac{1}{k} \int_\Omega \tilde{h}_k^2(x) \, dx = \lim_{k \to \infty} \frac{1}{k} = 0,$$

which contradicts (4.24).

(2) $\Longrightarrow$ (1). Let us suppose that (4.21) is not satisfied. Then for every $\tau' > 0$ there exists $h_{\tau'} \in C^{\tau'}_{\bar{u}, L^2(\Omega)}$ such that $\|h_{\tau'}\|_{L^2(\Omega)} = 1$ and

$$\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\lambda}) h_{\tau'}^2 < \tau'.$$

Since $\{h_{\tau'}\}$ is bounded in $L^2(\Omega)$, there exists a subsequence, denoted in the same way, such that $h_{\tau'} \rightharpoonup h$ weakly in $L^2(\Omega)$. We have that $h \in C^0_{\bar{u}, L^2(\Omega)}$. Indeed relations (3.10) are obtained for h by passing to the limit in the corresponding ones satisfied

by $h_{\tau'}$. Let us see that $h(x) = 0$ in Ω^0:

$$\int_\Omega |h(x)||d(x)|\,dx = \int_\Omega h(x)d(x)\,dx = \lim_{\tau'\to 0}\int_\Omega h_{\tau'}(x)d(x)\,dx$$
$$= \lim_{\tau'\to 0}\int_{\Omega\setminus\Omega^{\tau'}} |h_{\tau'}(x)||d(x)|\,dx$$
$$\le \lim_{\tau'\to 0}\tau'\int_\Omega |h_{\tau'}(x)|\,dx \le \lim_{\tau'\to 0}\tau'\sqrt{m(\Omega)}\|h_{\tau'}\|_{L^2(\Omega)} = 0;$$

hence $h(x)d(x) = 0$, and therefore $h(x) = 0$ for a.e. $x \in \Omega^0$.

Since $\bar H_{uu}(x) \ge \omega > 0$ in $\Omega \setminus \Omega^\tau$, $(\Omega \setminus \Omega^{\tau'}) \subset (\Omega \setminus \Omega^\tau)$ for $\tau' < \tau$, and $h_{\tau'} = 0$ in $\Omega_{\tau'}$, we have that

$$\liminf_{k\to\infty}\int_\Omega \bar H_{uu}(x)h^2_{\tau'}(x)dx = \liminf_{k\to\infty}\int_{\Omega\setminus\Omega^{\tau'}} \bar H_{uu}(x)h^2_{\tau'}(x)dx$$
$$\ge \int_{\Omega\setminus\Omega^{\tau'}} \bar H_{uu}(x)h^2(x)dx = \int_\Omega \bar H_{uu}(x)h^2(x)dx.$$

Therefore, using the definition of $h_{\tau'}$ along with the strong convergence $z_{h_{\tau'}} \to z_h$ in $W^{1,q}(\Omega)$, we get

$$0 \ge \limsup_{\tau'\to 0}\frac{\partial^2\mathcal{L}}{\partial u^2}(\bar u,\bar\lambda)h^2_{\tau'} \ge \liminf_{\tau'\to 0}\frac{\partial^2\mathcal{L}}{\partial u^2}(\bar u,\bar\lambda)h^2_{\tau'}$$
$$= \liminf_{\tau'\to 0}\left\{\int_\Omega \bar H_{uu}(x)h^2_{\tau'}(x)\,dx + \int_\Omega \bar H_{yy}(x)z^2_{h_{\tau'}}(x)\,dx\right.$$
$$\left. + 2\int_\Omega \bar H_{yu}(x)h_{\tau'}(x)z_{h_{\tau'}}(x)\,dx + \sum_{j=1}^{n_e+n_i}\bar\lambda_j F_j''(\bar y)z^2_{h_{\tau'}}\right\} \ge \frac{\partial^2\mathcal{L}}{\partial u^2}(\bar u,\bar\lambda)h^2,$$

which, together with (4.22), implies that $h = 0$. Finally, using the weak convergence $h_{\tau'} \rightharpoonup 0$ in $L^2(\Omega)$ and the strong convergence $z_{h_{\tau'}} \to 0$ in $W^{1,q}(\Omega)$, we conclude that

$$\omega = \omega\limsup_{\tau'\to 0}\|h_{\tau'}\|^2_{L^2(\Omega)} \le \limsup_{\tau'\to 0}\int_\Omega \bar H_{uu}(x)h^2_{\tau'}(x)dx$$
$$\le \limsup_{\tau'\to 0}\left\{\frac{\partial^2\mathcal{L}}{\partial u^2}(\bar u,\bar\lambda)h^2_{\tau'}dx - \int_\Omega \bar H_{yy}(x)z^2_{h_{\tau'}}(x)\,dx\right.$$
$$\left. - 2\int_\Omega \bar H_{yu}(x)h_{\tau'}(x)z_{h_{\tau'}}(x)\,dx - \sum_{j=1}^{n_e+n_i}\bar\lambda_j F_j''(\bar y)z^2_{h_{\tau'}}\right\} \le 0,$$

and we have a contradiction. ∎

REFERENCES

[1] J. Bonnans and E. Casas, *Contrôle de systèmes elliptiques semilinéaires comportant des contraintes sur l'état*, in Nonlinear Partial Differential Equations and Their Applications, vol. 8, Collège de France Seminar, H. Brezis and J. Lions, eds., Longman Scientific and Technical, New York, 1988, pp. 69–86.

[2] J. F. Bonnans and H. Zidani, *Optimal control problems with partially polyhedric constraints*, SIAM J. Control Optim., 37 (1999), pp. 1726–1741.
[3] H. Cartan, *Calcul Différentiel*, Hermann, Paris, 1967.
[4] E. Casas, *Pontryagin's principle for optimal control problems governed by semilinear elliptic equations*, in International Conference on Control and Estimation of Distributed Parameter Systems: Nonlinear Phenomena, vol. 118, Internat. Ser. Numer. Math. 118, Birkhäuser, Basel, 1994, pp. 97–114.
[5] E. Casas, M. Mateos, and L. Fernández, *Second-order optimality conditions for semilinear elliptic control problems with constraints on the gradient of the state*, Control Cybernet., 28 (1999), pp. 463–479.
[6] E. Casas, J.-P. Raymond, and H. Zidani, *Pontryagin's principle for local solutions of control problems with mixed control-state constraints*, SIAM J. Control Optim., 39 (2000), pp. 1182–1203.
[7] E. Casas and F. Tröltzsch, *Second order necessary and sufficient optimality conditions for optimization problems and applications to control theory*, SIAM J. Optim., to appear.
[8] E. Casas and F. Tröltzsch, *Second order necessary optimality conditions for some state-constrained control problems of semilinear elliptic equations*, Appl. Math. Optim., 39 (1999), pp. 211–227.
[9] E. Casas, F. Tröltzsch, and A. Unger, *Second order sufficient optimality conditions for a nonlinear elliptic control problem*, Z. Anal. Anwendungen, 15 (1996), pp. 687–707.
[10] E. Casas, F. Tröltzsch, and A. Unger, *Second order sufficient optimality conditions for some state-constrained control problems of semilinear elliptic equations*, SIAM J. Control Optim., 38 (2000), pp. 1369–1391.
[11] F. Clarke, *A new approach to Lagrange multipliers*, Math. Oper. Res., 1 (1976), pp. 165–174.
[12] J. C. Dunn, *On second order sufficient optimality conditions for structured nonlinear programs in infinite-dimensional function spaces*, in Mathematical Programming with Data Perturbations, Lecture Notes in Pure and Appl. Math. 195, Marcel Dekker, New York, 1998, pp. 83–107.
[13] H. Goldberg and F. Tröltzsch, *Second-order sufficient optimality conditions for a class of nonlinear parabolic boundary control problems*, SIAM J. Control Optim., 31 (1993), pp. 1007–1025.
[14] E. Hewitt and K. Stromberg, *Real and Abstract Analysis*, Springer-Verlag, Berlin, Heidelberg, New York, 1965.
[15] A. D. Ioffe, *Necessary and sufficient conditions for a local minimum* 3*: Second order conditions and augmented duality*, SIAM J. Control Optim., 17 (1979), pp. 266–288.
[16] X. Li and J. Yong, *Optimal Control Theory for Infinite Dimensional Systems*, Birkhäuser Boston, Boston, 1995.
[17] M. Mateos, *Problemas de control óptimo gobernados por ecuaciones semilineales con restricciones de tipo integral sobre el gradiente del estado*, Ph.D. thesis, University of Cantabria, Santander, Spain, 2000.
[18] H. Maurer, *First and second order sufficient optimality conditions in mathematical programming and optimal control*, Math. Programming Study, 14 (1981), pp. 163–177.
[19] C. B. Morrey, Jr., *Multiple integrals in the calculus of variations*, Springer-Verlag, New York, 1966.
[20] J. Raymond and F. Tröltzsch, *Second order sufficient optimality conditions for nonlinear parabolic control problems with state-constraints*, Discrete Contin. Dynam. Systems, 6 (2000), pp. 431–450.
[21] G. M. Troianiello, *Elliptic Differential Equations and Obstacle Problems*, Plenum Press, New York , 1987.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 5, pp. 1455–1472

OPTIMAL CONTROL OF THE SOLID FUEL IGNITION MODEL WITH H^1-COST*

KAZUFUMI ITO† AND KARL KUNISCH‡

Abstract. Optimal control problems for the stationary as well as the time-dependent solid fuel ignition model are investigated. Existence of optimal controls is proved, and optimality systems are derived. The analysis is based on a closedness lemma for the exponential function in L^1 and a generalization of Aubin's lemma.

Key words. optimal control, control of exponential nonlinearity, explosion phenomena, optimality conditions

AMS subject classifications. 49B22, 35K55

PII. S0363012900366042

1. Introduction.

This paper is concerned with optimal control problems for systems governed by

$$y_t = \Delta y + \delta e^y + u \tag{1.1}$$

on the time-space domain $(0,T] \times \Omega$, together with appropriate initial and boundary conditions. Here δ is a fixed positive constant, and u denotes the control. System (1.1) arises, for example, in the theory of combustion, where it is referred to as the solid fuel ignition model [BE]. More precisely, combustion models describe rapid exothermic chemical reactions in combustible materials. The modeling process is based upon conservation of mass, species, momentum, and energy. Due to the Arrhenius law relating the production rate of reactants to their concentration and to temperature, exponential terms enter these models. Even for single one-step irreversible reactions involving fuel and an oxidant, the resulting model is a complicated system of partial differential equations for the variables temperature, density, pressure, and oxidant mass fraction; see [BE, p. 6]. The model can be significantly simplified if, instead of considering a compressible gas, one focuses on the thermal reaction formulated for a nondeformable material of constant density. Expressed in dimensionless form, this results in the following system of equations describing the combustion of a solid fuel [BE]:

$$\begin{cases} T_t = \Delta T + \epsilon\delta v^m \exp\left(\frac{T-1}{\epsilon T}\right), \\ v_t = \beta\Delta v - \epsilon\delta\Gamma v^m \exp\left(\frac{T-1}{\epsilon T}\right), \end{cases} \tag{1.2}$$

with initial and boundary conditions given, e.g., by

$$\begin{cases} T(0,\cdot) = T_0,\ v(0,\cdot) = v_0 \ \text{ in } \ \Omega, \\ T(t,x) = 1,\ \frac{\partial v}{\partial n}(t,x) = 0 \ \text{ on } \ (0,\infty) \times \partial\Omega, \end{cases} \tag{1.3}$$

*Received by the editors January 11, 2000; accepted for publication (in revised form) July 7, 2001; published electronically January 18, 2002.
http://www.siam.org/journals/sicon/40-5/36604.html
†Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205 (kito@eos.ncsu.edu).
‡Institut für Mathematik, Karl-Franzens-Universität Graz, A-8010 Graz, Austria (karl.kunisch@uni-graz.at). The research of this author was supported by the Fonds zur Förderung der wissenschaftlichen Forschung under SFB 03 "Optimierung und Kontrolle."

where T and v denote the temperature and oxidant mass fraction, respectively. Further, β, Γ, and δ are nonnegative constants, m is the mass of a single molecule, $\frac{\partial v}{\partial n}$ stands for the outer normal to $\partial\Omega$, and $\epsilon = \frac{RT_0}{E}$, with R the gas constant, E the total energy of the system, and T_0 the temperature corresponding to an equilibrium state. Under appropriate conditions (see [BE, BK]), a first-order approximation with respect to changes in T and v is justified. We set $T = 1 + \epsilon\, y$ and $v = 1 - \epsilon\, c$, where y and c are two auxiliary quantities expressing temperature and oxidant mass fraction changes. If, further, $\epsilon \ll 1$, the equations in y and c decouple and the equation for y is obtained:

$$\begin{cases} y_t = \Delta y + \delta \exp y, \\ y(0,\cdot) = y_0 \ \text{ in } \ \Omega, \quad y(t,x) = 0 \ \text{ on } \ (0,\infty) \times \partial\Omega. \end{cases} \tag{1.4}$$

This is the uncontrolled solid fuel ignition model. In (1.1) we added a source term to describe the control mechanism. The stationary version of (1.4) is called the steady state solid fuel ignition model [BE] and is given by

$$\begin{cases} \Delta y + \delta \exp y = 0 \ \text{ in } \Omega, \\ y = 0 \ \text{ on } \partial\Omega. \end{cases} \tag{1.5}$$

Let us mention that (1.4)–(1.5) also play an important role in the study of stellar structures; we refer to [FK] and the references given there.

An additional motivation for the study of optimal control problems involving (1.1) is the recent development of diverse numerical methods for solving open loop optimal control problems for nonlinear partial differential equations. Due to the strong nonlinearity in (1.1) it serves as an excellent test example. The choice of a numerical method is guided by, among other considerations, a proper function space framework for the nonlinear optimal control problem. For (1.1) this choice is not evident. For numerical results concerning optimal control of (1.1) and its stationary version we refer to [BK, GH, KK].

Problems of existence and uniqueness for (1.4) as well as the qualitative behavior of the solutions to (1.1) are rather well understood; see [BE, F] and the references given there. Depending on δ and the dimension of Ω, the solutions to (1.4) typically exist only locally and exhibit finite time blow-up at a distinct point within the spatial domain Ω. The stationary uncontrolled equation (1.5) is also well investigated. To give the reader a taste of the multitude of properties linked with (1.5), we recall a specific result for the case in which Ω is the unit ball in $\mathbb{R}^2$: There exists a real number $\delta^* > 0$ (see [BE, F]) such that (i) if $\delta > \delta^*$, then (1.5) admits no solution; (ii) if $\delta \in (0, \delta^*)$, then there exist exactly two solutions; (iii) if $\delta = \delta^*$, then there exists exactly one solution to (1.5). Related results also hold in general domains.

Optimal control of equations with the property that they admit multiple states in the stationary case and finite blow-up in the evolutionary case are called singular control problems in [L]. Thus studying optimal control for the control system (1.1) means analyzing a singular control problem with the most severe nonlinearity. The techniques developed in [L] are not applicable to the optimal control problems of this paper. In fact most of the work in [L] is concerned with nonlinearities of the power-law type ($\pm y^3$). In this case, existence of optimal controls and optimality systems can be derived if the cost functionals are coercive (radially unbounded) in appropriate L^p-norms, with $p > 2$. This would suggest using cost functionals that are coercive with respect to the L^∞-norm. Such a choice would, however, be neither practical from the point of view of discretization and numerical realization nor mathematically

appealing. We shall demonstrate that it is possible to obtain the existence of optimal controls as well as to derive optimality conditions for the stationary and for the evolutionary case for cost functionals which are coercive in a Hilbert-space setting involving H^1-norms. Let us also point out that optimal control for (1.1) was posed as an open problem in [L, problem 25, p. 501]. The difficulty of studying just the existence problem for optimal control of (1.1) arises from the fact that a priori bounds on the control u do not imply appropriate bounds on the state y. Even if bounds on both u and y are implied by the choice of the cost functional, it is difficult to pass to the limit on minimizing subsequences in (1.1), which appears as a constraint in the optimal control problem. It is worth observing that the introduction of controls into (1.4)–(1.5) gives additional freedom since it allows the existence of control-state pairs satisfying (1.4) (resp., (1.5)) for δ values for which the uncontrolled equation does not admit a solution. Concerning optimal control of (1.1), the dynamical system behavior of (1.4) suggests that we distinguish two classes of problems. The first class of problems excludes the situation in which blow-up actually occurs. Control can be used to avoid blow-up entirely or to steer close to blow-up, but actual blow-up is avoided. Typical optimal control formulations result in tracking-type problems. The second conceivable class of control problems is the one which allows blow-up and uses control to influence blow-up time or location. No research effort so far has focused on the second class of problems.

Thus, optimal control for (1.1) is a rich subject for a well-motivated class of problems. Nevertheless only little attention has been paid to it so far. We are only aware of the two contributions [CKP] and [KK]. To cope with the exponential nonlinearity in (1.1), the cost functionals in [KK] are chosen such that they contain an exponential term. In the present paper we do not utilize such a technique. The formulation in [CKP] uses an implicit constraint on the state variable which allows the determination of an a priori bound on $\exp(y)$ for admissible (y, u) pairs if $n \geq 3$. The analysis in [CKP] is restricted to the stationary case with distributed controls. Let us also mention [GH], in which optimal control techniques based on a linearization of (1.1) are studied numerically. Newton methods for optimal control of (1.1) are treated in [IK].

The present paper develops a different framework for optimal control of (1.1), which we believe to be more practical than those in [CKP] and [KK], and which avoids linearization as used in [GH]. It is based on cost functionals that are coercive with respect to the $L^2((0,T)\times\Omega)$-norm for the control and the $L^2(0,T;H_0^1(\Omega))$-norm for the state. The impossibility of obtaining a priori estimates on y in terms of u in (1.1) requires us to employ several nonstandard methods to guarantee existence for the optimal control problems that we formulate. In the stationary case the essential technical tool consists of a closedness lemma of the exponential function in L^1-spaces; for the evolutionary case a generalization of Aubin's lemma is developed. To derive optimality systems we use a perturbation argument as well as the Zowe–Kurcyusz Lagrangian theory. Both distributed as well as boundary control problems are considered. Section 2 is devoted to optimal control problems for the time-independent version of (1.1). Section 3 focuses on optimal control for (1.1). The analysis of multigrid methods applied to the optimality systems obtained in this paper and a comparison to the approach taken in [KK] is given in [BK].

2. Stationary optimal control problems. This section is devoted to the stationary control problem

$$\min\ J(y,u) \tag{2.1}$$

subject to $(y,u) \in H_0^1(\Omega) \times L^2(\Omega)$ and

$$\left\{ \begin{array}{l} -\Delta y - \delta e^y = u \ \text{ in } \Omega, \\ y = 0 \ \text{ on } \partial\Omega, \end{array} \right. \tag{2.2}$$

with Ω a bounded domain in $\mathbb{R}^n$ with smooth boundary $\partial\Omega$ and $\delta > 0$.

It is assumed that

$$\left\{ \begin{array}{l} J : H^1(\Omega) \times L^2(\Omega) \to \mathbb{R} \text{ is bounded from below, weakly} \\ \text{ lower semicontinuous, and coercive, in the sense that} \\ J(y,u) \to \infty \ \text{ if } \ |y|_{H_0^1(\Omega)} \to \infty \text{ or } |u|_{L^2(\Omega)} \to \infty. \end{array} \right. \tag{A1}$$

A typical example for a cost functional satisfying (A1) is given by tracking functionals of the form

$$J(y,u) = \frac{1}{2} \int_\Omega |\nabla(y-z)|^2 dx + \frac{\alpha}{2} \int_\Omega |u|^2 dx,$$

with $\alpha > 0$ and $z \in H^1(\Omega)$.

The notion of a solution to (2.2) is defined next. Let $\mathcal{D}(\Omega)$ denote the space of functions in $C^\infty(\Omega)$ with compact support in Ω, endowed with the usual topology, and let $\mathcal{D}'(\Omega)$ be its topological dual. The duality pairing between $\mathcal{D}(\Omega)$ and $\mathcal{D}'(\Omega)$ is denoted by $(\cdot,\cdot)_{\mathcal{D}',\mathcal{D}}$.

DEFINITION 2.1. *For $u \in L^2(\Omega)$ a function $y \in H_0^1(\Omega)$ is called a solution to equation* (2.2) *if $e^y \in \mathcal{D}'(\Omega)$ and*

$$(\nabla y, \nabla v)_{L_n^2} - \delta(e^y, v)_{\mathcal{D}',\mathcal{D}} = (u,v) \ \text{ for all } v \in \mathcal{D}(\Omega). \tag{2.3}$$

Note that if y is a solution to (2.2), then necessarily

$$e^y \in H^{-1}(\Omega) \cap L^1(\Omega).$$

In fact $e^y \in H^{-1}(\Omega)$ follows from (2.3) and the density of $\mathcal{D}(\Omega)$ in $H_0^1(\Omega)$. Moreover, e^y is measurable and nonnegative so that $e^y \in L^1(\Omega)$, provided that $\int_\Omega e^y dx < \infty$. Set $\Omega_1 = \{x \in \Omega : y(x) \le 1\}$ and observe that from (2.3) with $v = |y|$ we have $\int_\Omega |y e^y| dx < \infty$. It follows that

$$\int_\Omega e^y dx \le \int_{\Omega_1} e^y dx + \int_{\Omega \setminus \Omega_1} \frac{1}{y} y e^y dx \le \text{ meas } \Omega + \int_\Omega |y e^y| dx < \infty,$$

and hence $e^y \in L^1(\Omega)$.

For the proof of existence of a solution to (2.1)–(2.2) we require the following lemma on maximal monotone operators in $L^1(\Omega)$ (see [Br, p. 126]).

LEMMA 2.2. *Let γ be a maximal monotone graph in $\mathbb{R} \times \mathbb{R}$, and let f_n and v_n be measurable functions from Ω to $\mathbb{R}$. Assume that $v_n \to v$ a.e. on Ω, and $f_n \to f$ weakly in $L^1(\Omega)$. If $f_n(x) \in \gamma(v_n(x))$ a.e. on Ω, then $f(x) \in \gamma(v(x))$ a.e. on Ω.*

THEOREM 2.3. *If* (A1) *holds, then* (2.1)–(2.2) *admits a solution $(y^*, u^*) \in H_0^1(\Omega) \times L^2(\Omega)$ with $e^{y^*} \in H^{-1}(\Omega) \cap L^1(\Omega)$.*

Proof. There exists a pair $(y,u) \in H_0^1(\Omega) \times L^2(\Omega)$ satisfying (2.2), and hence the feasible set for (2.1)–(2.2) is nonempty. Since by (A1) the functional J is bounded from below, there exists a minimizing sequence $(y_n, u_n) \in H_0^1(\Omega) \times L^2(\Omega)$ to (2.1)–(2.2). Due to the coercivity of J, it follows that $\{(y_n,u_n)\}_{n=1}^\infty$ is bounded in $H_0^1(\Omega) \times L^2(\Omega)$.

Since y_n is a solution to (2.2), with u replaced by u_n for every n, it follows that $\{e^{y_n}\}_{n=1}^{\infty}$ is bounded in $H^{-1}(\Omega)$. Hence there exist a subsequence of $\{(y_n, u_n, e^{y_n})\}$, denoted by the same symbol, and $(y^*, u^*, w^*) \in H_0^1(\Omega) \times L^2(\Omega) \times H^{-1}(\Omega)$ such that $y_n \rightharpoonup y^*$ in $H_0^1(\Omega), u_n^* \rightharpoonup u^*$ in $L^2(\Omega), y_n \to y^*$ a.e. in Ω, and $e^{y_n} \rightharpoonup w$ in H^{-1}. Due to the weak lower semicontinuity of J we have

$$J(y^*, u^*) \le \liminf_{n\to\infty} J(y_n, u_n),$$

and hence the proof will be complete if we show that y^* is a solution to (2.2) with $u = u^*$.

For this purpose we show that $e^{y_n} \rightharpoonup w^*$ weakly in $L^1(\Omega)$ by utilizing the Dunford–Pettis theorem. This requires us to show that the integrals $\int_\Omega |e^{y_n}| dx$ are uniformly absolutely convergent. Due to the boundedness of $\{(y_n, u_n)\}_{n=1}^{\infty}$ in $H_0^1(\Omega) \times L^2(\Omega)$ and (2.2) there exists C such that $\int_\Omega e^{y_n} y_n dx \le C$ for all n. Let $\epsilon > 0$ be arbitrary, set $R = \frac{2C}{\epsilon}$, and choose $\delta < \frac{\epsilon}{2}\, e^{-2C/\epsilon}$. Then for every measurable set E with $|E| < \delta$ we find

$$\begin{aligned}\int_E e^{y_n} dx &= \int_{E\cap\{y_n \le R\}} e^{y_n} dx + \int_{E\cap\{y_n > R\}} e^{y_n} dx \\ &\le |E| e^R + \int_{E\cap\{y > R\}} \frac{e^{y_n}}{y_n} y_n dx \le |E| e^R + \frac{C}{R} < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,\end{aligned}$$

where $\{y_n < R\}$ is short for $\{x\colon y_n(x) < R \text{ a.e.}\}$. Hence the integrals $\int_\Omega |e^{y_n}| dx$ are uniformly absolutely continuous, and there exists a subsequence of $\{y_n\}$, again denoted by $\{y_n\}$, such that $e^{y_n} \rightharpoonup \widetilde{w}$ weakly in $L^1(\Omega)$. Since $e^{y_n} \rightharpoonup w^*$ in $H^{-1}(\Omega)$, we have $e^{y_n} \rightharpoonup w^*$ weakly in $L^1(\Omega)$. Lemma 2.2 implies that $w^* = e^{y^*}$, and thus $e^{y_n} \rightharpoonup e^{y^*}$ weakly in $L^1(\Omega)$. Now we can take the limit in

$$(\nabla y_n, \nabla v)_{L_n^2} - \delta(e^{y_n}, v)_{\mathcal{D}',\mathcal{D}} = (u_n, v) \text{ for all } v \in \mathcal{D}(\Omega)$$

and obtain

$$(\nabla y^*, \nabla v)_{L_n^2} - \delta(e^{y^*}, v)_{\mathcal{D}',\mathcal{D}} = (u^*, v) \text{ for all } v \in \mathcal{D}(\Omega).$$

Hence y^* is a solution to (2.2) with $u = u^*$, and the proof is complete. □

We turn to optimality conditions satisfied by a solution (y^*, u^*) of (2.1)–(2.2). It is simple to formally derive the first-order conditions

$$\text{(2.4)} \qquad \begin{cases} -\Delta y^* - \delta e^{y^*} = u^* & \text{in } \Omega, \\ y^* = 0 & \text{on } \partial\Omega, \\ -\Delta \lambda^* - \delta e^{y^*} \lambda^* = J_y(y^*, u^*) & \text{in } \Omega, \\ \lambda^* = 0 & \text{on } \partial\Omega, \\ \lambda^* = J_u(y^*, u^*) & \text{in } \Omega, \end{cases}$$

where J_y and J_u denote the Riesz representative of the partial derivatives of J. We next verify (2.4) in the case in which $\Omega \subset \mathbb{R}^2$.

For this purpose let (y^*, u^*) denote a solution to (2.1)–(2.2). We recall from [GT, p. 155] that

$$\text{(2.5)} \qquad \begin{aligned}&\{|e^y|_{L^4}\colon y \in \mathcal{B}\} \text{ is bounded for every bounded set } \mathcal{B} \subset H_0^1(\Omega), \\ &\text{with } \Omega \subset \mathbb{R}^2.\end{aligned}$$

For this result the restriction to $n = 2$ is essential. It follows in particular that $e^{y^*} \in L^4(\Omega)$ and, utilizing (2.2) and [T, Theorem 2.2.7], we conclude that $y^* \in L^\infty(\Omega) \cap H_0^1(\Omega)$. Let $G: H_0^1(\Omega) \to H^{-1}$ denote the operator

$$Gy = -\Delta y - \delta e^{y^*} y.$$

Note that G is an operator with compact resolvent so that its spectrum consists of eigenvalues only. We shall require the following assumptions.

(A2) $$0 \text{ is not an eigenvalue of } G.$$

To interpret (A2) recall that, because (1.5) is a singular system (see [L]), the existence of a solution to the nonlinear control system, given by the first two equations in (2.4) here, does not imply well-posedness of the linearized system $Gy = g$, with $y \in H_0^1(\Omega)$, $g \in H^{-1}$. The simple structure of the linearization admits a straightforward characterization of the well-posedness condition by means of (A2). In the evolution case such a condition will not be necessary.

(A3) $$\begin{cases} J: H_0^1(\Omega) \times L^2(\Omega) \to \mathbb{R} \text{ is continuously Fréchet-differentiable} \\ \text{in a neighborhood } U(y^*, u^*) \subset H_0^1(\Omega) \times L^2(\Omega) \text{ of } (y^*, u^*). \end{cases}$$

The tracking functional specified below (A1) satisfies (A3).

THEOREM 2.4. *If* (A1)–(A3) *hold,* $n = 2$, *and* (y^*, u^*) *is a solution to* (2.1)–(2.2), *then there exists* $\lambda^* \in H_0^1(\Omega)$ *such that* (2.4) *holds.*

Proof. (i) We first argue by means of the implicit function theorem that there exists a convex neighborhood $U(u^*) \subset L^2(\Omega)$ of u^* and a constant k such that (2.2) has a solution $y(u)$ for every $u \in U(u^*)$ and

$$|y(u) - y^*|_{H_0^1} \le k|u - u^*|_{L^2} \quad \text{for all } u \in U(u^*), \tag{2.6}$$

where $y^* = y(u^*)$. The implicit function theorem is applied to the mapping $g: H_0^1 \times L^2(\Omega) \to H^{-1}(\Omega)$ defined by

$$g(y, u) = -\Delta y - \delta e^y - u.$$

Since (y^*, u^*) is a solution to (2.1)–(2.2), we have $g(y^*, u^*) = 0$. Moreover, by (2.5) the mapping g is continuous and Fréchet-differentiable with respect to y, with Fréchet derivative

$$g_y(y, u)\delta y = -\Delta \delta y - \delta e^y \delta y.$$

Note that $g_y(y^*, u^*) = G$. By (A2) the operator G is continuously invertible from $H^{-1}(\Omega)$ to $H_0^1(\Omega)$, and hence the implicit function theorem allows us to ascertain the existence of $U(u^*)$ and k such that (2.6) holds.

(ii) Let $u \in U(u^*)$ be arbitrary, and set $v = u - u^*$ and $y(t) = y(u^* + tv)$ for $t \in [0, 1]$. Observe that

$$\begin{aligned} 0 &= g(y(t), u^* + tv) - g(y^*, u^*) \\ &= -\Delta(y(t) - y^*) - tv - \delta e^{y(t)} + \delta e^{y^*}. \end{aligned} \tag{2.7}$$

Since by assumption $J_y(y^*, u^*) \in H^{-1}$ and due to (A2), there exists $\lambda^* \in H_0^1(\Omega)$ such that the third and fourth equations in (2.4) hold. Observe that

$$\begin{aligned} &0 \le J(y(t), u^* + tv) - J(y^*, u^*) = J_y(y^*, u^*)(y(t) - y^*) + tJ_u(y^*, u^*)v \\ &+ \textstyle\int_0^1 [J'(y^* + s(y(t) - y^*), u^* + stv) - J'(y^*, u^*)](y(t) - y^*, tv)ds, \end{aligned}$$

and hence by (2.6) and the definition of $y(t)$

$$0 \le \liminf_{t\to 0^+} \tfrac{1}{t}\,\{J_y(y^*,u^*)(y(t)-y^*) + t\,J_u(y^*,u^*)v\}.$$

By (2.7) we obtain

$$\begin{aligned} 0 \le \liminf_{t\to 0^+} \tfrac{1}{t}\,\big\{ & J_y(y^*,u^*)(y(t)-y^*) + t\,J_u(y^*,u^*)v \\ & - \langle y(t)-y^*, \Delta\lambda^*\rangle_{H_0^1,H^{-1}} - t(v,\lambda^*)_{L^2} \\ & - \delta\left(\textstyle\int_0^1 (e^{y^*+s(y(t)-y^*)} - e^{y^*})(y(t)-y^*)ds, \lambda^*\right)_{L^2} \\ & - \delta(e^{y^*}(y(t)-y^*),\lambda^*)_{L^2}\big\}. \end{aligned}$$

Using (2.4), (2.6), and (2.7), we can pass to the limit in this last inequality to obtain $0 \le J_u(y^*,u^*)v - (v,\lambda^*)_{L^2}$.

Since $u \in U(u^*)$ was arbitrary, it follows that

$$\lambda^* = J_u(y^*,u^*).$$

Here we do not distinguish in notation between the Fréchet-derivative and its Riesz representative in $L^2(\Omega)$. □

Let us turn to the boundary control problem

$$\min\ J(y,u) \tag{2.8}$$

subject to $(y,u) \in H^1(\Omega)\times L^2(\partial\Omega)$ and

$$\begin{cases} -\Delta y - \delta e^y = f & \text{in } \Omega, \\ \frac{\partial y}{\partial n} = u & \text{on } \partial\Omega, \end{cases} \tag{2.9}$$

where $f \in L^2(\Omega)$ is fixed, $\Omega \subset \mathbb{R}^2$, and (2.9) is understood in the variational sense, i.e.,

$$(\nabla y, \nabla v)_{L_n^2} - \delta(e^y, v)_{L^2} = (f,v)_{L^2} + (u,v)_{\partial\Omega} \ \text{ for all } \ v \in H^1(\Omega). \tag{2.10}$$

We shall require the following modifications of (A1)–(A3).

(A1′) This is (A1) with $L^2(\Omega)$ replaced by $L^2(\partial\Omega)$.

(A2′) $\begin{cases} \text{This is (A2) with } G \text{ replaced by } \widetilde{G}\colon H^1(\Omega) \to H^1(\Omega)^*, \\ \text{where } \widetilde{G}(\delta y) \text{ is the element of } H^1(\Omega)^* \text{ characterized by} \\ v \to (\nabla\delta y, \nabla v)_{L_n^2} - \delta(e^{y^*}\delta y, v) \text{ for all } v \in H^1(\Omega). \end{cases}$

(A3′) This is (A3) with $H_0^1(\Omega)\times L^2(\Omega)$ replaced by $H^1(\Omega)\times L^2(\partial\Omega)$.

If (A1′) holds, then assuming that the feasible set is nonempty, it is simple to argue the existence of a solution $(y^*,u^*) \in H^1(\Omega)\times L^2(\partial\Omega)$ with $e^{y^*} \in H^1(\Omega)^* \cap L^1(\Omega)$. The formal first-order necessary optimality condition for (2.8)–(2.9) is given by

$$\begin{cases} -\Delta y^* - \delta e^{y^*} = f & \text{in } \Omega, \\ \frac{\partial y^*}{\partial n} = u^* & \text{on } \partial\Omega, \\ -\Delta\lambda^* - \delta e^{y^*}\lambda^* = J_y(y^*,u^*) & \text{in } \Omega, \\ \frac{\partial\lambda^*}{\partial n} = 0 & \text{on } \partial\Omega, \\ \lambda^* = J_u(y^*,u^*) & \text{on } \partial\Omega. \end{cases} \tag{2.11}$$

THEOREM 2.5. *If* (A1′)–(A3′) *hold,* $n = 2$*, and* (y^*, u^*) *is a solution to* (2.8)–(2.9)*, then there exists* $\lambda^* \in H^1(\Omega)$ *such that* (2.11) *holds.*

Proof. The proof is similar to that of Theorem 2.4, so we only give the necessary changes. For $\widetilde{g}\colon H^1(\Omega) \times L^2(\partial\Omega) \to (H^1(\Omega))^*$ defined by

$$\widetilde{g}(y,u) = (\nabla y, \nabla v)_{L^2_n} - \delta(e^y, v)_{L^2} - (u,v)_{L^2(\partial\Omega)} \quad \text{for all} \quad v \in H^1(\Omega),$$

one argues that due to (A2′) and (2.5) the implicit function theorem is applicable, and thus there exists a neighborhood $U(u^*)$ of u^* in $L^2(\partial\Omega)$ and a constant $k > 0$ such that for all $u \in U(u^*)$ there exists a solution $y(u) \in H^1(\Omega)$ to (2.9) and

$$|y(u) - y^*|_{H^1} \le |u - u^*|_{L^2(\partial\Omega)} \quad \text{for all} \quad u \in U(u^*).$$

Due to (A2′) there exists a solution λ^* to the variational form of the third and fourth equations in (2.11), i.e., $(\nabla\lambda^*, \nabla v)_{L^2_n} - \delta(e^{y^*}\lambda^*, v)_{L^2} = \langle J_y(y^*, u^*), v\rangle_{(H^1)^*, H^1}$ for all $v \in H^1(\Omega)$.

For arbitrary $u \in U(u^*)$ we find, as in (ii) of the proof of Theorem 2.4,

$$\begin{aligned} 0 \le \liminf_{t\to 0^+} \tfrac{1}{t}\, & J_y(y^*,u^*)(y(t)-y^*) + t\, J_u(y^*,u^*)(u-u^*) \\ & + (\nabla(y(t)-y^*), \nabla\lambda^*)_{L^2_n} - t\,(u-u^*, \lambda^*)_{L^2(\partial\Omega)} \\ & - \delta\left(\int_0^1 (e^{y^* + s(y(t)-y^*)} - e^{y^*})(y(t)-y^*)ds, \lambda^*\right)_{L^2} - \delta(e^{y^*}(y(t)-y^*), \lambda^*)_{L^2} \\ = {} & J_u(y^*,u^*)(u-u^*) - (u-u^*, \lambda^*)_{L^2(\partial\Omega)}, \end{aligned}$$

and thus (2.11) follows.

3. Evolutionary optimal control problems. This section is devoted to optimal control problems of the type

$$\min J(y, y(T,\cdot), u) \tag{3.1}$$

subject to $(y, y(T,\cdot), u) \in W \times L^2(\Omega) \times L^2(Q)$ and

$$\begin{cases} y_t = \Delta y + \delta e^y + u & \text{in } Q = (0,T] \times \Omega, \\ y(0,\cdot) = \varphi & \text{in } \Omega, \\ y = 0 & \text{on } \Sigma = (0,T] \times \partial\Omega, \end{cases} \tag{3.2}$$

where Ω is a bounded domain in $\mathbb{R}^n$ with smooth boundary $\partial\Omega$, $\varphi \in L^2(\Omega)$, and

$$W = \{y \in L^2(0,T; H^1_0(\Omega))\colon y_t \in L^1(0,T; X_1^*)\}$$

endowed with $|y|_W = |y|_{L^2(H^1_0(\Omega))} + |y_t|_{L^1(x_1^*)}$ as norm. Here $|y|_{L^2(H^1_0(\Omega))}$ stands for $|y|_{L^2(0,T;H^1_0(\Omega))}$. X_1^* denotes the topological dual of a reflexive Banach space X_1 with the property that

$$\mathcal{D}(\Omega) \hookrightarrow X_1 \hookrightarrow (L^\infty(\Omega) \cap H^1_0(\Omega))$$

with continuous and dense injections. It follows that $(L^1(\Omega) \cup H^{-1}(\Omega)) \hookrightarrow X_1^* \subset \mathcal{D}'(\Omega)$ with continuous and dense injections as well. The choice $X_1 = W^{1,p}_0(\Omega)$, $p > n$, satisfies the specified properties. Henceforth u is assumed to be an element of $L^2(Q)$. Contents permitting, $L^p(X)$ stands for $L^p(0,T;X)$ with X a Banach space and $p \in [1,\infty]$.

DEFINITION 3.1. *A function $y \in W$ is called a solution to* (3.2) *if* $e^y \in L^1(Q)$ *and*

$$\text{(3.3)} \qquad \begin{cases} (y_t, v)_{L^1(X_1^*),L^\infty(X_1)} + (\nabla y, \nabla v)_{L^2(L^2_n)} = \delta(e^y, v)_{L^1(Q),L^\infty(Q)} \\ \qquad\qquad + (u, v)_{L^2(Q)} \quad \textit{for all } v \in L^\infty(0,T;X_1), \\ y(0,\cdot) = \varphi. \end{cases}$$

Recall that every element of W, and hence every solution to (3.2), can be identified a.e. with respect to $t \in [0,T]$ with a function in $C([0,T];X_1^*)$; see [Ba]. Since $L^\infty(0,T;X_1)$ is the dual space to $L^1(0,T;X_1^*)$, a function $y \in W$ is a solution to (3.2) if and only if $e^y \in L^1(Q)$, $y(0,\cdot) = \varphi$, and

$$\text{(3.4)} \qquad y_t = \Delta y + \delta e^y + u \text{ in } L^1(0,T;X_1^*).$$

We shall require the following a priori bound on $e^y y$ when y is a solution to (3.2).

PROPOSITION 3.2. *Let $u \in L^2(Q)$ and let $y = y(u)$ be a solution to* (3.2) *with $y(T,\cdot) \in L^2(\Omega)$. Then $e^y y \in L^1(Q)$ and*

$$\text{(3.5)} \qquad \begin{aligned} &\delta \textstyle\int_Q e^y y dx + \frac{1}{2}|\varphi|^2_{L^2(\Omega)} \\ &\le \tfrac{1}{2}|y(T,\cdot)|^2_{L^2(\Omega)} + |y|^2_{L^2(0,T;H_0^1)} + \tfrac{1}{2}|u|^2_{L^2(Q)} + \tfrac{1}{2}|y|^2_{L^2(Q)}. \end{aligned}$$

Formally, (3.5) is obtained by taking the inner product in $L^2(Q)$ of (3.4) with y. A detailed proof is given in the appendix.

To establish the existence of a solution to (3.1)–(3.2) the following assumption is required:

$$\text{(A4)} \qquad \begin{cases} J\colon \mathcal{X} = L^2(0,T;H_0^1(\Omega)) \times L^2(\Omega) \times L^2(Q) \to \mathbb{R} \text{ is bounded from} \\ \text{below, weakly lower semicontinuous, and coercive in the sense that} \\ J(y, y(T,\cdot), u) \to \infty \text{ if } |y|_{L^2(0,T;H_0^1)} \to \infty \text{ or } |y(T)|_{L^2(\Omega)} \to \infty \\ \text{or } |u|_{L^2(Q)} \to \infty. \end{cases}$$

THEOREM 3.3. *If* (A4) *holds, then* (3.1)–(3.2) *admits a solution* $(y^*, y^*(T,\cdot), u^*) \in W \times L^2(\Omega) \times L^2(Q)$.

Proof. The set of feasible points for (3.1)–(3.2) is nonempty. By assumption (A4) the functional J is bounded from below and there exists a minimizing sequence $\{(y_n, y_n(T,\cdot), u_n\}_{n=1}^\infty \in W \times L^2(\Omega) \times L^2(Q)$ for (3.1)–(3.2). Due to (A4) there exists a subsequence, denoted by the same symbol, and $(y^*, z^*, u^*) \in \mathcal{X}$ such that

$$\text{(3.6)} \qquad (y_n, y_n(T,\cdot), u_n) \rightharpoonup (y^*, z^*, u^*) \text{ in } \mathcal{X}.$$

From (3.5), moreover,

$$\text{(3.7)} \qquad \{e^{y_n} y_n\}_{n=1}^\infty \text{ is bounded in } L^1(Q).$$

Due to the weak lower semicontinuity of J we have

$$J(y^*, y^*(T,\cdot), u^*) \le J(y, y(T,\cdot), u)$$

for all pairs $(y, y(T,\cdot), u) \in W \times L^2(\Omega) \times L^2(Q)$ which satisfy (3.2). It thus remains to show that $(y^*, u^*) \in W \times L^2(Q)$ satisfies (3.2). Since $\{y_n\}_{n=1}^\infty$ is bounded in $L^2(0,T;H_0^1(\Omega))$ and since (3.7) with $y_n = y(u_n)$ a solution to (3.2) holds, a generalization of Aubin's lemma given in Lemma 3.4 below implies the existence of a subsequence of $\{y_n\}_{n=1}^\infty$ denoted by the same symbol and of an element $\tilde{y} \in L^1(0,T;L^2(\Omega))$

such that $y_n \to \tilde{y}$ in $L^1(0,T;L^2(\Omega))$. Since $y_n \rightharpoonup y^*$ in $L^2(0,T;H_0^1(\Omega))$, it follows that

$$y_n \rightharpoonup y^* \text{ in } L^1(0,T;L^2(\Omega)). \tag{3.8}$$

As a consequence there exists a further subsequence, again denoted by y_n, converging a.e. in Q to y^*. Due to (3.7) the Dunford–Pettis theorem can be used as in the proof of Theorem 2.3 to argue that

$$e^{y_n} \rightharpoonup e^{y^*} \text{ in } L^1(Q). \tag{3.9}$$

Finally, we consider the convergence of $\{\frac{d}{dt}y_n\}_{n=1}^{\infty}$. Due to (3.6), (3.9), and (3.4), with u replaced by u_n, there exists $z \in L^1(0,T;X_1^*)$ such that

$$\frac{d}{dt}y_n \rightharpoonup z \text{ in } L^1(0,T;X_1^*). \tag{3.10}$$

To argue that $z = \frac{d}{dt}y^*$, we introduce the differentiation operator

$$D\colon \text{dom } (D) \subset L^1(0,T;X_1^*) \to L^1(0,T;X_1^*),$$

with dom $(D) = \{y \in L^1(0,T;X_1^*), \frac{d}{dt}y \in L^1(0,T;X_1^*), y(0) = 0\}$. Since the inverse of D is a bounded linear operator, D is closed. Due to (3.8) and (3.10),

$$y_n - \varphi \to y^* - \varphi \quad \text{and} \quad \frac{d}{dt}y_n \rightharpoonup z \text{ in } L^1(0,T;X_1^*),$$

and hence $y^* = z$ and

$$\frac{d}{dt}y_n \rightharpoonup z \text{ in } L^1(0,T;X_1^*). \tag{3.11}$$

Taking the limit in

$$\begin{cases} (\frac{d}{dt}y_n, v)_{L^1(X_1^*),L^\infty(X_1)} + (\nabla y_n, \nabla v)_{L^2(L_n^2)} = \delta(e^{y_n}, v)_{L^1(Q),L^\infty(Q)} + (u,v)_{L^2(Q)} \\ \qquad\qquad \text{for all } v \in L^\infty(0,T;X_1), \\ y_n(0,\cdot) = \varphi, \end{cases}$$

and utilizing (3.6), (3.9), and (3.11), we find

$$\begin{cases} (\frac{d}{dt}y^*, v)_{L^1(X_1^*),L^\infty(X_1)} + (\nabla y^*, \nabla v)_{L^2(L_n^2)} = \delta(e^{y^*}, v)_{L^1(Q),L^\infty(Q)} + (u,v)_{L^2(Q)} \\ \qquad\qquad \text{for all } v \in L^\infty(0,T;X_1), \\ y^*(0,\cdot) = \varphi. \end{cases}$$

Thus y^* is a solution to (3.2) with $u = u^*$ as desired.

LEMMA 3.4. *Let y_n denote solutions to* (3.2) *with u replaced by $u_n \in L^2(Q)$, $n \in \mathbb{N}$, and assume that $\{u_n\}_{n=1}^{\infty}$ is bounded in $L^2(Q)$, that $\{y_n\}_{n=1}^{\infty}$ is bounded in $L^2(0,T;H_0^1(\Omega))$, and that $\{e^{y_n}y_n\}_{n=1}^{\infty}$ is bounded in $L^1(Q)$. Then there exists a subsequence $\{y_{n_k}\}_{k=1}^{\infty}$ of $\{y_n\}_{n=1}^{\infty}$ and $y^* \in L^2(0,T;H_0^1)$ such that $y_{n_k} \to y^*$ in $L^1(0,T;L^2(\Omega))$ as $k \to \infty$.*

Proof. For the proof we follow essentially an Aubin-lemma argument [CF].

(i) Since $H_0^1(\Omega) \subset L^2(\Omega) \subset X_1^*$ with $H_0^1(\Omega)$, compact in $L^2(\Omega)$, there exists for every $\epsilon > 0$ a constant $c_\epsilon > 0$ such that

$$|y|_{L^2(\Omega)} \le \epsilon |y|_{H_0^1(\Omega)} + c_\epsilon |y|_{X_1^*} \text{ for all } y \in H_0^1(\Omega). \tag{3.12}$$

(ii) Due to the assumptions on the boundedness of $\{y_n\}_{n=1}^{\infty}$, there exists $y^* \in L^2(0,T;H_0^1)$ and a subsequence of $\{y_n\}$, denoted by the same symbol, such that $y_n \rightharpoonup y^*$ in $L^2(0,T;H_0^1)$. We need to show that $y_n \to y^*$ in $L^1(0,T;L^2(\Omega))$. By assumption there exists a constant $C>0$ such that

$$|y_n|_{L^2(H_0^1(\Omega))} \le C, \quad |u_n|_{L^2(Q)} \le C, \quad \text{and} \quad |e^{y_n} y_n|_{L^1(Q)} \le C \tag{3.13}$$

for all $n \in \mathbb{N}$. Due to (3.12) and (3.13) we have for every $\epsilon > 0$

$$\begin{aligned} &\textstyle\int_0^T |y_n - y^*|_{L^2(\Omega)} dt \le \epsilon \int_0^T |y_n - y^*|_{H_0^1(\Omega)} dt + c_\epsilon \int_0^T |y_n - y^*|_{X_1^*} dt \\ &\le \epsilon\sqrt{T}\left(C + |y^*|_{L^2(H_0^1(\Omega))}\right) + c_\epsilon \textstyle\int_0^T |y_n - y^*|_{X_1^*} dt \text{ for all } n. \end{aligned} \tag{3.14}$$

(iii) We next show that $y_n(t) \to y^*(t)$ in X_1^* for almost every $t \in (0,T)$. First observe that for every interval $I \subset [0,T]$

$$\int_I y_n(t,\cdot)dt \to \int_I y^*(t,\cdot)dt \text{ in } L^2(\Omega). \tag{3.15}$$

In fact, let χ_I denote the characteristic function of I and let $\ell \in H^{-1}(\Omega)$ be arbitrary. Then

$$\left(\int_I (y_n(t,\cdot) - y^*(t,\cdot))dt, \ell\right)_{H_0^1, H^{-1}} = \int_0^T (y_n(t,\cdot) - y^*(t,\cdot), \ell)_{H_0^1, H^{-1}} \chi_I \, dt \to 0,$$

since $y_n \rightharpoonup y^*$ in $L^2(0,T;H_0^1)$. Hence $\int_I y_n(t,\cdot)dt \rightharpoonup \int_I y^*(t,\cdot)dt$ in $H_0^1(\Omega)$, and (3.15) follows by the compactness of $H_0^1(\Omega)$ in $L^2(\Omega)$. To verify that $y_n(t) \to y^*(t)$ in X_1^* for almost every $t \in (0,T)$, let $\delta > o$ be arbitrary. We have the following equation in X_1^*:

$$y_n(t) = \frac{1}{\delta}\int_{t-\delta}^t y_n(s)ds + \frac{1}{\delta}\int_{t-\delta}^t (s-t+\delta)\frac{d}{ds}\, y_n(s)ds. \tag{3.16}$$

We turn to the second term on the right-hand side of (3.16). Utilizing (3.15) with u replaced by u_n, and denoting by k the embedding constant of $L^1(\Omega)$ as well as $H^{-1}(\Omega)$ into X_1^*, we find

$$\begin{aligned} &\tfrac{1}{\delta}\left|\textstyle\int_{t-\delta}^t (s-t+\delta)\frac{d}{ds} y_n(s)ds\right|_{X_1^*} \le \tfrac{k}{\delta}\textstyle\int_{t-\delta}^t |(s-t+\delta)\Delta y_n(s)|_{H^{-1}}\, ds \\ &+ \tfrac{k}{\delta}\textstyle\int_{t-\delta}^t |(s-t+\delta)e^{y_n(s)}|_{L^1} ds + \tfrac{k}{\delta}\textstyle\int_{t-\delta}^t |(s-t+\delta)u_n(s)|_{L^2}\, ds \\ &\le \tfrac{k}{\delta}\Big[(\textstyle\int_{t-\delta}^t (s-t+\delta)^2 ds)^{1/2}((\int_{t-\delta}^t |\Delta y_n(s)|_{H^{-1}}^2 ds)^{1/2} + (\int_{t-\delta}^t |u_n(s)|_{L^2}^2 ds)^{1/2}) \\ &\qquad + \textstyle\int_{t-\delta}^t |(s-t+\delta)e^{y_n(s)}|_{L^1} ds\Big] \\ &\le \tfrac{k}{\delta}\left[(\tfrac{\delta^3}{3})^{1/2} C(1+|\Omega|^{1/2}) + \int_0^\delta \left|\tfrac{s}{y_n(t+s-\delta)}\, e^{y_n(t+s-\delta)}\, y_n(t+s-\delta)\right|_{L^1} ds\right], \end{aligned}$$

where we used (3.13). For $R>0$ we decompose Q into $Q_R = \{(t,x) \in Q : y_n(t,x) \le R\}$, where the dependence of Q_R on n is supressed, and its complement. Utilizing the last inequality in (3.13), we obtain

$$\begin{aligned} &\tfrac{1}{\delta}\left|\textstyle\int_{t-\delta}^t (s-t+\delta)\frac{d}{ds}\, y_n(s)ds\right|_{X_1^*} \le \tfrac{kC}{\sqrt{3}}\sqrt{\delta}(1+|\Omega|^{1/2}) \\ &+ \tfrac{k}{\delta}\big[\textstyle\int_0^\delta s \int_\Omega e^R dx\, ds + \frac{1}{R}\int_0^\delta \int_\Omega y_n(t+s-\delta)e^{y_n(t+s-\delta)} ds\big] \\ &\le \tfrac{kC}{\sqrt{3}}\sqrt{\delta}(1+|\Omega|^{1/2}) + \tfrac{1}{2}\delta e^R|\Omega| + \tfrac{C}{R}. \end{aligned} \tag{3.17}$$

Henceforth let t be a Lebesgue point for y^*, so that

$$\lim_{\delta\to 0^+} \frac{1}{\delta}\int_{t-\delta}^{t} y^*(s)ds = y^*(t). \tag{3.18}$$

Let $\epsilon > 0$ be arbitrary. Due to (3.17) and (3.18), one can choose R and subsequently δ such that $\frac{1}{\delta}|\int_{t-\delta}^{t}(s-t+\delta)\frac{d}{ds}\, y_n(s)ds|_{X_1^*} \le \epsilon_0$ and $|\frac{1}{\delta}\int_{t-\delta}^{t} y^*(s)ds - y^*(t)| \le \epsilon$ for all n. Combined with (3.16) we find $|y_n(t)-y^*(t)|_{X_1^*} \le |\frac{1}{\delta}\int_{t-\delta}^{t}(y_n(s)-y^*(s))ds|_{X_1^*} + 2\epsilon_0$, and, since δ is fixed in this last inequality, (3.15) implies that $y_n(t) \to y^*(t)$ in X_1^*. The set of Lebesgue points of y^* is dense in $(0,T)$, and thus $y_n(t) \to y^*(t)$ in X_1^* for almost every $t \in (0,T)$.

(iv) Due to (3.16) and (3.17) Lebesgue's bounded convergence theorem is applicable to passing to the limit on the right-hand side of (3.14). This implies the desired result. □

Lemma 3.4 differs from the classical Aubin lemma [CF] in that $\{\frac{d}{dt}\, y_n\}_{n=1}^{\infty}$ is not bounded in $L^p(0,T;X_1^*)$ for some $p > 1$. Rather, the boundedness of $\{e^{y_n}y_n\}_{n=1}^{\infty}$ in $L^1(Q)$ is used as well as the assumption that the functions y_n are solutions of (3.2).

We turn to a modification of (3.1)–(3.2) and consider a cost functional without atom at T as in (3.1). For this purpose we define

$$W_G = \left\{y \in L^2(0,T;H_0^1(\Omega))\colon y_t \in L_G^1(0,T;X_1^*)\right\},$$

where

$$L_G^1(0,T;X_1^*) = \left\{y\colon (0,T) \to X_1^* \text{ measurable} : \int_0^T |y(t)|_{X_1^*}(T-t)dt < \infty\right\}.$$

For $u \in L^2(Q)$ a function $y = y(u) \in W_G$ is called the solution in W_G of (3.2) if $e^y \in L_G^1$ and

$$\left\{\begin{array}{rcl} y_t & = & \Delta y + \delta e^y + u \text{ in } L_G^1, \\ y(0) & = & \varphi. \end{array}\right. \tag{3.19}$$

Let us consider

$$\left\{\begin{array}{l} \min J(y,u) \\ \text{ subject to } (y,u) \in W_G \times L^2(Q) \text{ with } y \text{ a solution to (3.2)}, \end{array}\right. \tag{3.20}$$

where Ω satisfies the properties specified at the beginning of this section and $\varphi \in L^2(\Omega)$. We require

$$\left\{\begin{array}{l} J\colon L^2(0,T;H_0^1(\Omega)) \times L^2(Q) \to \mathbb{R} \text{ is bounded from below,} \\ \text{ weakly lower semicontinuous, and coercive in the sense that} \\ J(y,u) \to \infty \text{ if } |y|_{L^2(0,T;H_0^1)} \to \infty \text{ or } |u|_{L^2(Q)} \to \infty. \end{array}\right. \tag{A5}$$

THEOREM 3.5. *If* (A5) *holds, then* (3.20) *admits a solution* $(y^*,u^*) \in W_G \times L^2(Q)$.

Proof. The set of feasible points is nonempty. By (A5) there exists a minimizing sequence $\{(y_n,u_n)\} \in L^2(0,T;H_0^1(\Omega)) \times L^2(Q)$ for (3.20). Moreover, there exists a subsequence, denoted by the same symbol, as well as $(y^*,u^*) \in L^2(0,T;H_0^1(\Omega)) \times L^2(Q)$ such that

$$(y_n,u_n) \rightharpoonup (y^*,u^*) \text{ in } L^2(0,T;H_0^1(\Omega)) \times L^2(Q). \tag{3.21}$$

Due to the weak lower semicontinuity of J,

$$J(y^*, u^*) \leq J(y, u)$$

for all (y, u) satisfying (3.2). It thus remains to show that $y^* \in W_G$ and that (y^*, u^*) is a solution to (3.2).

Let $y = y(u)$ with $u \in L^2(Q)$ be a solution to (3.2). Taking the inner product of (3.2) in $L^2(Q)$ with $(T-t)y$, one computes that

$$\begin{aligned} &\delta \textstyle\int_Q e^y y(T-T)dQ + \frac{T}{2}|\varphi|^2_{L^2(\Omega)} \\ &\leq |y|^2_{L^2(0,T;H_0^1)} + |y|^2_{L^2(Q)} + \tfrac{T^2}{2}|u|^2_{L^2(Q)}. \end{aligned} \tag{3.22}$$

The detailed proof is similar to that of Proposition 3.2. Let k^0 be such that $T \geq \frac{1}{k_0}$. From (3.22) it follows that $\{e^{y_n} y_n|_{(0,T-\frac{1}{k})}\}_{n=1}^{\infty}$ is bounded in $L^2(0, T-\frac{1}{k}; H_0^1(\Omega))$ for every $k \geq k_0$. By construction, $\{y_n|_{(0,T-\frac{1}{k})}\}_{n=1}^{\infty}$ is bounded in $L^2(0, T-\frac{1}{k}; H_0^1(\Omega))$ for every $k \geq k_0$ as well. Arguing as in the proof of Theorem 3.3, there exists a nested sequence of subsequences satisfying $\{n_{k+1}\} \prec \{n_k\} \prec \{n\}, k \geq k_0$, and

$$\begin{aligned} y_{n_k} &\rightharpoonup y^* \text{ in } L^1\left(0, T-\frac{1}{k}; L^2(\Omega)\right), \\ e^{y_{n_k}} &\rightharpoonup e^{y^*} \text{ in } L^1\left(\left(0, T-\frac{1}{k}\right) \times \Omega\right), \\ \frac{d}{dt} y_{n_k} &\rightharpoonup \frac{d}{dt} y^* \text{ in } L^1\left(0, T-\frac{1}{k}; X_1^*\right) \text{ as } n_k \to \infty, \end{aligned}$$

and y^* is a solution to (3.2) on $(0, T-\frac{1}{k})$ for $k \geq k_0$. In particular, this implies that $\frac{d}{dt}y^* = \Delta y^* + \delta e^{y^*} + u^*$ for almost every $t \in [0,T]$, and $y^*(0) = \varphi$. It remains to argue that $y^* \in W_G$. Since $y^*|_{(0,T-\frac{1}{k})}$ is a solution to (3.2) on $(0, T-\frac{1}{k})$, for $k \geq k_0$, one argues as for (3.22) to obtain for every $k \geq k_0$

$$\begin{aligned} \delta \int_0^{T-\frac{1}{k}} \int_\Omega \left| e^{y^*} y^* \left(T - \tfrac{1}{k} - t\right)\right| dxdt + \tfrac{T-\frac{1}{k}}{2}|\varphi|^2_{L^2(\Omega)} \\ \leq |y^*|^2_{L^2(0,T;H_0^1)} + |y^*|^2_{L^2(Q)} + \tfrac{T^2}{2}|u^*|^2_{L^2(Q)}. \end{aligned}$$

Taking the limit with respect to k, we find

$$\delta \int_Q e^{y^*} y^*(T-t)dQ + \tfrac{T}{2}|\varphi|^2_{L^2(\Omega)} \leq |y^*|^2_{L^2(0,T;H_0^1)} + |y^*|^2_{L^2(Q)} + \tfrac{T^2}{2}|u^*|^2_{L^2(Q)},$$

and consequently $\int_Q e^{y^*}(T-t)dQ < \infty$. Since $y^* \in L^2(0,T;H_0^1(\Omega)$, it follows that $y^* \in W_G$ and $\frac{d}{dt}y^* = \Delta y^* + \delta e^{y^*} + u^*$ in $L^1_G(0,T;x_1^*)$, as desired. ∎

3.1. Optimality conditions. We turn to the optimality condition satisfied by a solution $(y^*, y^*(T,\cdot), u^*)$ to (3.1)–(3.2). Again it is simple to formally derive first-order optimality conditions:

$$\begin{cases} y_t^* = \Delta y^* + \delta e^{y^*} + u^* \text{ in } Q, \\ y^*(0,\cdot) = \varphi \text{ in } \Omega, \quad y^* = 0 \text{ on } \Sigma, \\ -\lambda_t^* = \Delta\lambda^* + \delta e^{y^*}\lambda^* - J_y(y^*, y^*(T,\cdot), u^*) \text{ in } Q, \\ \lambda^*(T,\cdot) = J_{y(T,\cdot)}(y^*, y^*(T,\cdot), u^*) \text{ in } \Omega, \quad \lambda^* = 0 \text{ on } \Sigma, \\ \lambda^* = J_u(y^*, y^*(T,\cdot), u^*) \text{ in } Q. \end{cases} \tag{3.23}$$

We shall use (3.23)(i) to denote the first equation in (3.23), and similarly for the other equations in (3.23). Next (3.23) is justified in special cases. We recall the definition of $\mathcal{X}$ in (A4) and introduce a further assumption:

$$\text{(A6)} \quad \begin{cases} J\colon \mathcal{X} \to \mathbb{R} \text{ is continuously Fréchet-differentiable in a neighborhood} \\ U(y^*, y^*(T,\cdot), u^*) \subset \mathcal{X} \text{ of } (y^*, y^*(T,\cdot), u^*). \end{cases}$$

We set $\mathcal{W} = \{y \in L^2(0,T; H_0^1(\Omega))\colon y_t \in L^2(0,T; H^{-1}(\Omega))\}$, and recall from the theory of parabolic equations that there exists a solution $\lambda^* \in \mathcal{W}$ to (3.23)(iii) and (iv), provided that e^{y^*} is sufficiently smooth.

THEOREM 3.6. *If* (A4) *and* (A6) *hold,* $n \le 2, \varphi \in H_0^1(\Omega)$, *and* $(y^*, y^*(T,\cdot), u^*)$ *is a solution to* (3.1)–(3.2) *with* $e^{y^*} \in L^2(Q)$, *then there exists* $\lambda^* \in \mathcal{W}$ *such that* (3.23) *holds.*

Proof. For the proof we require the space

$$W^{2,1} = \{y \in L^2(0,T; H^2(\Omega) \cap H_0^1(\Omega))\colon y_t \in L^2(0,T; L^2(\Omega))\},$$

endowed with the natural Hilbert-space norm. Recall that $W^{2,1}$ is continuously embedded in $C(0,T; H_0^1(\Omega))$, and hence for $y \in W^{2,1}$ we find $e^y \in C(0,T; L^4(\Omega))$. The assumptions $e^{y^*} \in L^2(Q)$ and $\varphi \in H_0^1(\Omega)$ imply that $y^* \in W^{2,1}$.

(i) We now follow the proof of Theorem 2.4 and argue that there exists a convex neighborhood $U(u^*) \subset L^2(Q)$ of u^* and a constant k such that (3.2) has a solution $y(u)$ for every $u \in U(u^*)$ and

$$|y(u) - y^*|_{W^{2,1}} \le k|u - u^*|_{L^2(Q)} \text{ for every } u \in U(u^*). \tag{3.24}$$

The implicit function theorem is applied to the mapping $g\colon W^{2,1} \times L^2(Q) \to L^2(Q) \times L^2(\Omega)$ defined by

$$g(y,u) = (y_t - \Delta y - \delta e^y - u, y(0,\cdot) - \varphi).$$

Note that $g(y^*, u^*) = 0$. Utilizing (2.5), one can show (see, e.g., [KK]) that $y \to e^y$ is continuously Fréchet-differentiable from $W^{2,1}$ to $L^\infty(0,T; L^4(\Omega))$. It follows that g is continuously Fréchet-differentiable with its partial derivative with respect to y given by

$$g_y(y,u)\delta y = ((\delta y)_t - \Delta \delta y - \delta e^y \delta y, \delta y(0,\cdot)).$$

The theory of parabolic equations implies that $g_y(y^*, u^*)\colon \mathcal{W} \to L^2(Q) \times L^2(\Omega)$ is continuously invertible, and hence (3.24) follows.

(ii) Proceeding as in the proof of Theorem 2.4, we choose $u \in U(u^*)$ arbitrarily and set $v = u - u^*$ and $y(t) = y(u^* + tv)$ for $t \in [0,1]$. Observe that

$$0 = g(y(t), u^* + tv) - g(y^*, u^*) = \frac{d}{dt}(y(t) - y^*) - \Delta(y(t) - y^*) - tv - \delta e^{y(t)} + \delta e^{y^*(t)}. \tag{3.25}$$

By assumption, $J_y(y^*, y^*(T,\cdot), u^*) \in L^2(0,T; H^{-1}(\Omega))$, and hence there exists $\lambda^* \in \mathcal{W}$ such that (3.23)(iii) and (iv) hold. It remains to verify the last equation of (3.23). Due to (A6) and (3.25) we find

$$
\begin{aligned}
0 \le \liminf_{t\to 0^+} \tfrac{1}{t} \Big\{ & J_y(y^*, y^*(T,\cdot), u^*)(y(t) - y^*) \\
& + J_{y(T,\cdot)}(y^*, y^*(T,\cdot), u^*)(y(t)(T,\cdot) - y^*(T,\cdot)) + tJ_u(y^*, y^*(T,\cdot), u^*)v \\
& + (y(t)(T,\cdot) - y^*(T,\cdot), \lambda^*(T,\cdot))_{L^2(\Omega)} \\
& - \langle y(t) - y^*, \lambda_t^* \rangle_{L^2(0,T;H_0^1), L^2(0,T;H^{-1})} \\
& - \langle y(t) - y^*, \Delta\lambda^* \rangle_{L^2(0,T;H_0^1), L^2(0,T;H^{-1})} - t(v, \lambda^*)_{L^2(Q)} \\
& - \delta \left(\textstyle\int_0^1 (e^{y^* + s(y(t)-y^*)} - e^{y^*})(y(t) - y^*) ds, \lambda^* \right)_{L^2(Q)} \\
& - \delta (e^{y^*}(y(t) - y^*), \lambda^*)_{L^2(Q)} \Big\} = J_u(y^*, y^*(T,\cdot), u^*)v - (v, \lambda^*)_{L^2(Q)},
\end{aligned}
$$

where we used (3.23)(iii) and (iv) and the fact that $\{|e^{(y^* + s(y(t)-y^*))(\sigma,\cdot)}|_{L^4(\Omega)}: s \in [0,1], \sigma, t \in [0,T]\}$ is bounded. Since $u \in U(u^*)$ is chosen arbitrarily, the last equation in (3.23) holds as well. □

We next consider a variation of problem (3.1)–(3.2) aimed at eliminating the regularity condition $e^{y^*} \in L^2(Q)$ in Theorem 3.6. This can be achieved at the expense of introducing the bound

$$\int_\Omega |\nabla y(t,\cdot)|^2 dx \le M \text{ a.e. } t \in [0,T],$$

for some $M > 0$, to (3.1)–(3.2). This implies that the optimal state is necessarily an element of $W^{2,1}$. This motivates us to consider

$$\min J(y,u) \tag{3.26}$$

subject to $(y,u) \in W^{2,1} \times L^2(Q)$, and

$$\begin{cases} (y,u) \text{ satisfies (3.2) with} \\ \int_\Omega |\nabla y(t,\cdot)|^2 dx \le M \text{ for almost every } t \in [0,T]. \end{cases} \tag{3.27}$$

Here $n = 2, M > 0$ is fixed, and $\varphi \in H_0^1(\Omega)$ with $|\nabla\varphi|_{L_n^2} \le M$. Let (y^*, u^*) denote a solution to (3.26)–(3.27), which is readily shown to exist if $J: W^{2,1} \times L^2(Q)$ is bounded below, weakly lower semicontinuous, and radially unbounded in the sense that $J(y,u) \to \infty$ if $|y|_{W^{2,1}} \to \infty$ or $|u|_{L^2} \to \infty$. We require

$$\begin{cases} J: W^{2,1} \to \mathbb{R} \text{ is continuously Fréchet-differentiable in a} \\ \text{neighborhood } U(y^*, u^*) \subset W^{2,1} \times L^2(Q) \text{ of } (y^*, u^*). \end{cases} \tag{A7}$$

THEOREM 3.7. *If (y^*, u^*) is a solution to* (3.26)–(3.27) *and* (A7) *is satisfied, then there exists $\lambda^* \in L^2(Q)$ such that* (3.23)(i), (ii), *and* (v) *hold, while* (iii) *and* (iv) *are replaced by*

$$J_y(y^*, u^*)(y - y^*) + (\lambda^*, (y - y^*)_t - \Delta(y - y^*) - \delta e^{y^*}(y - y^*))_{L^2(Q)} \ge 0 \tag{3.28}$$

for all $y \in W^{2,1}$ with $\int_\Omega |\nabla y(t,\cdot)|^2 \le M$ for almost every $t \in [0,T]$, and $y = y(0,\cdot) = \varphi$.

Proof. The result follows from an abstract theorem on the existence of Lagrange multipliers for optimization problems [ZK]. Here a Lagrange multiplier is only introduced for the constraint $g: W^{2,1} \times L^2(Q) \to L^2(Q)$, given by

$$g(y,u) = y_t - \Delta y - \delta e^y - u,$$

while $y(0,\cdot) = \varphi$ and $|\nabla y(t,\cdot)|^2_{L^2_n} \leq M$, for almost every $t \in [0,T]$, are treated as explicit constraints. Since $W^{2,1}$ is continuously embedded in $C(0,T;H^1_0(\Omega))$, it follows that $\{y \in W^{2,1}: y(0,\cdot) = \varphi, |\nabla y(t,\cdot)|^2_{L^2_n} \leq M \text{ a.e. } t \in [0,T]\}$ is a closed convex subset of $W^{2,1}$. Moreover $g_u: L^2(Q) \to L^2(Q)$ is surjective, and hence (3.28) follows from Theorem 3.3 in [ZK]. □

As a direct consequence of Theorem 3.6, we observe that if (y^*, u^*) is a solution to (3.26)–(3.27) such that $|\nabla y^*(t,\cdot)|^2 < M$ for all t and (A7) holds, then (y^*, u^*) satisfies the optimality system (3.23).

Let us close the section with a remark on boundary control problems.

Remark 3.8. The technique of Theorem 3.3 is also applicable to the treatment of certain boundary control problems. Let us consider

$$\min J(y, y(T,\cdot), u) \tag{3.29}$$

subject to $(y, y(T,\cdot), u) \in \widetilde{W} \times L^2(\Omega) \times L^2(\Sigma)$ and

$$\begin{cases} y_t = \Delta y + \delta e^y & \text{in } Q, \\ \frac{\partial y}{\partial n} = u & \text{on } \Sigma, \\ y(0,\cdot) = \varphi & \text{in } \Omega, \end{cases} \tag{3.30}$$

where

$$\widetilde{W} = \{y \in L^2(0,T;H^1(\Omega)): y_t \in L^1(0,T;X_1^*)\}.$$

For $u \in L^2(\Sigma)$ a function y is called a solution to (3.30) if $e^y \in L^1(Q)$ and

$$\begin{cases} (y_t, v)_{L^1(X_1^*),L^\infty(X_1)} + (\nabla y, \nabla v)_{L^2(L^2_n)} = \delta(e^y, v)_{L^1(Q),L^\infty(Q)} \\ \qquad\qquad + (u, v)_{L^2(\Sigma)} \quad \text{for all } v \in L^\infty(0,T;X_1), \\ y(0,\cdot) = \varphi. \end{cases} \tag{3.31}$$

We introduce the condition

$$\begin{cases} J: L^2(0,T;H^1(\Omega)) \times L^2(\Omega) \times L^2(\Sigma) \to \mathbb{R} \text{ is bounded from below,} \\ \text{weakly lower semicontinuous, and coercive in the sense that} \\ J(y, y(T,\cdot), u) \to \infty \text{ if } |y|_{L^2(0,T;H^1)} \to \infty \text{ or } |y(T,\cdot)|_{L^2(\Omega)} \to \infty \text{ or} \\ |u|_{L^2(\Sigma)} \to \infty. \end{cases} \tag{A8}$$

If (A8) holds and there exists $u \in L^2(\Sigma)$ such that (3.30) admits a solution $y(u)$, then there exists a solution $(y^*, y^*(T,\cdot), u^*) \in \widetilde{W} \times L^2(\Omega) \times L^1(\Sigma)$. In fact, one first argues, in a manner similar to that used for Proposition 3.2, that every solution y to (3.30) satisfies

$$\begin{aligned} &\delta \textstyle\int_Q e^y y dQ + \frac{1}{2}|\varphi|^2_{L^2(\Omega)} \\ &\leq \tfrac{1}{2}|y(T,\cdot)|^2_{L^2(\Omega)} + |\nabla y|^2_{L^2(L^2_n)} + \tfrac{1}{2}|u|^2_{L^2(\Sigma)} + \tfrac{K}{2}|y|^2_{L^2(H^1)}, \end{aligned} \tag{3.32}$$

where K is the embedding constant of $L^2(\partial\Omega)$ into $H^1(\Omega)$. The existence proof is then quite similar to that of Theorem 3.3, with $H^1_0(\Omega)$ and W replaced by $H^1(\Omega)$ and $\widetilde{W}$. Formally it simple to derive a first-order optimality condition for (3.29)–(3.30). The detailed analysis of optimality conditions for boundary control problems is not within the scope of this paper, however.

Appendix. *Proof of Proposition* 3.2. Let $\{S_\epsilon : \epsilon > 0\}$ denote the semigroup generated by the Laplacian with homogenous Dirichlet boundary conditions in $L^1(\Omega)$. Recall that $S_\epsilon(L^1(\Omega)) \subset L^\infty(\Omega)$ for every $\epsilon > 0$. The restriction of S_ϵ to $L^2(\Omega)$ will be denoted by S_ϵ as well. S_ϵ is used as a spatial smoothing operation. For the temporal smoothing we use

$$T_\epsilon : L^1(0,T;X_1^*) \to W^{1,1}(0,T;X_1^*)$$

given by

$$T_\epsilon v(t) = \frac{1}{\epsilon}\int_t^{t+\epsilon} v(s)\,ds,$$

where v is extended by 0 for $s > T$. Applying $T_\epsilon S_\epsilon$ to (3.4), we obtain for every $\epsilon > 0$

$$(T_\epsilon S_\epsilon y)_t = \Delta(T_\epsilon S_\epsilon y) + \delta T_\epsilon S_\epsilon e^y + T_\epsilon S_\epsilon u, \tag{A.1}$$

where $T_\epsilon S_\epsilon y \in W^{1,2}(0,T;H_0^1 \cap L^\infty)$. In fact, $y \in L^2(0,T;H_0^1), S_\epsilon y \in L^2(0,T;H_0^1 \cap L^\infty)$, and hence $T_\epsilon S_\epsilon y \in W^{1,2}(0,T;H_0^1 \cap L^\infty)$. In addition we have $\Delta(T_\epsilon S_\epsilon y) \in L^2(0,T;L^2), T_\epsilon S_\epsilon u \in L^2(0,T;L^2)$, and $T_\epsilon S_\epsilon e^y \in L^\infty(Q)$. Set $z_\epsilon = T_\epsilon S_\epsilon y \in L^\infty(Q)$ and note that z_ϵ can be identified a.e. with an element of $H^1(Q)$. Since the positive part z_ϵ^+ of z_ϵ is an element of $H^1(Q)$ and $L^2(0,T;H_0^1)$ as well, we can integrate (A.1) against z^+ on Q. Let us first consider

$$\int_Q z_\epsilon z_{\epsilon,t}^+ = \int_\Omega \int_0^T z_\epsilon z_{\epsilon,t}^+ = \frac{1}{2}\int_\Omega |z_\epsilon^+(T,\cdot)|^2 - \frac{1}{2}\int_\Omega |z_\epsilon^+(0,\cdot)|^2.$$

We thus obtain from (A.1)

$$\begin{aligned}
&\delta \textstyle\int_Q T_\epsilon S_\epsilon e^y (T_\epsilon S_\epsilon y)^+ dx\,dt + \frac{1}{2}\int_\Omega |z_\epsilon^+(0,\cdot)|^2 dx \\
&\le \tfrac{1}{2}\textstyle\int_\Omega |z_\epsilon^+(T,\cdot)|^2 dx + \int_0^T (\nabla z_\epsilon, \nabla z_\epsilon^+)_{L_n^2(\Omega)} dt \\
&\qquad - \textstyle\int_0^T (T_\epsilon S_\epsilon u, z_\epsilon^+)_{L^2(\Omega)} dt.
\end{aligned}$$

Applying Fatou's lemma, we obtain

$$\begin{aligned}
&\delta \textstyle\int_Q e^y y^+ dx\,dt + \frac{1}{2}\int_\Omega |\varphi^+|^2 dx \\
&\le \liminf_{\epsilon \to 0^+} \Big[\tfrac{1}{2}\textstyle\int_\Omega |z_\epsilon^+(T,\cdot)|^2 dx + \int_0^T (\nabla z_\epsilon, \nabla z_\epsilon^+)_{L_n^2(\Omega)} dt \\
&\qquad - \textstyle\int_0^T (T_\epsilon S_\epsilon u, z_\epsilon^+)_{L^2(\Omega)} dt \Big].
\end{aligned} \tag{A.2}$$

Next Lebesque's bounded convergence theorem is applied to the right-hand side of (A.2). It is applicable, since S_ϵ is a contraction semigroup and T_ϵ is a contraction as well. To pass to the limit in the expression $(\nabla z_\epsilon(t,\cdot), \nabla z_\epsilon^+(t,\cdot))_{L_n^2(\Omega)}$ for fixed $t > 0$, we recall that the mapping $u \to u^+$ is continuous from $H^1(\Omega)$ to itself (see [T, p. 79]). Hence we have

$$\begin{aligned}
&\delta \textstyle\int_Q e^y y\,dx\,dt + \frac{1}{2}\int_\Omega |\varphi^+|^2 dx \\
&\le \tfrac{1}{2}\textstyle\int_\Omega |y^+(T,\cdot)|^2_{L^2(\Omega)} dx + \int_0^T |\nabla y^+|^2_{L_n^2(\Omega)} dt - \int_0^T (u, y^+)_{L^2(\Omega)} dt.
\end{aligned} \tag{A.3}$$

The analogous estimate holds for y^-, and hence (3.5) follows. ∎

REFERENCES

[Ba] V. Barbu, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Noordhoff, Leyden, The Netherlands, 1976.

[Br] H. Brezis, *Monotonicity methods in Hilbert spaces and some applications to nonlinear partial differential equations*, in Contributions to Nonlinear Functional Analysis, E. H. Zarantonello, ed., Academic Press, New York, 1971, pp. 101–156.

[BE] J. Bebernes and D. Eberly, *Mathematical Problems from Combustion Theory*, Springer-Verlag, Berlin, Germany, 1989.

[BK] A. Borzi and K. Kunisch, *The numerical solution of the steady state solid fuel ignition model and its optimal control*, SIAM J. Sci. Comp., 22 (2000) pp. 263–284.

[CF] P. Constantin and C. Foias, *Navier-Stokes Equations*, The University of Chicago Press, Chicago, IL, 1988.

[CKP] E. Casas, O. Kavian, and J.-P. Puel, *Optimal control of ill-posed elliptic semilinear equation with an exponential nonlinearity*, ESAIM Control Optim. Calc. Var., 3 (1998), pp. 361–380.

[F] H. Fujita, *On the nonlinear equation* $\Delta u + e^u = 0$ *and* $\frac{\partial u}{\partial t} = \Delta u + e^u$, Bull. Amer. Math. Soc., 75 (1969), pp. 132–135.

[FK] D. A. Franck-Kamenetskii, *Diffusion and Heat Transfer in Chemical Kinetics*, Plenum Press, New York, 1969.

[GH] R. Glowinski and J.-W. He, *On Control Problems for Some Advection-Reaction-Diffusion Systems*, University of Houston, Houston, TX, preprint.

[GT] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, Germany, 1977.

[IK] K. Ito and K. Kunisch, *Newton's method for a class of weakly singular optimal control problems*, SIAM J. Control Optim., 109 (1999), pp. 896–916.

[KK] A. Kauffmann and K. Kunisch, *Optimal control of the solid fuel ignition model*, in Contrôle des systèmes gouvernés par des équations aux dérivées partielles, ESIAM Proc. 8, Soc. Math. Appl. Indust., Paris, 2000, pp. 65–76.

[L] J. L. Lions, *Control of Distributed Singular Systems*, Gauthier–Villars, Paris, France, 1985.

[T] G. T. Troianello, *Elliptic Differential Equations and Obstacle Problems*, Plenum Press, New York, 1987.

[ZK] J. Zowe and S. Kurcyusz, *Regularity and stability for the mathematical programming problem in Banach spaces*, Appl. Math. Optim., 5 (1979), pp. 49–62.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 5, pp. 1473–1490

VARIATIONAL PROBLEMS WITH NONCONVEX, NONCOERCIVE, HIGHLY DISCONTINUOUS INTEGRANDS: CHARACTERIZATION AND EXISTENCE OF MINIMIZERS*

CRISTINA MARCELLI†

Abstract. We consider the functional $F(v) = \int_a^b f(t, v'(t))dt$ in $\mathcal{H}_p = \{v \in W^{1,p} : v(a) = 0, v(b) = d\}$, $p \in [1, +\infty]$. Under only the assumption that the integrand is $\mathcal{L} \otimes \mathcal{B}_n$-measurable, we prove characterizations of strong and weak minimizers both in terms of the minimizers of the relaxed functional and by means of the Euler–Lagrange inclusion.

As an application, we provide necessary and sufficient conditions for the existence of the minimum, expressed in terms of a limitation on the width of the slope d.

Key words. strong and weak minimizers, Euler–Lagrange condition, convexification, subdifferential

AMS subject classifications. 49K05, 49J05

PII. S036301299936141X

1. Introduction.

The study of nonsmooth and nonconvex variational problems has been widely developed in recent years, with regard to both necessary conditions for optimality and sufficient conditions for existence of minimizers. The development of nonsmooth analysis has allowed researchers to deal with nonregular problems. In particular, the introduction of various types of subdifferentials that are sharper and sharper, starting with Clarke's and continuing with the more refined Mordukhovich's subdifferential, permitted the achievement of nonsmooth versions of the classical Euler–Lagrange condition (see [7], [8], [12], [13], [21], [22], [23], [24] [25], [31]). In particular, in the paper by Ioffe and Rockafellar [13], which deals with nonsmooth integrands $f = f(t, v(t), v'(t))$ everywhere finite, the Euler–Lagrange condition is proved for weak local minimizers (w.l.m.) by means of this last type of subdifferential.

In the same paper the authors also prove the Weierstrass condition for minimizers in the $W^{1,1}$-norm (see [13, Theorem 1]), which are usually called *intermediate minimizers*. Of course, this condition can not hold in general for w.l.m., since it has a global character.

In this paper we consider the optimization problem

$$\text{minimize} \quad F(v) = \int_a^b f(t, v'(t))dt$$

in the class

$$\mathcal{H}_p = \left\{ v \in W^{1,p}([a,b]; \mathbb{R}^n) : \int_a^b f(t, v'(t))dt \text{ is well-defined in } \tilde{\mathbb{R}}, v(a) = 0, v(b) = d \right\}$$

for $p \in [1, +\infty]$, where $\tilde{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$ and the integrand $f : [a,b] \times \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is assumed to be only $\mathcal{L} \bigotimes \mathcal{B}_n$-measurable, hence, in general, neither convex nor coercive nor continuous.

*Received by the editors September 9, 1999; accepted for publication (in revised form) June 8, 2001; published electronically January 18, 2002.
http://www.siam.org/journals/sicon/40-5/36141.html

†Dipartimento di Matematica "V. Volterra," Universitá di Ancona, Via Brecce Bianche, 60131 Ancona, Italy (marcelli@dipmat.unian.it).

The first aim of this paper is to provide, under these minimal assumptions, a localized version of the Weierstrass condition for w.l.m., which allows us to obtain a Euler–Lagrange inclusion expressed in terms of a localized version of the subdifferential in the sense of convex analysis. This type of subgradient, which in general is strictly contained in the other types of subdifferential, is the best one in this context in which the integrand is of the type $f = f(t, v'(t))$. Indeed, the present version of the Euler–Lagrange condition also provides a sufficient condition for optimality.

In order to provide more detail, our first main result (see Theorem 3.2) is a characterization of $W^{1,p}$-strong local minimizers, that is, functions $u_0 \in W^{1,1}$ for which there exists $\delta > 0$ such that $F(u_0) \leq F(v)$ for every $v \in u_0 + W_0^{1,p}$ with $|v(t) - u_0(t)| < \delta$ for every $t \in [a, b]$. In particular, we prove that $u_0 \in \mathcal{H}_1$ is a $W^{1,p}$-strong local minimizer if and only if $F(u_0) = \min_{v \in \mathcal{H}_p} \int_a^b f^{**}(t, v'(t))dt$, and this condition is equivalent to the Euler–Lagrange inclusion $c \in \partial f(t, u_0'(t))$ a.e. in $[a, b]$, where $\partial f(t, \cdot)$ denotes the usual subgradient in the sense of convex analysis.

As a consequence, note that if u_0 is a $W^{1,p}$-strong local minimizer, then it is also a global minimizer. Moreover, when the minimum exists, the integrand f coincides with its convex envelope along the minimizer, even if it is nonconvex and highly discontinuous. Furthermore, the Euler–Lagrange inclusion involves the usual subgradient in the sense of convex analysis, even if the integrand is not convex in general. Finally, this characterization extends a result in [1] where the Euler–Lagrange inclusion is obtained for nonconvex integrands, but under a suitable growth assumption.

We also obtain a similar characterization for w.l.m., that is, functions u_0 such that $F(u_0) \leq F(v)$ for every $v \in u_0 + W_0^{1,p}$ with $|v'(t) - u_0'(t)| < \epsilon$, for almost every $t \in [a, b]$, for some $\epsilon > 0$, by means of a localization of the subgradient (see Corollary 3.4).

We wish to underline that, with regard to the regularity of the integrand, we only assume the measurability with respect to the last variable, contrary to most of the papers on this subject, in which lower semicontinuity and limitations by summable functions are needed, such as in [13], [17]. In particular, in [17] a version of the Euler–Lagrange condition in terms of the subdifferential of convex analysis is obtained for complete integrands $f = f(t, v(t), v'(t))$, everywhere finite. In the present paper, we limit ourselves to dealing with integrands of the type $f = f(t, v'(t))$, taking values on $\mathbb{R} \cup \{+\infty\}$, for which we require weaker regularity assumptions.

As a consequence of these results, we are able to prove a general existence theorem for nonconvex problems, which will allow us to get many applications concerning the existence of minimizers in the second part of the paper.

In more detail, denoting by $C_t = \{z \in \mathbb{R}^n : f(t, z) = f^{**}(t, z)\}$ the set where $f(t, \cdot)$ coincides with its convex envelope, we prove (see Theorem 3.5) that the functional F admits a minimum in $\mathcal{H}_p$ if and only if there exists $u_0 \in \mathcal{H}_p$ such that $\int_a^b f^{**}(t, u_0'(t))dt = \min_{v \in \mathcal{H}_p} \int_a^b f^{**}(t, v'(t))dt$ and $u_0'(t) \in co(C_t)$ a.e. in $[a, b]$. By applying this result, we are able to give very operative necessary and sufficient conditions for the existence of the minimum in the scalar case ($n = 1$).

We recall that in [18] Marcellini proved the existence of the minimum for the functional F under the assumption that $f(t, \cdot)$ is coercive but not necessarily convex. As is well known, when the integrand is not coercive the functional F could admit no minimum. A classical example is provided by the weighed length functional, whose integrand has the form $f(t, z) = \phi(t)\sqrt{1 + z^2}$. For such a functional, Kaiser proved in [14] that the existence of the minimum depends on the width of the slope d. More precisely, he obtained a necessary and sufficient condition for the existence of the

minimum, expressed in terms of a limitation on the slope d.

This result was extended by Brandi in [4] to the case of more general convex integrands. Subsequently in [15] a necessary and sufficient condition for the existence of the minimum of the functional F with integrand convex but not coercive was established. This condition consists of a limitation on the slope d in such a way that the range of the values of d for which the minimum exists is exactly determined.

Such results take a boundary condition (the width of d) into consideration as a specific parameter of the variational problem. This approach to the existence of the minimum is quite different from the classical *direct method*, which indeed presents well-known limitations in the treatment of noncoercive or nonconvex problems.

The importance of the role played by boundary conditions and other specific parameters of the problem is discussed by Mordukhovich in [26], where *individual existence theorems* are presented in the setting of optimal control problems, together with examples, a survey, and a wide bibliography on this subject.

In this paper we are able to discuss the existence of the minimum related to the parameter d in the case of nonconvex and noncoercive integrands. Indeed, we exactly determine the set $D = \{d \in \mathbb{R} : F(v) \text{ admits minimum}\}$ (see Theorem 4.2).

The result we obtain becomes particularly expressive when the integrand has the structure $f(t,z) = \phi(t)h(z)$. Indeed, set $C = \{z : h(z) = h^{**}(z)\}$, and in general we have $D \subset co(C)$, but strict inclusion may occur, as we show in Example 4.1.

More precisely, for the case in which h is not convex at ∞, i.e., the set C is bounded, we show that if h^{**} is strictly convex in C, then D is strictly contained in $co(C)$. Whereas, when h^{**} is affine in $co(C)$, it can happen that $D = co(C)$. We prove that this situation occurs if and only if a precise relation, which links the minimum and the maximum of the function $\phi(t)$ and the left and right derivatives of h^{**} in $\min C$, $\max C$, is satisfied (see Corollary 4.5).

Finally, in the case when $f^{**}(t,\cdot)$ is strictly convex at ∞ (but not necessarily coercive), we show that the range D is exactly the same as that obtained for the convex case in [15].

With regard to the methodology used in order to prove our main result about the characterization of minimizers, we wish to remark that a key tool is a localized version of classical relaxation theorems. Specifically, set $S_p(u_0,\epsilon) = \{v \in u_0 + W_0^{1,p} : |u_0(t) - v(t)| < \epsilon \text{ for every } t \in [a,b]\}$; we prove (see Theorem 2.1)

$$\inf_{v \in S_p(u_0,\epsilon)} F(v) = \inf_{v \in S_q(u_0,\epsilon)} \int_a^b f^{**}(t, v'(t))dt \qquad \text{for every } \ \epsilon > 0, \quad p, q \in [1, +\infty],$$

from which one deduces, as a particular case, $\inf_{v \in \mathcal{H}_p} F(v) = \inf_{v \in \mathcal{H}_p} \int_a^b f^{**}(t, v'(t))dt$.

2. Notations and relaxation theorem. Let $f : [a,b] \times \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a given $\mathcal{L} \bigotimes \mathcal{B}_n$-measurable function, where $\mathcal{L}$, $\mathcal{B}_n$ denote, respectively, the Lebesgue and Borel σ-fields in $[a,b]$ and $\mathbb{R}^n$. As usual, we denote by $\mathrm{dom} f = \{(t,z) : f(t,z) < +\infty\}$ and by $\mathrm{dom} f(t,\cdot) = \{z \in \mathbb{R}^n : f(t,z) < +\infty\}$.

Let $f^{**}(t,\cdot)$ and $\mathcal{C}f(t,\cdot)$ be the bipolar function and the convex envelope of the function $f(t,\cdot)$, respectively. As is well known (see, e.g., [29, Corollary 17.1.5]), we have

$$\mathcal{C}f(t,z) = \inf \left\{ \sum_{j=1}^{n+1} \lambda_j f(t,\xi_j) : \lambda_j \in [0,1], \ \sum_{j=1}^{n+1} \lambda_j = 1, \ \sum_{j=1}^{n+1} \lambda_j \xi_j = z \right\}. \tag{2.1}$$

Moreover, f^{**} is convex and lower semicontinuous; hence we have

$$f^{**}(t,\cdot) \le \mathcal{C}f(t,\cdot) \le f(t,\cdot), \tag{2.2}$$

and if dom $f(t,\cdot) = \mathbb{R}^n$, then we have $f^{**} \equiv \mathcal{C}f$ (see, e.g., [10, Theorem 2.2.5]). But note that in general we have

$$\mathcal{C}f(t,z) = f^{**}(t,z) \quad \text{for every } z \in \text{int}(\text{co}(\text{dom}f(t,\cdot))).$$

Of course, dom $f^{**}(t,\cdot) = \overline{co}(\text{dom}f(t,\cdot))$, where $\overline{co}(\cdot)$ denotes the closure of the convex hull.

We denote by $\partial f(t,z)$ the subgradient of $f(t,\cdot)$ at the point z in the sense of convex analysis, i.e.,

$$\partial f(t,z) = \{\zeta \in \mathbb{R}^n \ : \ f(t,w) \ge f(t,z) + \langle \zeta, w - z\rangle \ \text{ for every } \ w \in \mathbb{R}^n\}.$$

In what follows we consider for $p \in [1,+\infty]$ the classes

$$\mathcal{H}_p = \left\{ v \in W^{1,p} \ : \ \int_a^b f(t,v'(t))dt \text{ is well-defined in } \tilde{\mathbb{R}}, v(a) = 0, v(b) = d \right\}.$$

Let $F : \mathcal{H}_1 \to \tilde{\mathbb{R}}$ be the functional defined by $F(v) = \int_a^b f(t,v'(t))dt$.

Let $u_0 \in \mathcal{H}_1$ be given. For every $\epsilon > 0$, we denote by

$$S_p(u_0,\epsilon) = \{v \in u_0 + W_0^{1,p} : |u_0(t) - v(t)| < \epsilon \quad \text{for every } t \in [a,b]\},$$

$$W_p(u_0,\epsilon) = \{v \in u_0 + W_0^{1,p} : \|u_0 - v\|_{W_0^{1,p}} < \epsilon\}$$

the strong and weak ϵ-neighborhoods of u_0.

Recall that a function $u_0 \in W^{1,p}(a,b)$ is said to be a *(global) minimizer* for the functional F in $\mathcal{H}_p$ if $F(u_0) \le F(v)$ for every $v \in \mathcal{H}_p$, whereas it is said to be a $W^{1,q}$*-strong* (respectively, $W^{1,q}$*-weak*) *local minimizer*, with $q \ge p$, if a constant $\epsilon > 0$ exists such that $F(u_0) \le F(v)$ for every $v \in S_q(u_0,\epsilon)$ (respectively, $v \in W_q(u_0,\epsilon)$).

Functions u_0, which are $W^{1,\infty}$-weak local minimizers, are usually simply called *w.l.m.* In contrast, $W^{1,p}$-weak local minimizers for $p < +\infty$ are called *intermediate local minimizers*. In [25], [30] the relation between strong, intermediate, and weak local minimizers is discussed, showing that they are different concepts in general.

In the present paper we characterize strong and weak local minimizers.

The following theorem is a relaxation-type result for infima taken over strong neighborhoods.

THEOREM 2.1. *Let* $f : [a,b] \times \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ *be a* $\mathcal{L} \bigotimes \mathcal{B}_n$*-measurable function such that* dom$f(t,\cdot)$ *is a convex set for almost every* $t \in [a,b]$.

Assume that some $u_0 \in W^{1,1}(a,b)$ *satisfies both* $F(u_0) < +\infty$ *and*

$$u_0'(t) \in \text{int } (\text{dom}f(t,\cdot)) \qquad \text{a.e. in } [a,b]. \tag{2.3}$$

Then, for every $p,q \in [1,+\infty]$ *and every* $\epsilon > 0$,

$$\inf_{v \in S_p(u_0,\epsilon)} F(v) = \inf_{v \in S_q(u_0,\epsilon)} \int_a^b f^{**}(t,v'(t))dt.$$

In particular,

$$\inf_{v\in\mathcal{H}_p} F(v) = \inf_{v\in\mathcal{H}_p} \int_a^b f^{**}(t, v'(t))dt.$$

Let us now recall a classical Lyapunov-type result, whose proof can be found in [6, Theorem 16.1.v], which will play a fundamental role in the proof of Theorem 2.1.

LYAPUNOV THEOREM. *Let $g_j : A \to \mathbb{R}^m$, $j = 1, \dots, h$, be L-integrable functions on a set $A \subset \mathbb{R}$ with finite measure, and let $\lambda_j : A \to [0, 1]$, $j = 1, \dots, h$, be measurable weight functions with $\sum_{j=1}^h \lambda_j(t) = 1$.*

Then, there exists a decomposition $E_1, \dots, E_h$ of A into disjoint measurable subsets such that

$$\sum_{j=1}^h \int_{E_j} g_j(t)dt = \int_A \sum_{j=1}^h \lambda_j(t) g_j(t)dt.$$

Proof of Theorem 2.1. Taking equation (2.2) into account, it suffices to show that $\inf_{v\in S_p(u_0,\epsilon)} F(v) \le \inf_{v\in S_q(u_0,\epsilon)} \int_a^b f^{**}(t, v'(t))dt$ for every $\epsilon > 0$ and every $p, q \in [1, +\infty]$.

Let us assume, for contradiction, that for some $\epsilon > 0$ and some $p, q \in [1, +\infty]$ there exists $w \in S_q(u_0, \epsilon)$ such that

$$\int_a^b f^{**}(t, w'(t))dt < \inf_{v\in S_p(u_0,\epsilon)} F(v).$$

Let α, β be two real numbers such that

$$\int_a^b f^{**}(t, w'(t))dt < \alpha < \beta < \inf_{v\in S_p(u_0,\epsilon)} F(v).$$

Let $\delta \in]0, 1/2[$ be such that $B(w(t), \delta) \subset B(u_0(t), \epsilon)$ for every $t \in [0, 1]$ and

$$(1-\delta)\int_a^b f^{**}(t, w'(t))dt + \delta \int_a^b f^{**}(t, u_0'(t))dt < \alpha,$$

with obvious meaning in the case $\int_a^b f^{**}(t, w'(t))dt = -\infty$ or $\int_a^b f^{**}(t, u_0'(t))dt = -\infty$. (Notice that both of the previous integrals are less than $+\infty$.)

For $\tilde{w}(t) := (1-\delta)w(t) + \delta u_0(t)$, we have that $B(\tilde{w}(t), \delta) \subset B(u_0(t), \epsilon)$ for every $t \in [a, b]$, and by the convexity of the function f^{**} we deduce that

$$\int_a^b f^{**}(t, \tilde{w}'(t))dt \le (1-\delta)\int_a^b f^{**}(t, w'(t))dt + \delta \int_a^b f^{**}(t, u_0'(t))dt < \alpha.$$

Moreover, since $w'(t) \in \mathrm{dom} f^{**}(t, \cdot)$ for almost all t, by virtue of the convexity of $\mathrm{dom} f(t, \cdot)$ and by (2.3) we have that $\tilde{w}'(t) \in \mathrm{int}(\mathrm{dom} f(t, \cdot))$ a.e. in $[a, b]$. Hence,

$$f^{**}(t, \tilde{w}'(t)) = \mathcal{C}f(t, \tilde{w}'(t)) \quad \text{for a.a. (almost all) } t \in [a, b]. \tag{2.4}$$

Let $g : [a, b] \to \mathbb{R}$ be a summable function such that $g(t) > f^{**}(t, \tilde{w}'(t))$ and

$$\int_a^b g(t)dt < \alpha.$$

Combining (2.1) with (2.4) shows that the multifunction $Q : t \in [a,b] \mapsto Q(t) \subset [0,1]^{n+1} \times \mathbb{R}^{n(n+1)}$ defined by

$$Q(t) = \left\{ (\lambda, \xi) \ : \ \sum_{j=1}^{n+1} \lambda_j = 1, \sum_{j=1}^{n+1} \lambda_j \xi_j = \tilde{w}'(t), \sum_{j=1}^{n+1} \lambda_j f(t, \xi_j) \leq g(t), \ f(t, \xi_j) < +\infty \right\}$$

has nonempty values and is measurable. By virtue of the Aumann selection theorem [8, Theorem 7.2.1], there exist measurable functions $\lambda_j : [a,b] \to [0,1]$, $\xi_j : [a,b] \to \mathbb{R}^n$, $j = 1, \dots, n+1$, such that $\sum_{j=1}^{n+1} \lambda_j(t)\xi_j(t) = \tilde{w}'(t)$ and

$$\sum_{j=1}^{n+1} \lambda_j(t) f(t, \xi_j(t)) \leq g(t) \quad \text{a.e. in } [a,b]. \tag{2.5}$$

Let $\rho > 0$ be a real number such that for every set $E \subset [a,b]$ with $|E| < 2\rho$ we have

$$\int_E (|u_0'(t)| + |\tilde{w}'(t)|)dt < \frac{\delta}{3}, \tag{2.6}$$

$$\int_E |f(t, u_0'(t))|dt < \frac{1}{2}(\beta - \alpha), \tag{2.7}$$

$$\int_{[a,b]\setminus E} g(t)dt < \alpha. \tag{2.8}$$

Let $r \in]0,1[$ be such that for $G = \{t : \overline{B(\tilde{w}'(t), nr)} \subset \mathrm{dom} f(t, \cdot)\}$ we have $|G| > b - a - \rho/2$.

Let $C \subset G$ be a compact set with $(b-a) - \rho < |C| < b - a$, and let $L > 0$ be a constant such that for almost every $t \in C$ and every $\lambda \in \mathbb{R}^n$ with $\lambda_i \in \{\tilde{w}_i'(t), \tilde{w}_i'(t) + r, \tilde{w}_i'(t) - r\}$, $i = 1, \dots, n$, we have

$$|u_0'(t)| + |\tilde{w}'(t)| + \sum_{j=1}^{n+1} |\xi_j(t)| + \sum_{j=1}^{n+1} |f(t, \xi_j(t))| + |f(t, \lambda)| \leq L. \tag{2.9}$$

Put $\sigma := \min\{\frac{\beta - \alpha}{2L}, \rho, |C|, \delta/3n\}$, let $A \subset [a,b] \setminus C$ be a set, and let $N > 0$ be a constant such that we have

$$\sum_{j=1}^{n+1} (|f(t, \xi_j(t))| + |\xi_j(t)|) + |\tilde{w}'(t)| + |u_0'(t)| \leq N \ \text{ for a.a. } t \in [a,b] \setminus (A \cup C), \tag{2.10}$$

$$\int_A |\tilde{w}'(t) - u_0'(t)|dt \leq r\sigma. \tag{2.11}$$

Let $C^* \subset C$ be a set with $|C^*| = \sigma$. Set $B = [a,b] \setminus (A \cup C^*)$ and let $B_1, \dots, B_s$ be a finite partition of B into disjoint measurable subsets such that

$$|B_k| < \rho, \quad \sup B_k \leq \inf B_{k+1}, \qquad k = 1, \dots, s,$$

and

$$\int_{B_k}\left(|\tilde{w}'(t)|+\sum_{j=1}^{n+1}|\xi_j(t)|\right)dt\le\frac{\delta}{3},\quad k=1,\dots,s. \tag{2.12}$$

For every $k\in\{1,\dots,s\}$ we can apply the Lyapunov theorem to the functions $h_j:B_k\to\mathbb{R}^{n+1}$, $j=1,\dots,n+1$, defined by $h_j(t)=(\xi_j(t),f(t,\xi_j(t)))$ to deduce that for every $k=1,\dots,s$ there exist disjoint sets $E_1^k,\dots,E_{n+1}^k\subset B_k$ such that for $\phi(t)=\sum_{k=1}^{s}\sum_{j=1}^{n+1}\chi_{E_j^k}(t)\xi_j(t)$, where $\chi_D(\cdot)$ denotes the characteristic function of the set D, we have

$$\int_{B_k}\phi(t)dt=\int_{B_k}\sum_{j=1}^{n+1}\lambda_j(t)\xi_j(t)dt=\int_{B_k}\tilde{w}'(t)dt,\quad k=1,\dots,s. \tag{2.13}$$

Notice that, taking (2.5), (2.8) into account,

$$\int_B f(t,\phi(t))dt=\int_B\sum_{j=1}^{n+1}\lambda_j(t)f(t,\xi_j(t))dt\le\int_B g(t)dt<\alpha. \tag{2.14}$$

For every $i\in\{1,\dots,n\}$ set $\gamma_i:=\int_A[\tilde{w}_i'(t)-u_{0i}'(t)]dt$. By (2.11) we can choose a set $C_i^*\subset C^*$ such that $|C_i^*|=|\gamma_i|/r$.

Finally, for every $i\in\{1,\dots,n\}$, set

$$\psi_i(t):=\begin{cases}u_{0i}'(t), & t\in A,\\ \phi_i(t), & t\in B,\\ \tilde{w}_i'(t)+r\ \operatorname{sgn}(\gamma_i), & t\in C_i^*,\\ \tilde{w}_i'(t), & t\in C^*\setminus C_i^*,\end{cases}$$

and let $\tilde{v}(t):=\int_a^t\psi(\tau)d\tau$. Note that by (2.9) we have $|f(t,\psi(t))|\le L$ for almost every $t\in C^*$. Moreover, from (2.13) it follows that

$$\int_a^b\tilde{v}_i'(t)dt=\int_A u_{0i}'(t)dt+\int_B\phi_i(t)dt+\int_{C^*}\tilde{w}_i'(t)dt+|\gamma_i|\operatorname{sgn}(\gamma_i)=\int_a^b\tilde{w}_i'(t)dt.$$

Hence, by (2.9) and (2.10) we have $u_0-\tilde{v}\in W_0^{1,\infty}(a,b)$.

Moreover, by (2.6), (2.12), (2.13), for every $t\in[a,b]$ we have

$$|\tilde{v}(t)-\tilde{w}(t)|\le\int_A|u_0'(\tau)-\tilde{w}'(\tau)|d\tau+\int_{B_k}|\phi(\tau)-\tilde{w}'(\tau)|d\tau+n\sigma\le\frac{\delta}{3}+\frac{\delta}{3}+\frac{\delta}{3},$$

where $k\in\{1,\dots,s\}$ is such that $\sup B_{k-1}<t\le\sup B_k$, and $B_0=\{a\}$.

Hence, $\tilde{v}\in S_p(u_0,\epsilon)$. Finally, from (2.7), (2.14) we deduce

$$F(\tilde{v})=\int_a^b f(t,\tilde{v}'(t))dt=\int_A f(t,u_0'(t))dt+\int_B f(t,\phi(t))dt+\int_{C^*}f(t,\psi(t))dt$$

$$<\frac{1}{2}(\beta-\alpha)+\alpha+L\sigma\le\beta<\inf_{v\in S_p(u_0,\epsilon)}F(v),$$

a contradiction. □

Remark 2.2. Assumption (2.3) in Theorem 2.1 can be weakened as follows:

$$\operatorname{meas}\{t:u_0'(t)\in\ \operatorname{int}\ (\operatorname{dom}\ f(t,\cdot))\}>0, \tag{2.3′}$$

provided we replace f^{**} with $\mathcal{C}f$ in the assertion. In fact, in view of the proof, in this case it suffices if we ensure that $C^*\subset G$, and this is possible, provided $|G|>0$.

3. Characterization of minimizers. As a consequence of the relaxation theorem, we now prove a characterization for strong minimizers both in terms of minimizers of the relaxed functional and in terms of the Euler–Lagrange inclusion. This will generalize the following result by Ambrosio, Ascenzi, and Buttazzo [1], which concerns convex integrands.

THEOREM 3.1 (see [1, Theorem 3.1]). *Let $f : [0,1] \times \mathbb{R}^n \to [0,+\infty]$ be an $\mathcal{L} \bigotimes \mathcal{B}_n$-measurable function lower semicontinuous and convex with respect to the last variable.*

Let $u_0 \in W^{1,1}(0,1)$ be a w.l.m. for the functional F such that

$$u_0'(t) \in \text{int } (\text{dom } f(t,\cdot)) \qquad a.e.\ in\ \ [0,1].$$

Then, there exists $c \in \mathbb{R}^n$ with

$$c \in \partial f(t, u_0'(t)) \qquad a.e.\ in\ [0,1].$$

In the next theorem, we consider nonconvex integrands which are only assumed to be $\mathcal{L} \bigotimes \mathcal{B}_n$-measurable. By using Theorem 2.1, we show that the Euler–Lagrange inclusion is a necessary and sufficient condition for the existence of the minimum.

THEOREM 3.2. *Let $f : [a,b] \times \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a given $\mathcal{L} \bigotimes \mathcal{B}_n$-measurable function such that* dom $f(t,\cdot)$ *is a convex set for almost every $t \in [a,b]$.*

Let $u_0 \in \mathcal{H}_1$ be such that $F(u_0) \in \mathbb{R}$ and

$$u_0'(t) \in \text{int } (\text{dom} f(t,\cdot)) \qquad a.e.\ in\ [a,b]. \tag{3.1}$$

Then, the following conditions are equivalent:

(i) *u_0 is a $W^{1,p}$-strong local minimizer for F;*

(ii) *$F(u_0) = \int_a^b f^{**}(t, u_0'(t))dt = \min_{v \in \mathcal{H}_p} \int_a^b f^{**}(t, v'(t))dt$;*

(iii) *a constant $c \in \mathbb{R}^n$ exists such that $c \in \partial f(t, u_0'(t))$ a.e. in $[a,b]$.*

Proof. The implication (i) $\Rightarrow$ (ii) is an immediate consequence of Theorem 2.1, taking the convexity of f^{**} into account.

Let us now prove the implication (ii) $\Rightarrow$ (iii). Set

$$\tilde{f}(t,z) := \max\{f^{**}(t,z), f^{**}(t,u_0'(t)) - 1\} - [f^{**}(t,u_0'(t)) - 1]$$

and $\tilde{F}(v) = \int_a^b \tilde{f}(t, v'(t))dt$.

Of course $\tilde{f}$ is a nonnegative function, convex and lower semicontinuous in the second argument. Moreover, since $\tilde{f}(t,z) \geq f^{**}(t,z) - f^{**}(t,u_0'(t)) + 1$, it is easy to see that $\tilde{F}(u_0) = \min_{v \in \mathcal{H}_p} \tilde{F}(v)$. Hence, we can apply Theorem 3.1 to deduce that a constant $c \in \mathbb{R}^n$ exists such that $c \in \partial \tilde{f}(t, u_0'(t))$ a.e. in $[a,b]$, i.e.,

$$\max\{f^{**}(t,z), f^{**}(t,u_0'(t)) - 1\} \geq f^{**}(t,u_0'(t)) + \langle c, z - u_0'(t)\rangle, \quad z \in \mathbb{R}^n, \text{ a.e. } t \in [a,b].$$

Therefore, for almost every $t \in [a,b]$ there exists a real number $\rho > 0$ such that

$$f^{**}(t,z) \geq f^{**}(t,u_0'(t)) + \langle c, z - u_0'(t)\rangle \qquad \text{for every } z \in B(u_0'(t), \rho);$$

hence, taking the convexity of f^{**} into account, we deduce that $c \in \partial f^{**}(t, u_0'(t))$ a.e. in $[a,b]$. Now, the assertion follows immediately, since

$$f(t,z) \geq f^{**}(t,z) \geq f^{**}(t,u_0'(t)) + \langle c, z - u_0'(t)\rangle = f(t,u_0'(t)) + \langle c, z - u_0'(t)\rangle.$$

(iii) $\Rightarrow$ (i). For every $v \in \mathcal{H}_p$, the subgradient inequality gives

$$\int_a^b f(t,v'(t))dt \geq \int_a^b f(t,u_0'(t))dt + \int_a^b \langle c, v'(t) - u_0'(t)\rangle dt = \int_a^b f(t,u_0'(t))dt. \qquad \square$$

Remark 3.3. The previous theorem also improves an analogous result established in [1] (see Proposition 3.6) for nonconvex integrands f which are $\mathcal{L}\bigotimes\mathcal{B}_n$-measurable, but under the following additional hypothesis:

$$(*) \qquad |z|^p \leq f(t,z) \leq K(1+|z|^p) \quad \text{for some } p > 1.$$

Now we will apply Theorem 3.2 to obtain an analogous equivalence result for w.l.m. as well.

For a given subset $S \subset \mathbb{R}^n$, let $\delta_S(\cdot)$ be the indicator function; that is, $\delta_S(z) = 0$ if $z \in S$, $\delta_S(z) = +\infty$ otherwise. If u_0 is a w.l.m. for functional F with respect to the weak ϵ-neighborhood $W_\infty(u_0,\epsilon)$, set

$$\tilde{f}_\epsilon(t,z) := f(t,z) + \delta_{B(u_0'(t),\epsilon)}(z) \quad \text{and} \quad \tilde{F}_\epsilon(u) := \int_a^b \tilde{f}_\epsilon(t,u'(t))dt;$$

then u_0 is a global minimizer for functional $\tilde{F}_\epsilon$. Hence, we have the following result.

COROLLARY 3.4. *Under the same assumptions of Theorem* 3.2, *the function* $u_0 \in \mathcal{H}_1$ *is a w.l.m. for the functional* F *in the weak* ϵ*-neighborhood* $W_\infty(u_0,\epsilon)$ *if and only if a constant* $c \in \mathbb{R}^n$ *exists such that for almost every* $t \in [a,b]$

$$c \in \partial\tilde{f}_\epsilon(t,u_0'(t)) = \{\zeta : f(t,w) \geq f(t,u_0'(t)) + \langle \zeta, w - u_0'(t)\rangle \text{ for every } w \in B(u_0'(t),\epsilon)\}.$$

Now we wish to note that, as an application of Theorem 3.2, the following result holds, which can be seen to be an extension to the case of integrands $f(t,z)$ of a classical result known for integrands of the type $f = f(z)$ (see [10, Theorem 5.2.6]).

THEOREM 3.5. *Let* $f : [a,b] \times \mathbb{R}^n \to [0,+\infty[$ *be an* $\mathcal{L}\bigotimes\mathcal{B}_n$*-measurable integrand. Denote by*

$$C_t = \{z \in \mathbb{R}^n : f(t,z) = f^{**}(t,z)\}$$

the set where $f(t,\cdot)$ *coincides with its convex envelope, and assume that there exists a function* $m \in L^p(a,b)$ *such that*

$$(3.2) \qquad Bd(C_t) \subset B(0,m(t)) \quad \text{for almost every } t \in [a,b],$$

where $Bd(C_t)$ *denotes the boundary of the set* C_t.

Then, the functional F *admits a minimum in* $\mathcal{H}_p$ *if and only if there exists* $u_0 \in \mathcal{H}_p$ *such that* $\int_a^b f^{**}(t,u_0'(t))dt = \min_{v\in\mathcal{H}_p}\int_a^b f^{**}(t,v'(t))dt$ *and*

$$(3.3) \qquad u_0'(t) \in co(C_t) \text{ a.e. in } [a,b].$$

Proof. The necessary part is contained in Theorem 3.2.

As regards the sufficient part, by (3.3) there exist measurable functions $\lambda_j : [a,b] \to [0,1]$, $\xi_j : [a,b] \to \mathbb{R}^n$, with $\xi_j(t) \in C_t$, $j = 1,\dots,n+1$, such that $\sum_{j=1}^{n+1}\lambda_j(t)\xi_j(t) = u_0'(t)$ and $f^{**}(t,\cdot)$ is affine in $co(\{\xi_j(t), j = 1,\dots,n+1\})$, a.e. in $[a,b]$. By applying the Lyapunov theorem to the functions $g_j(t) = (\xi_j(t), f^{**}(t,\xi_j(t)))$,

$j = 1, \dots, n+1$, we deduce that there exist measurable disjoint sets $E_1, \dots, E_{n+1}$ such that for $\phi(t) = \sum_{j=1}^{n+1} \chi_{E_j}(t)\xi_j(t)$ we have $\int_a^b \phi(t)dt = \int_a^b u_0'(t)dt$ and

$$\int_a^b f(t,\phi(t))dt = \sum_{j=1}^{n+1} \int_{E_j} f^{**}(t,\xi_j(t))dt = \int_a^b \sum_{j=1}^{n+1} \lambda_j(t) f^{**}(t,\xi_j(t))dt$$
$$= \int_a^b f^{**}(t,u_0'(t))dt.$$

Then the assertion follows, since by (3.2) we have $\phi(t) \in L^p$. □

We conclude this section by noting that, in view of Theorem 3.2, the Lipschitz regularity result established in [1, Theorem 3.2] can be extended to nonconvex integrands by using the same proof. (For the existence of the minimum, see Marcellini [18], Olech [28].)

THEOREM 3.6. *Let $f : [a,b] \times \mathbb{R}^n \to [0, +\infty[$ be an $\mathcal{L} \bigotimes \mathcal{B}_n$-measurable integrand. Assume that a function $\theta : \mathbb{R} \to \mathbb{R}$ exists with $\lim_{x\to+\infty} \theta(x)/x = +\infty$ such that*

$$f(t,z) \geq \theta(|z|) \quad \text{for every } z \in \mathbb{R}^n \text{ and almost every } t \in [a,b].$$

Moreover, assume that there exists $z_0 \in L^\infty(a,b)$ such that $f(t,z_0(t)) \in L^\infty(a,b)$.

Then, every (global) minimizer of the functional F is Lipschitz continuous on $[a,b]$.

4. Applications: Existence of the minimum for $f^{**}(t,\cdot)$ affine at ∞. Let $f : [0,1] \times \mathbb{R} \to \mathbb{R}$ be a normal integrand, that is, such that $f(t,\cdot)$ is lower semicontinuous for almost every t and there exists a Borel function $\tilde{f} : [0,1] \times \mathbb{R} \to \mathbb{R}$ such that $\tilde{f}(t,\cdot) = f(t,\cdot)$ for a.a. $t \in [0,1]$. Recall that a Carathéodory function is a normal integrand (see [11, Proposition 8.1.1]).

In what follows, for every $t \in [0,1]$ we denote by C_t the set where $f^{**}(t,\cdot)$ coincides with $f(t,\cdot)$, i.e.,

$$C_t = \{z \in \mathbb{R} : f^{**}(t,z) = f(t,z)\}, \quad t \in [0,1].$$

If the set C_t does not depend on t, we denote it by C.

As an immediate consequence of Theorem 3.2, it follows that if F admits a minimizer u_0, then the convex envelope $f^{**}(t,\cdot)$ takes real values and the set C_t is nonempty for almost every $t \in [0,1]$. For this reason, from now on we assume $C_t \neq \emptyset$ for almost every $t \in [0,1]$.

In this section we deal with the case of integrands whose convex envelope is affine at ∞. More precisely, we now assume that C_t is bounded for almost every $t \in [0,1]$. In this case, the convex envelope $f^{**}(t,\cdot)$ is affine outside the set C_t.

As is well known, if the integrand does not depend on t, i.e., $f(t,z) \equiv h(z)$, the minimum exists if and only if

$$\min C \leq d \leq \max C$$

(see [10, Theorem 5.2.6]). When the integrand also depends on t, from Theorem 3.1 it follows that a necessary condition for the existence of the minimum is

$$\int_0^1 [\min C_t]\ dt \leq d \leq \int_0^1 [\max C_t]\ dt, \tag{4.1}$$

provided that the integrals above are well-defined. But unfortunately, this condition is not sufficient in general, as the following example shows.

Example 4.1. Let $f(t,z) = \phi(t)h(z)$, where $h : \mathbb{R} \to \mathbb{R}$ is the function defined by

$$h(z) = \begin{cases} z^2 - 1 & \text{for } |z| \le 1, \\ \log|z|/|z| + |2z| - 2 & \text{for } |z| > 1, \end{cases}$$

and $\phi(t) = t + 1$, $t \in [0,1]$.

We have that

$$h^{**}(z) = \begin{cases} z^2 - 1 & \text{for } |z| \le 1, \\ |2z| - 2 & \text{for } |z| > 1 \end{cases}$$

is a C^1-function, and the set of coincidence points of h and h^{**} is given by $C = [-1,1]$.

Note that if F admits a minimum u_0, then $u_0'(t) \in [-1,1]$ and a constant c exists such that $c/(t+1) = h'(u_0'(t)) = 2u_0'(t)$ for almost every $t \in [0,1]$. Therefore, $c/(t+1) \in [-2,2]$ a.e. in $[0,1]$; hence $|c| \le 2$. Moreover,

$$|d| = \left| \int_0^1 u_0'(t)dt \right| = \frac{|c|}{2} \int_0^1 \frac{1}{(t+1)} dt = \frac{|c| \log 2}{2} \le \log 2 < 1.$$

Then, if $|d| > \log 2$, the minimum does not exist.

Our aim in this section is to establish the limitation on the slope d, stronger than (4.1), which is necessary and sufficient for the existence of the minimum. In other words, denote by

$$D = \{d \in \mathbb{R} \ : \ F(v) \text{ admits minimum}\}$$

the range of the value of d for which the minimum exists; we now exactly determine the set D.

In order to do this, let us introduce some notation.

Let $(f_z^{**})^+(t,\cdot), (f_z^{**})^-(t,\cdot)$ be the right and left derivatives of $f^{**}(t,\cdot)$. Set

$$\alpha(t) = (f_z^{**})^-(t, \min C_t), \quad \beta(t) = (f_z^{**})^+(t, \max C_t),$$

let $g^+, g^- : [0,1] \times \mathbb{R} \to \mathbb{R}$ be the functions defined by

$$g^+(t,\zeta) = \begin{cases} \max \partial f^*(t,\zeta) & \text{for } \zeta < \beta(t), \\ \max C_t & \text{for } \zeta \ge \beta(t), \end{cases}$$

$$g^-(t,\zeta) = \begin{cases} \min \partial f^*(t,\zeta) & \text{for } \zeta > \alpha(t), \\ \min C_t & \text{for } \zeta \le \alpha(t), \end{cases}$$

where $\partial f^*(t,\cdot)$ denotes the subgradient of $f^*(t,\cdot)$.

Note that when $\alpha(t) < \zeta < \beta(t)$, the functions g^+, g^-, respectively, are the right and left derivative of the polar function $f^*(t,\cdot)$. We cut off these functions in such a way that they assume values in $[\min C_t, \max C_t]$.

In spite of this modification, the functions g^+, g^- satisfy the same properties of the derivatives of $f^*(t,\cdot)$. More precisely (see [29]), $g^+(t,\cdot), g^-(t,\cdot)$ are monotone nondecreasing, and $g^+(t,\cdot)$ is right-continuous, whereas $g^-(t,\cdot)$ is left-continuous. Moreover, if $\zeta_1 < \zeta_2$, then $g^-(t,\zeta_1) \le g^+(t,\zeta_1) \le g^-(t,\zeta_2) \le g^+(t,\zeta_2)$.

Finally, note that $\min C_t \in \partial f^*(t, \alpha(t))$ and $\max C_t \in \partial f^*(t, \beta(t))$.

Set

$$l = \operatorname*{esssup}_{t\in[0,1]} (f_z^{**})^-(t, \min C_t), \quad L = \operatorname*{ess\,inf}_{t\in[0,1]} (f_z^{**})^+(t, \max C_t).$$

The necessary and sufficient condition for the existence of the minimum is expressed by the following result.

THEOREM 4.2. *Assume that C_t is bounded for almost every $t \in \mathbb{R}$ and $\max C_t$, $\min C_t \in L^p(0,1)$. Then, the functional F admits a minimum if and only if $l \le L$ and*

$$\int_0^1 g^-(t,l)dt \le d \le \int_0^1 g^+(t,L)dt. \tag{4.2}$$

This condition becomes more expressive when the integrand f has the particular structure $f(t,z) = \phi(t)h(z)$, with $\phi(t)$ measurable and nonnegative and h lower semicontinuous. Note that in this case the set C does not depend on t.

Set

$$m = \operatorname*{ess\,inf}_{t\in[0,1]} \phi(t), \quad M = \operatorname*{ess\,sup}_{t\in[0,1]} \phi(t), \quad h_1 = \lim_{z\to-\infty} (h^{**})'(z), \quad h_2 = \lim_{z\to+\infty} (h^{**})'(z).$$

Moreover, let $\tilde g^+, \tilde g^- : [h_1, h_2] \to \mathbb{R}$ be the functions defined by

$$\tilde g^+(\zeta) = \begin{cases} \max \partial h^*(\zeta) & \text{for } \zeta < h_2, \\ \max C & \text{for } \zeta = h_2, \end{cases}$$

$$\tilde g^-(\zeta) = \begin{cases} \min \partial h^*(\zeta) & \text{for } \zeta > h_1, \\ \min C & \text{for } \zeta = h_1. \end{cases}$$

In the following result we will assume $m > 0$. The case $m = 0$ will be tackled in Corollary 4.11.

COROLLARY 4.3. *Assume that C is bounded and $m = \operatorname{essinf}_{t\in[0,1]}\phi(t) > 0$. Then, the functional F admits a minimum if and only if the following condition* (4.3) *(expressed according to the sign of h_1, h_2) is satisfied:*

$$\int_0^1 \tilde g^- \left(\frac{mh_1}{\phi(t)}\right) dt \le d \le \int_0^1 \tilde g^+ \left(\frac{mh_2}{\phi(t)}\right) dt \qquad \text{if} \quad h_1 \le 0 \le h_2, \tag{4.3$_1$}$$

$$M < +\infty \quad \text{and} \quad \int_0^1 \tilde g^- \left(\frac{Mh_1}{\phi(t)}\right) dt \le d \le \int_0^1 \tilde g^+ \left(\frac{mh_2}{\phi(t)}\right) dt \ \text{if} \quad 0 < h_1 \le h_2, \tag{4.3$_2$}$$

$$M < +\infty \quad \text{and} \quad \int_0^1 \tilde g^- \left(\frac{mh_1}{\phi(t)}\right) dt \le d \le \int_0^1 \tilde g^+ \left(\frac{Mh_2}{\phi(t)}\right) dt \ \text{if} \quad h_1 \le h_2 < 0. \tag{4.3$_3$}$$

Moreover the minimum, if it exists, is Lipschitz continuous.

Remark 4.4. Taking into account the definition of the functions $\tilde g^+, \tilde g^-$, the integrands which appear in inequalities (4.3$_i$), $i = 1, 2, 3$, assume values in $[\min C, \max C]$. Note that if $h_1 = 0$, then the integral on the left-hand side of (4.3$_1$) coincides with $\min C$, whereas, if $h_2 = 0$, then the integral on the right-hand side coincides with $\max C$.

Moreover, when ϕ is not constant, if h^{**} is strictly convex in a left neighborhood of $\max C$, with $h_2 \neq 0$, then the integrals in the right-hand side of (4.3$_i$) are strictly less than $\max C$. Similarly, if h^{**} is strictly convex in a right neighborhood of $\min C$,

with $h_1 \neq 0$, then the integrals in the left-hand side of (4.3_i) are strictly greater than $\min C$. Hence, when these cases occur, conditions (4.3_i), $i = 1, 2, 3$, actually are more restrictive than (4.1), and then D is strictly contained in $co(C)$.

Whereas, when h^{**} is affine in $co(C)$ it can happen that conditions (4.3_i) coincide with (4.1), that is, the range D coincides with $co(C)$. The following result shows under what conditions this situation occurs.

COROLLARY 4.5. *Let C be bounded and $m = \operatorname{essinf}_{t\in[0,1]}\phi(t) > 0$. Moreover, let h^{**} be affine in $co(C)$ with slope h_0.*

Assume that the following condition (4.4) *holds (expressed according to the sign of h_1, h_2):*

$$mh_1 \leq Mh_0 \leq mh_2 \quad \text{if} \quad h_1 \leq 0 \leq h_2, \tag{4.4$_1$}$$

$$Mh_1 \leq mh_0 \leq Mh_0 \leq mh_2 \quad \text{if} \quad 0 < h_1 \leq h_2, \tag{4.4$_2$}$$

$$mh_1 \leq Mh_0 \leq mh_0 \leq Mh_2 \quad \text{if} \quad h_1 \leq h_2 < 0. \tag{4.4$_3$}$$

Then, the minimum exists if and only if $\min C \leq d \leq \max C$, i.e., $D = co(C)$.

Vice versa, if (4.4) *is not satisfied and C is not a singleton, then condition* (4.3) *is actually more restrictive than* (4.1)*, i.e., D is strictly contained in $co(C)$.*

Remark 4.6. In the case $h_1h_2 > 0$, according to (4.3_2), (4.3_3), $M < +\infty$ is a necessary condition for the existence of the minimum. Therefore, in (4.4_2), (4.4_3), M is assumed to be real.

When $h_1 \leq 0 \leq h_2$, M could be $+\infty$. Hence, in this case if $h_0 \neq 0$, then condition (4.4_1) is not satisfied, but if $h_0 = 0$, then it is, with the convention $+\infty \cdot 0 = 0$. In contrast, if h^{**} is constant in co(C), that is, $h_1 \leq h_0 = 0 \leq h_2$, condition (4.4_1) is satisfied whatever the function ϕ may be, and then $D = co(C)$.

We now provide some examples of applications of Corollaries 4.3, 4.5.

Example 4.7. Let $f(t, z)$ be the integrand defined in Example 4.1. We have $C = [-1, 1]$, $h_1 = -2$, $h_2 = 2$, $m = 1$. Moreover, since h^{**} is C^1, we have $g^-(\zeta) = g^+(\zeta) = \zeta/2$ for $|\zeta| < 2$. Hence, condition (4.3_1) becomes

$$|d| \leq \int_0^1 \frac{1}{(t+1)} dt = \log 2,$$

i.e., the minimum exists if and only if $|d| \leq \log 2$. Here, since h^{**} is strictly convex in C, the range D is a proper subset of C.

Example 4.8. If we modify the definition of function h by putting $h(z) = 1 - z^2$ in $[-1, 1]$, then $h^{**}(z) = 0$ for every $z \in [-1, 1]$. In this case, by virtue of Corollary 4.5 (see also Remark 4.6), the minimum exists if and only if $|d| \leq 1$.

Example 4.9. Let $h : \mathbb{R} \to \mathbb{R}$ be the function defined by

$$h(z) = \begin{cases} z^2 - 1 & \text{for } |z| \leq 1, \\ \log z/z + 2z - 2 & \text{for } z > 1, \\ \log |z|/|z| & \text{for } z < -1, \end{cases}$$

and let $\phi(t) = t + 1$, $t \in [0, 1]$. We have that

$$h^{**}(z) = \begin{cases} z^2 - 1 & \text{for } 0 \leq z \leq 1, \\ 2z - 2 & \text{for } z > 1, \\ -1 & \text{for } z < 0 \end{cases}$$

is a C^1-function and $C = [0, 1]$. In this case $h_1 = 0, h_2 = 2, m = 1$; hence by applying condition (4.3_1), we obtain

$$0 \leq d \leq \int_0^1 \frac{1}{(t+1)} dt = \log 2.$$

Example 4.10. Let $\tilde{h}(z) = h(z) + z$, where h is as in Example 4.8. In this case, we still have $co(C) = [-1, 1]$ and $\tilde{h}^{**}$ affine in $co(C)$ with slope $h_0 = 1$. Hence, by virtue of Corollary 4.5, if $\phi(t)$ is such that $M \leq 3m$, then the minimum exists if and only if $|d| \leq 1$, that is, the range D coincides with $co(C)$. In fact, we have $h_1 = -1$, $h_2 = 3$, and then condition (4.4_1) is satisfied. Whereas, if $M > 3m$, then D is strictly contained in $[-1, 1]$.

Let us now consider the case in which $C = \{z_0\}$ is a singleton. Of course, if the minimum u_0 exists, then necessarily we have $d = z_0$ and $u_0(t) = td$.

Note that in the case $h_1 \leq 0 \leq h_2$ the linear function is actually the minimizer since $0 \in \partial h^{**}(d)$. Whereas, in the case $0 < h_1 \leq h_2$ $(h_1 \leq h_2 < 0)$ it is the minimizer if and only if $Mh_1 \leq mh_2$ $(mh_1 \leq Mh_2)$. In fact, since $\partial h^{**}(d) = [h_1, h_2]$, we have $h_1\phi(t) \leq c \leq h_2\phi(t)$ for some constant c.

In the following result we will consider the case in which $m = \operatorname{essinf}_{t\in[0,1]} \phi(t) = 0$. In this case, the necessary and sufficient condition on the slope d is much more restrictive than the one obtained in the case $m > 0$.

COROLLARY 4.11. *Assume that C is bounded and*

$$0 = m < \phi(t) \quad \text{a.e. in } [0, 1]. \tag{4.5}$$

Then, the functional F admits a minimum if and only if

$$\min C \leq d \leq \max C \quad \text{and} \quad h^{**}(d) = \min_{z\in\mathbb{R}} h^{**}(z),$$

that is, $D = co(C) \cap \partial h^*(0)$.

As an application of this result, note that if we replace function $\phi(t) = t + 1$ with function $\tilde{\phi}(t) = t$ in Examples 4.7, 4.8, 4.9, the necessary and sufficient condition for the existence of the minimum becomes $d = 0$ for Examples 4.7 and 4.9, $|d| \leq 1$ for Example 4.8.

Remark 4.12. In the previous result we do not take into consideration the case in which the set $Z = \{t : \phi(t) = 0\}$ has positive measure (but is less than 1), since in this case the existence of the minimum does not depend on the value of the slope d.

In fact, it is easy to check that the functional F admits a minimum if and only if the function h admits a minimum, and a minimizer u_0 for F is the function such that $u_0'(t) = z_0$ for $t \notin Z$, $u_0'(t) = \frac{d - z_0(1-\mu)}{\mu}$ for $t \in Z$, where $\mu = \text{meas(Z)}$ and z_0 is any point such that $h(z_0) = \min_{z\in\mathbb{R}} h(z)$.

We conclude this section by providing proofs of the results stated above.

Proof of Theorem 4.2. (*Necessary condition*) By virtue of Theorem 3.2, if F admits a minimum, then a function u_0 and a constant c exist such that $\int_0^1 u_0'(t)dt = d$ and

$$c \in \partial f^{**}(t, u_0'(t)) \text{ a.e. in } [0, 1]. \tag{4.6}$$

Moreover, we have $u_0'(t) \in C_t$ a.e. in $[0, 1]$. Therefore,

$$\alpha(t) \leq (f_z^{**})^-(t, u_0'(t)) \leq c \leq (f_z^{**})^+(t, u_0'(t)) \leq \beta(t) \text{ a.e. in } [0, 1],$$

so $l \le c \le L$.

Finally, from (4.6) we have

$$g^-(t,l) \le g^-(t,c) \le u_0'(t) \le g^+(t,c) \le g^+(t,L),$$

from which (4.2) follows.

(*Sufficient condition*) Since g^+ and g^-, respectively, are right-continuous and left-continuous, we deduce that the functions $G^+(s) := \int_0^1 g^+(t,s)dt$ and $G^-(s) := \int_0^1 g^-(t,s)dt$, respectively, are right-continuous and left-continuous in $[l,L]$. Therefore, for

$$c = \sup\{s : G^-(s) \le d\}$$

we have

$$\int_0^1 g^-(t,c)dt \le d \le \int_0^1 g^+(t,c)dt.$$

Then, it is easy to prove that a constant $r \in [0,1]$ exists such that the function

$$\psi(t) = \begin{cases} g^-(t,c) & \text{for } t \in [0,r], \\ g^+(t,c) & \text{for } t \in [r,1] \end{cases}$$

satisfies $\int_0^1 \psi(t)dt = d$.

Set $u_0(t) := \int_0^t \psi(\tau)d\tau$. Then, from the definition of ψ it follows that $u_0'(t) \in \partial f^*(t,c)$, that is, $c \in \partial f^{**}(t,u_0'(t))$ for a.a. $t \in [0,1]$. Hence, by Theorem 3.5, we deduce that u_0 is a minimizer for functional F, and this concludes the proof. ∎

Proof of Corollary 4.3. When $h_1 \le 0 \le h_2$ we have $l = mh_1$ and $L = mh_2$. Moreover, it is easy to verify that $\partial f^*(t,\zeta) = \partial h^*(\frac{\zeta}{\phi(t)})$; then $g^+(t,\zeta) = \tilde{g}^+(\frac{\zeta}{\phi(t)})$ and $g^-(t,\zeta) = \tilde{g}^-(\frac{\zeta}{\phi(t)})$ for $\zeta \in [l,L]$ and a.a. $t \in [0,1]$. Therefore, condition (4.2) is equivalent to (4.3_1).

Assume now $h_1 > 0$. In this case we have $l = Mh_1$ and $L = mh_2$. Therefore, if $l \le L$, then $M < +\infty$. Moreover, by virtue of the observations above, we have $g^+(t,L) = \tilde{g}^+(\frac{mh_2}{\phi(t)})$ and $g^-(t,l) = \tilde{g}^-(\frac{Mh_1}{\phi(t)})$. As a consequence, (4.2) is equivalent to (4.3_2).

The case $h_2 < 0$ can be treated in a similar way.

Finally, since C is bounded, if a minimizer exists, it is Lipschitz continuous. ∎

Proof of Corollary 4.5. Assume $h_1 \le 0 \le h_2$; the other cases can be proved in a similar way.

If (4.4_1) holds, then $\frac{mh_2}{\phi(t)} \ge h_0$ for a.a. $t \in [0,1]$; hence $\tilde{g}^+(\frac{mh_2}{\phi(t)}) \ge \tilde{g}^+(h_0) \ge \max C$, that is, $\int_0^1 \tilde{g}^+(\frac{mh_2}{\phi(t)})dt = \max C$.

Similarly, we have $\frac{mh_1}{\phi(t)} \le h_0$ for a.a. $t \in [0,1]$. Hence, $\tilde{g}^-(\frac{mh_1}{\phi(t)}) \le \tilde{g}^-(h_0) \le \min C$, and the assertion is proved.

Finally, note that if (4.4_1) does not hold, for example if $mh_2 < Mh_0$, then we have $\tilde{g}^+(\frac{mh_2}{\phi(t)}) \le \tilde{g}^-(h_0) \le \min C < \max C$ in a set of positive measure. Hence we have $\int_0^1 \tilde{g}^+(\frac{mh_2}{\phi(t)})dt < \max C$, and condition (4.3_1) is more restrictive than (4.1). ∎

Proof of Corollary 4.11. (*Necessary condition*) First note that if F admits a minimum, then $h_1 \le 0 \le h_2$. In fact, if $h_1 > 0$, then $l = Mh_1 > 0 = L$, which is a contradiction by Theorem 4.2. Similarly we exclude that $h_2 < 0$.

Hence, since $l = mh_1 = 0 = mh_2 = L$, condition (4.2) becomes $\tilde{g}^-(0) \le d \le \tilde{g}^+(0)$; that is, $0 \in \partial h^{**}(d)$ and then $h^{**}(d) = \min_{z\in\mathbb{R}} h^{**}(z)$.

Finally, from (4.1) we deduce that $\min C \le d \le \max C$.

(*Sufficient condition*) From the assumptions it can be immediately verified that the linear function $u_0(t) = td$ is a minimizer for the relaxed functional F^{**}, and since $d \in co(C)$, the assertion follows from Theorem 3.5. □

5. Applications: Existence of the minimum for $f^{**}(t,\cdot)$ strictly convex at ∞. In this section we deal with integrands $f : [0,1] \times \mathbb{R} \to \mathbb{R}$ whose convex envelope is strictly convex at ∞. More precisely, we now assume that the set $\mathbb{R} \setminus C_t$ is bounded for a.a. $t \in \mathbb{R}$ and that $f(t,\cdot)$ is strictly convex outside the set $co(\mathbb{R} \setminus C_t)$.

We now give some notations.

Let $f_z^+(t,\cdot), f_z^-(t,\cdot)$ be the right and left derivatives of $f(t,\cdot)$, and let $f_\zeta^{*+}(t,\cdot)$, $f_\zeta^{*-}(t,\cdot)$ be the right and left derivatives of the polar function $f^*(t,\cdot)$.

Set $f_z^-(t,-\infty) = f_z^+(t,-\infty) = \lim_{\xi\to-\infty} f_z^-(t,\xi) = \lim_{\xi\to-\infty} f_z^+(t,\xi)$ and $f_z^-(t,+\infty) = f_z^+(t,+\infty) = \lim_{\xi\to+\infty} f_z^-(t,\xi) = \lim_{\xi\to+\infty} f_z^+(t,\xi)$.

Finally, let

$$l = \operatorname*{ess\,sup}_{t\in[0,1]} f_z^-(t,-\infty) \quad \text{and} \quad L = \operatorname*{ess\,inf}_{t\in[0,1]} f_z^+(t,+\infty),$$

$$T_p^+ = \{s \in [l,L] \cap \mathbb{R} : f_\zeta^{*+}(t,s) \in L^p(0,1)\},$$

$$T_p^- = \{s \in [l,L] \cap \mathbb{R} : f_\zeta^{*-}(t,s) \in L^p(0,1)\}.$$

The following result extends Theorem 3′ in [15] to the case of nonconvex integrands.

THEOREM 5.1. *Assume that $\mathbb{R} \setminus C_t$ is bounded for a.a. $t \in [0,1]$ with*

$$\min(\mathbb{R} \setminus C_t), \max(\mathbb{R} \setminus C_t) \in L^p(0,1) \tag{5.1}$$

and $f(t,\cdot)$ strictly convex outside the set $co(\mathbb{R} \setminus C_t)$.

Then, the functional F admits a minimum if and only if one of the following conditions is satisfied:

(a) $T_p^+ \cap T_p^- \neq \emptyset$ *and* $\displaystyle\inf_{s\in T_p^-} \int_0^1 f_\zeta^{*-}(t,s)dt < d < \sup_{s\in T_p^+} \int_0^1 f_\zeta^{*+}(t,s)dt,$

(b) $T_p^- \neq \emptyset$ *and* $\displaystyle d = \min_{s\in T_p^-} \int_0^1 f_\zeta^{*-}(t,s)dt,$

(c) $T_p^+ \neq \emptyset$ *and* $\displaystyle d = \max_{s\in T_p^+} \int_0^1 f_\zeta^{*+}(t,s)dt.$

We wish to remark that conditions (a), (b), (c) are exactly the same as those we obtained in the convex case in Theorem 3′ in [15].

Proof. First we observe that in [15] the function $f(t,\cdot)$ is assumed to be C^1, as well as convex. But, as we showed in [16], it is easy to prove that the assumption C^1 can be weakened by taking $f(t,\cdot)$ only continuous.

(*Sufficient condition*) By Theorem 3′ in [15] we deduce that the relaxed functional $F^{**}(v)$ admits a minimum v_0. Then, if we show that there exists a function $u_0 \in v_0 + W_0^{1,p}(0,1)$ such that $F(u_0) = F^{**}(v_0)$, the assertion will be proved.

Since the set $\mathbb{R} \setminus C_t$ is bounded, we have (see [11, Lemma 8.3.3])

$$f^{**}(t,z) = \min\{\lambda_1 f(t,\xi_1) + \lambda_2 f(t,\xi_2) \ : \ \lambda_1,\lambda_2 \in [0,1],\ \lambda_1+\lambda_2 = 1,\ \lambda_1\xi_1+\lambda_2\xi_2 = z\}.$$

Therefore, by virtue of Proposition 8.3.1 in [11], there exist measurable functions $\lambda_1, \lambda_2 : [0,1] \to [0,1]$, $\xi_1, \xi_2 : [0,1] \to \mathbb{R}$ such that

$$f^{**}(t, v_0'(t)) = \lambda_1(t) f(t, \xi_1(t)) + \lambda_2(t) f(t, \xi_2(t)), \quad \lambda_1(t)\xi_1(t) + \lambda_2(t)\xi_2(t) = v_0'(t).$$

Moreover, by (5.1) we deduce that $\xi_1, \xi_2 \in L^p(0,1)$.

Now, by applying the Lyapunov theorem to the functions $g_j(t) = (\xi_j(t), f(t, \xi_j(t)))$, $j = 1, 2$, we deduce that there exist two measurable disjoint sets E_1, E_2 such that for $\psi(t) = \chi_{E_1}(t)\xi_1(t) + \chi_{E_2}(t)\xi_2(t)$ we have $\int_0^1 \psi(t)dt = \int_0^1 v_0'(t)dt$ and

$$\int_0^1 f(t, \psi(t))dt = \int_{E_1} f(t, \xi_1(t))dt + \int_{E_2} f(t, \xi_2(t))dt$$

$$= \int_0^1 [\lambda_1(t) f(t, \xi_1(t)) + \lambda_2(t) f(t, \xi_2(t))]dt = \int_0^1 f^{**}(t, v_0'(t))dt = F^{**}(v_0).$$

Hence, for $u_0(t) = \int_0^t \psi(\tau)d\tau$ we have $u_0 \in v_0 + W_0^{1,p}(0,1)$ and $F(u_0) = F^{**}(v_0)$.

(*Necessary condition*) The necessary condition is an immediate consequence of Theorem 3.1 and of Theorem 3′ in [15], taking into account the relation $f^* = (f^{**})^*$. □

From this result it is possible to derive very operative necessary and sufficient conditions in the case in which the integrand has the particular structure $f(t,z) = \phi(t)h(z)$, as in [15]. In view of the present results, we can see that Theorems 5–9 in [15] hold even for nonconvex integrands, provided that the set $\mathbb{R} \setminus C$ is bounded and that h is strictly convex outside the set C.

These conditions emphasize a strict link between the exponent p, the behavior at ∞ of function h, and the infinitesimal order of $[\phi(t) - m]$.

We now quote only two significant results in a slightly less general setting, referring to [15] for proofs and details. For the sake of simplicity, in the following results the function ϕ is assumed to be continuous, and h to be C^1.

COROLLARY 5.2 (quasi-coercive case). *Assume that* $\min \phi = \phi(t_0) = 0$ *and* $\phi(t) > 0$ *for* $t \neq t_0$. *Moreover, let* h *be coercive.*

Furthermore, assume that $\phi \in O(\alpha)$ *when* $t \to t_0$ *and* $1/h' \in O(\beta)$ *when* $|z| \to +\infty$, *where* $O(\cdot)$ *denotes the infinitesimal order.*

Then, if $p\alpha < \beta$, *the functional* F *admits a minimum for every* $d \in \mathbb{R}$. *Whereas, if* $p\alpha \geq \beta$, *the minimum exists if and only if* $h^{**}(d) = \min_{z \in \mathbb{R}} h^{**}(z)$.

COROLLARY 5.3 (noncoercive case). *Set* $\lim_{z \to -\infty} h'(z) = h_1, \lim_{z \to +\infty} h'(z) = h_2$, *and assume that* $-\infty < h_1 < 0 < h_2 < +\infty$. *Moreover, assume that* $m = \min \phi(t) > 0$.

Finally, suppose that $[\phi(t) - m] \in O(\alpha)$ *and* $[h'(z) - h_2], [h'(z) - h_1] \in O(\beta)$.

Then, if $\alpha \geq \beta$, *the functional* F *admits a minimum for every* $d \in \mathbb{R}$. *Whereas, if* $\alpha < \beta$, *then the range* D *of the values of the slope* d *for which the minimum exists is a bounded interval.*

Remark 5.4. For the sake of brevity, we do not explicitly treat mixed cases, for example when h^{**} is strictly convex at $-\infty$ but is linear at $+\infty$, that is, when $\inf C = -\infty$ but $\sup C < +\infty$. In fact, we think it is now clear how to combine the conditions of Corollaries 4.3–4.5 and Theorem 5.1 in order to deal with these cases.

REFERENCES

[1] L. AMBROSIO, O. ASCENZI, AND G. BUTTAZZO, *Lipschitz regularity for minimizers of integral functionals with highly discontinuous integrands*, J. Math. Anal. Appl., 142 (1989), pp. 301–316.

[2] J. Ball, *Minimizers and the Euler-Lagrange equations*, in Trends and Applications of Pure Mathematics to Mechanics, Lecture Notes in Phys. 195, Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1984, pp. 1–4.
[3] J. Ball and Mizel, *One-dimensional variational problems whose minimizers do not satisfy the Euler-Lagrange equation*, Arch. Rat. Mech. Anal., 90 (1985), pp. 325–388.
[4] P. Brandi, *Sul problema libero unidimensionale del calcolo delle variazioni*, Atti Sem. Mat. Fis. Univ. Modena, 28 (1979), pp. 15–32.
[5] C. Castaing and M. Valadier, *Convex Analysis and Measurable Multifunctions*, Lecture Notes in Math. 580, Springer-Verlag, Berlin, 1977.
[6] L. Cesari, *Optimization Theory and Applications*, Springer-Verlag, New York, Heidelberg, Berlin, 1983.
[7] F. H. Clarke, *The Euler-Lagrange differential inclusion*, J. Differential Equations, 19 (1975), pp. 80–90.
[8] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Canad. Math. Soc. Ser. Monogr. Adv. Texts, Wiley-Interscience, New York, 1983.
[9] F. H. Clarke and R. B. Vinter, *On the condition under which the Euler-Lagrange equation or the maximum principle hold*, Appl. Math. Optim., 12 (1984), pp. 73–79.
[10] B. Dacorogna, *Direct Methods in the Calculus of Variations*, Springer-Verlag, New York, Heidelberg, Berlin, 1989.
[11] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, Stud. Math. Appl., North–Holland, Amsterdam, Oxford, 1976.
[12] A. Ioffe, *Proximal analysis and approximate subdifferential*, J. London Math. Soc., 41 (1990), pp. 175–192.
[13] A. Ioffe and R. T. Rockafellar, *The Euler and Weierstrass conditions for nonsmooth variational problems*, Calc. Var. Partial Differential Equations, 4 (1996), pp. 59–87.
[14] P. J. Kaiser, *A problem of slow growth in the Calculus of Variations*, Atti Sem. Mat. Fis. Univ. Modena, 24 (1975), pp. 236–246.
[15] C. Marcelli, *One-dimensional non coercive problems of the Calculus of Variations*, Ann. Mat. Pura Appl. (IV), 173 (1997), pp. 145–161.
[16] C. Marcelli, *Non-coercive variational problems with constraints on the derivatives*, J. Convex Anal., 5 (1998), pp. 1–17.
[17] C. Marcelli, E. Outkine, and M. Sytchev, *Remarks on necessary conditions for minimizers of one-dimensional variational problems*, Nonlinear Anal., to appear.
[18] P. Marcellini, *Alcune osservazioni sull'esistenza del minimo di integrali del calcolo delle variazioni senza ipotesi di convessitá*, Rend. Mat., 13 (1980), pp. 271–281.
[19] P. Marcellini, *Non-convex integrals of the calculus of variations*, in Methods of Non-convex Analysis, Cellina et al., eds., Springer-Verlag, New York, Heidelberg, Berlin, 1990.
[20] P. Marcellini and C. Sbordone, *Relaxation of non convex variational problems*, Atti Acad. Naz. Lincei, 63 (1977), pp. 341–344.
[21] B. Mordukhovich, *Maximum principle in problems of time optimal control with nonsmooth constraints*, J. Appl. Math. Mech., 40 (1976), pp. 960–969.
[22] B. Mordukhovich, *Approximation Methods in Problems of Optimization and Control*, Nauka, Moscow, 1988 (in Russian).
[23] B. Mordukhovich, *On variational analysis of differential inclusions*, in Optimization and Nonlinear Analysis, A. Ioffe, M. Marcus, S. Reich, eds., Pitman Res. Notes Math. 244, 1992, pp. 199–213.
[24] B. Mordukhovich, *Generalized differential calculus for nonsmooth and set valued mappings*, J. Math. Anal. Appl., 183 (1994), pp. 250–288.
[25] B. S. Mordukhovich, *Discrete approximations and refined Euler–Lagrange conditions for nonconvex differential inclusions*, SIAM J. Control Optim., 33 (1995), pp. 882–915.
[26] B. Mordukhovich, *Existence theorems in non-convex optimal control*, in Calculus of Variations and Optimal Control, Chapman and Hall/CRC Res. Notes Math. 411, Boca Raton, FL, 1999, pp. 173–197.
[27] B. Mordukhovich, *Optimal control of difference, differential and differential-difference inclusions*, J. Math. Sci., 100 (2000), pp. 2613–2632.
[28] C. Olech, *Existence theory in optimal control*, in Control Theory and Topics in Functional Analysis, Vol. I, International Atomic Energy Agency, Vienna, 1976, pp. 291–328.
[29] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[30] R. B. Vinter and P. D. Woodford, *On the occurrence of intermediate local minimizers that are not strong local minimizers*, Systems Control Lett., 31 (1997), pp. 235–242.
[31] R. Vinter and H. Zheng, *The extended Euler–Lagrange condition for nonconvex variational problems*, SIAM J. Control Optim., 35 (1997), pp. 56–77.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 5, pp. 1491–1504

CONVEXITY IN ZERO-SUM DIFFERENTIAL GAMES*

RAFAL GOEBEL†

Abstract. A new approach to two-player zero-sum differential games with convex-concave cost function is presented. It employs the tools of convex and variational analysis. A necessary and sufficient condition on controls to be an open-loop saddle point of the game is given. Explicit formulas for saddle controls are derived in terms of the subdifferential of the function conjugate to the cost. Existence of saddle controls is concluded under very general assumptions, not requiring the compactness of control sets. A Hamiltonian inclusion, new to the field of differential games, is shown to describe equilibrium trajectories of the game.

Key words. two-player zero-sum differential games, open-loop controls, convex-concave functions, generalized Hamiltonian systems

AMS subject classifications. 90D25, 49J35, 49K35

PII. S0363012999360737

1. Introduction.

Consider the following two-player zero-sum differential game. The trajectory of the game, $x(\cdot)$, is described by a linear differential equation

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + C(t)v(t), \tag{1}$$

and the initial condition is

$$x(\tau) = \xi \tag{2}$$

for some $\tau \in (-\infty, T]$ and $\xi \in \mathbb{R}^n$. Controls $u(\cdot)$ and $v(\cdot)$ are functions on $[\tau, T]$ chosen, respectively, by Player One and Player Two from some control sets $\mathcal{U}(\tau, \xi)$ and $\mathcal{V}(\tau, \xi)$, subject to the constraints

$$u(t) \in P(t) \text{ and } v(t) \in Q(t) \text{ for almost all } t \in [\tau, T], \tag{3}$$

for given convex sets $P(t) \in \mathbb{R}^k$, $Q(t) \in \mathbb{R}^l$. The cost functional $\Phi(\tau, \xi, \cdot, \cdot)$ is given by

$$\Phi(\tau, \xi, u(\cdot), v(\cdot)) = \int_\tau^T f(t, u(t), v(t))dt + d \cdot x(T). \tag{4}$$

Player One tries to minimize the cost $\Phi(\tau, \xi, u(\cdot), v(\cdot))$, while Player Two attempts to maximize it. The saddle value of the functional $\Phi(\tau, \xi, \cdot, \cdot)$ is called the value of the game. Controls $\bar{u}(\cdot)$ and $\bar{v}(\cdot)$, corresponding to a saddle point $(\bar{u}(\cdot), \bar{v}(\cdot))$ of this functional are referred to as saddle controls, or open-loop solutions of the game. The trajectory generated by them is an equilibrium trajectory.

Two-player zero-sum games and their generalization, N-player games, have seen extensive treatment in literature; a good reference is Basar and Olsder [1]. Assumptions of convexity of the cost for each player allowed Varaiya [14], Scalzo [11], and

*Received by the editors August 30, 1999; accepted for publication (in revised form) September 5, 2001; published electronically January 18, 2002. This work was supported in part by the National Science Foundation under grant DMS-9803089.

http://www.siam.org/journals/sicon/40-5/36073.html

†Centre for Experimental and Constructive Mathematics, Simon Fraser University and Department of Mathematics, University of British Columbia. Current address: Department of Mathematics, University of British Columbia, 1984 Mathematics Road, Vancouver, B.C., Canada V6T 1Z2 (rafal@cecm.sfu.ca).

Tolwinski [13] to obtain results on the existence of open-loop solutions for N-player games. Reference [13] contains a discussion of other related results. For two-player zero-sum games, an assumption of convexity of costs yields a convex-concave cost functional, which one player minimizes and the other maximizes. A special case of such a game was treated by Berkovitz [3].

All of the mentioned works rely on the compactness of sets of feasible control functions, achieved either by assumption of boundedness of control constraint sets, or by a priori integral bounds on control functions. This allows for the use of fixed point and saddle point theorems in infinite-dimensional function spaces to guarantee the existence of equilibrium trajectories. Here we approach the existence issue directly by analyzing explicit descriptions of saddle controls. We demonstrate that the strong assumptions of the compactness of control functions are not necessary in the convex-concave setting—sufficient "compactness-like" properties are guaranteed through the so-called Isaacs condition. In the language of convex analysis, finiteness of the saddle function conjugate to the cost reflects the desired growth properties of the cost function itself. Details are discussed in section 2.

In section 3 we show that for the controls $(\bar{u}(\cdot), \bar{v}(\cdot))$ to furnish a saddle point of the game, it is necessary and sufficient for $(\bar{u}(t), \bar{v}(t))$ to be a saddle point on $P(t) \times Q(t)$ of an auxiliary function

$$S(t,u,v) = f(t,u,v) + d \cdot \mathcal{A}(T,t)[B(t)u + C(t)v]. \tag{5}$$

The matrix $\mathcal{A}(T,t)$ is the fundamental matrix for $\dot{w}(t) = A(t)w(t)$, with $\mathcal{A}(T,T)$ being the identity matrix. In a different setting, a characterization of saddle controls as solutions of an instantaneous saddle problem involving a pre-Hamiltonian function was given by Subbotin [12]. A pre-Hamiltonian function has also been used by Berkovitz [2] to give a necessary condition for saddle controls and by Leitmann [5] to state a sufficient condition. The special properties of (5) when the cost is convex-concave seem not to have been analyzed, however.

We explore the convex-concave structure of (5) in section 4 and obtain explicit formulas for saddle points of this auxiliary function. This leads to expressions for saddle controls—see (14) and (15). Formula (15) is of special value, as it describes saddle controls as subgradients of a particular saddle function. Such a description seems more practical in our setting than any min/max expression. It leads, through the analysis of possible growth of subgradients of saddle functions, to results on the existence of saddle controls in Theorem 4.1 and, once the existence is guaranteed, allows for a detailed study of their structure. (See the comments at the end of section 4.) Let us also note that the saddle controls, when they exist, turn out not to depend on the initial condition.

A characterization of an equilibrium trajectory, new to the field of differential games, is given in Theorem 5.1 with the help of the associated Hamiltonian function $H(t,x,y)$. Much in the spirit of optimal control theory, solutions to a nonsmooth Hamiltonian dynamical system, posed in terms of the subgradients of $H(t,x,y)$, turn out to describe the equilibria of the game.

We conclude the introduction with remarks on differential games in the more standard setting of closed-loop controls. There, players choose strategies as functions of both time and state, $U : (-\infty, T] \times \mathbb{R}^n \to \mathbb{R}^k$ and $V : (-\infty, T] \times \mathbb{R}^n \to \mathbb{R}^l$. Given an initial condition (2), these strategies and the dynamics (1) determine the instantaneous controls by $u(t) = U(t, x(t))$ and $v(t) = V(t, x(t))$ and the cost $\Phi(\tau, \xi, U(\cdot,\cdot), V(\cdot,\cdot))$. The closed-loop solutions of the game are the strategies $U(\cdot,\cdot)$

and $V(\cdot,\cdot)$ that provide a saddle point of the cost functional for every (τ,ξ). It is known that if $\bar{u}(\cdot)$ and $\bar{v}(\cdot)$ are open-loop solutions of the game with a specified initial condition, then, by taking

$$U(t,x)=\bar{u}(t), \qquad V(t,x)=\bar{v}(t),$$

one obtains closed-loop saddle strategies for this game; see Berkovitz [3]. Our results yield open-loop controls independent of the initial condition, and so the above equations can be used to define closed-loop strategies on $(-\infty,T]\times\mathbb{R}^n$.

2. Assumptions. We now present the assumptions that are in effect in the remaining sections. The game described by (1)–(4) and the assumptions below will be referred to as $\mathcal{G}(\tau,\xi)$.

ASSUMPTION 2.1 (general assumptions). *The matrices $A(t)\in\mathbb{R}^{n\times n}$, $B(t)\in\mathbb{R}^{n\times k}$, and $C(t)\in\mathbb{R}^{n\times l}$ in (1) depend continuously on $t\in(-\infty,T]$. Control sets $\mathcal{U}(\tau,\xi)$ and $\mathcal{V}(\tau,\xi)$ are subsets of the space of measurable and locally integrable functions on $[\tau,T]$. The constraint sets in (3), $P(t)\subset\mathbb{R}^k$ and $Q(t)\subset\mathbb{R}^l$, are nonempty, closed (but not necessarily bounded), convex, and measurably dependent on t.*

Measurable dependence of sets on time is defined and discussed in chapter 14 of Rockafellar and Wets [9]. Sets depending continuously on time, constant sets in particular, have this property.

A natural generalization of the assumptions on the constraint sets P and Q might be to allow their dependence on the state variable, that is, to let $u(t)\in P(t,x(t))$ and $v(t)\in Q(t,x(t))$. However, this can lead to the set of feasible pairs of controls $(u(\cdot),v(\cdot))$ not being a product set in $\mathcal{U}(\tau,\xi)\times\mathcal{V}(\tau,\xi)$, as opposed to the case where the constraints have the form (3). For example, look at the game where $x(0)=0$, $\dot{x}(t)=u(t)$, and the controls are constrained by $u(t)\in[0,1]$ and $v(t)=x(t)$. In such cases, it is unclear what the notions of an equilibrium and the value of the game should be. We do not address this issue, preferring to work with the constraints given by (3).

The following assumption guarantees that the cost functional $\Phi(\tau,\xi,u(\cdot),v(\cdot))$ is convex in the control $u(\cdot)$ for fixed $(\tau,\xi,v(\cdot))$ and concave in the control $v(\cdot)$ for fixed $(\tau,\xi,u(\cdot))$.

ASSUMPTION 2.2 (convexity-concavity). *The function $f:(-\infty,T]\times P(t)\times Q(t)\to\mathbb{R}$ has the following properties: $f(t,u,v)$ is measurable in t for every fixed (u,v), continuous in $(u,v)\in P(t)\times Q(t)$ for every t, convex in u for every (t,v), and concave in v for every (t,u).*

For purposes of convex analysis, it is convenient to extend the function f to $(-\infty,T]\times\mathbb{R}^k\times\mathbb{R}^l$ by appropriately chosen infinite values. We define

$$\tilde{f}(t,u,v)=\begin{cases} f(t,u,v), & u\in P(t) \text{ and } v\in Q(t),\\ +\infty, & u\notin P(t),\\ -\infty, & u\in P(t) \text{ and } v\notin Q(t).\end{cases} \tag{6}$$

In what follows, we will often consider the function f on $(-\infty,T]\times\mathbb{R}^k\times\mathbb{R}^l$.

As a matter of fact, a game with linear dynamics and a cost function $\hat{f}(t,x,u,v)$ that is convex in (x,u) for fixed (t,v) and concave in (x,v) for fixed (t,u) can be reduced to a game with cost (4). Indeed, these properties imply that

$$\hat{f}(t,x,u,v)=\alpha(t)\cdot x(t)+f(t,u,v)$$

for some function α and a function $f(t,u,v)$ convex in u and concave in v. Skipping the technicalities, we point out that, through an inclusion of an additional one-dimensional variable in the dynamics, we can reformulate the game under discussion in the format of this paper.

ASSUMPTION 2.3 (finiteness of cost). *For any $u(\cdot) \in \mathcal{U}(\tau,\xi)$ and $v(\cdot) \in \mathcal{V}(\tau,\xi)$ satisfying* (3), $\int_\tau^T f(t,u(t),v(t))dt$ *is finite.*

This restrictive-looking assumption is satisfied, in particular, when $f(t,u,v)$ is continuous and the controls are essentially bounded or when $f(t,u,v)$ is a quadratic expression in u and v with bounded coefficients and the controls are L^2 functions.

Under these assumptions, the cost $\Phi(\tau,\xi,u(\cdot),v(\cdot))$ is well defined. A pair of controls $(\bar u(\cdot),\bar v(\cdot))$ is a saddle point (a Nash equilibrium) of $\Phi(\tau,\xi,u(\cdot),v(\cdot))$ over $\mathcal{U}(\tau,\xi)\times\mathcal{V}(\tau,\xi)$ if $(\bar u(\cdot),\bar v(\cdot)) \in \mathcal{U}(\tau,\xi)\times\mathcal{V}(\tau,\xi)$, and for any $(u(\cdot),v(\cdot)) \in \mathcal{U}(\tau,\xi)\times\mathcal{V}(\tau,\xi)$ the following is satisfied:

$$\Phi(\tau,\xi,\bar u(\cdot),v(\cdot)) \le \Phi(\tau,\xi,\bar u(\cdot),\bar v(\cdot)) \le \Phi(\tau,\xi,u(\cdot),\bar v(\cdot)). \tag{7}$$

ASSUMPTION 2.4 (Isaacs condition). *For every $t \in (-\infty,T]$ and every $(p,q) \in \mathbb{R}^k\times\mathbb{R}^l$*

$$\sup_{u\in P(t)}\inf_{v\in Q(t)}\{p\cdot u+q\cdot v-f(t,u,v)\} = \inf_{v\in Q(t)}\sup_{u\in P(t)}\{p\cdot u+q\cdot v-f(t,u,v)\}, \tag{8}$$

and the common value is finite.

The assumption is certainly satisfied whenever the sets $P(t)$, $Q(t)$ are compact (37.6.1 in [6]) but also holds in several other interesting cases.

Example 2.5. Let $P(t)=\mathbb{R}^k$, $Q(t)=\mathbb{R}^l$, and consider $f(t,u,v)=u\cdot v$. A direct calculation shows that both sides of (8) are equal $p\cdot q$. In particular, Assumption 2.4 holds. Note that not only the control constraint sets are unbounded, but the function $f(t,\cdot,\cdot)$ does not seem—at the first glance—to display favorable growth properties.

Example 2.6 (quadratic cost and polyhedral control constraints). Consider a game with polyhedral control constraint sets $P(t)$, $Q(t)$ and the cost function given by

$$f(t,u,v) = \frac{1}{2}u\cdot E(t)u - \frac{1}{2}v\cdot F(t)v + v\cdot G(t)u, \tag{9}$$

where $E(t)$, $F(t)$, and $G(t)$ are matrices of appropriate dimensions, with $E(t)$ and $F(t)$ symmetric and positive semidefinite. Note that a second order approximation of any smooth saddle cost would yield such a function. (We skip the linear terms for simplicity of presentation.) Consider the following condition:

$$\text{for every } t\le T, \qquad (P(t))^\infty \cap \ker E(t) = \{0\}, \qquad (Q(t))^\infty \cap \ker F(t) = \{0\}. \tag{10}$$

Above, C^∞ denotes the recession cone of a convex set C, defined as $\{y\in\mathbb{R}^m \mid C+y\subset C\}$. If (10) holds, then Assumption 2.4 is satisfied for the game under discussion. Indeed, pick any $\bar v(t)\in Q(t)$. We have

$$\begin{aligned}
&\inf_{v\in Q(t)}\sup_{u\in P(t)}\{p\cdot u+q\cdot v-f(t,u,v)\}\\
&\quad\le \sup_{u\in P(t)}\left\{p\cdot u+q\cdot v-\frac{1}{2}u\cdot E(t)u+\frac{1}{2}\bar v\cdot F(t)\bar v-\bar v\cdot G(t)u\right\}\\
&\quad= \frac{1}{2}\bar v\cdot F(t)\bar v+\sup_{u\in P(t)}\left\{(p-G^*(t)\bar v)\cdot u-\frac{1}{2}u\cdot E(t)u\right\}.
\end{aligned}$$

If Θ is a symmetric and positive semidefinite matrix and $\ker(\Theta) \cap C^\infty = \{0\}$, then for any $z \in \mathbb{R}^m$, $\sup_{y \in C}\{z \cdot y - \frac{1}{2} y \cdot \Theta y\}$ is finite; see 11.18 in Rockafellar and Wets [9]. Thus

$$\inf_{v \in Q(t)} \sup_{u \in P(t)} \{p \cdot u + q \cdot v - f(t,u,v)\} < +\infty.$$

A symmetric argument shows that $-\infty < \sup_{u \in P(t)} \inf_{v \in Q(t)} \{p \cdot u + q \cdot v - f(t,u,v)\}$, and since

$$\sup_{u \in P(t)} \inf_{v \in Q(t)} \{p \cdot u + q \cdot v - f(t,u,v)\} \leq \inf_{v \in Q(t)} \sup_{u \in P(t)} \{p \cdot u + q \cdot v - f(t,u,v)\},$$

both expressions are finite. This guarantees that they are actually equal to each other (34.2.1 in Rockafellar [6]).

Note that when matrices $E(t)$, $F(t)$, and $G(t)$ depend continuously on t and the sets $\mathcal{U}(\tau,\xi)$, $\mathcal{V}(\tau,\xi)$ are subsets of $L^2[\tau,T]$, Assumption 2.3 is satisfied (as well as Assumptions 2.1 and 2.2). Let us finish the discussion of this special case with two remarks.

- Conditions (10) are clearly satisfied when sets $P(t)$, $Q(t)$ are bounded (then their recession cones reduce to $\{0\}$) or when matrices $E(t)$, $F(t)$ are positive definite (and thus invertible).
- Conditions (10) take a particularly simple form when sets $P(t)$ and $Q(t)$ are cones: $P(t) \cap \ker E(t) = \{0\}$, $Q(t) \cap \ker F(t) = \{0\}$. Cones include the "nonnegative orthants" $\mathbb{R}^k_+$, $\mathbb{R}^l_+$, and, more generally, finite intersections of half-spaces given by linear (not just affine) subspaces.

A general result, describing all convex-concave functions for which Assumption 2.4 holds, will be given at the end of this section.

We now present the basic notions of convex analysis that will be heavily used in what follows. Extending the function f to $(-\infty,T] \times \mathbb{R}^k \times \mathbb{R}^l$, as described in (6), allows for defining the class of functions conjugate in the convex-concave sense to $f(t,\cdot,\cdot)$. Under the condition (8), this class consists of one function $f^* : (-\infty,T] \times \mathbb{R}^k \times \mathbb{R}^l \to \mathbb{R}$ given by

$$\begin{aligned} f^*(t,p,q) &= \sup_{u \in \mathbb{R}^k} \inf_{v \in \mathbb{R}^l} \{p \cdot u + q \cdot v - \tilde{f}(t,u,v)\} \\ &= \inf_{v \in \mathbb{R}^l} \sup_{u \in \mathbb{R}^k} \{p \cdot u + q \cdot v - \tilde{f}(t,u,v)\}. \end{aligned}$$

Equivalently, $f^*(t,p,q)$ is the common value in (8). The function $f^*(t,p,q)$ is convex in p, concave in q, and locally Lipschitz continuous in (p,q).

The subdifferential $\partial\phi(\bar{p},\bar{q})$ of a convex-concave function $\phi(\cdot,\cdot)$, in the sense of Rockafellar [6], is defined as

$$\partial\phi(\bar{p},\bar{q}) = \partial_1\phi(\bar{p},\bar{q}) \times \partial_2\phi(\bar{p},\bar{q}),$$

where

$$\partial_1\phi(\bar{p},\bar{q}) = \{r \mid \phi(p,\bar{q}) \geq \phi(\bar{p},\bar{q}) + r \cdot (p - \bar{p})\}$$

is the subdifferential in the sense of convex analysis of the convex function $\phi(\cdot,\bar{q})$, and

$$\partial_2\phi(\bar{p},\bar{q}) = \{s \mid \phi(\bar{p},q) \leq \phi(\bar{p},\bar{q}) + s \cdot (q - \bar{q})\}$$

is the subdifferential of the concave function $\phi(\bar{p},\cdot)$. The subdifferential $\partial f^*(t,p,q)$ is then by definition the subdifferential of the convex-concave function $(p,q) \mapsto f^*(t,p,q)$. The function $f^*(t,p,q)$ is differentiable in (p,q), in particular when the cost function $f(t,u,v)$ is strictly convex in p and strictly concave in q. Then the subgradient $\partial f^*(t,p,q)$ equals $\nabla f^*(t,p,q)$, where the gradient is taken with respect to (p,q). If, in addition, the cost function is continuous in (t,u,v) and the sets $P(t)$ and $Q(t)$ evolve continuously in time, as is the case when they are constant, then both $f^*(t,p,q)$ and $\nabla f^*(t,p,q)$ are continuous in (t,p,q). An example where strict convexity and strict concavity are present is provided by quadratic cost function (9) with positive definite matrices $E(t)$ and $F(t)$. For a complete presentation of saddle function theory, see Rockafellar [6].

The section concludes with an equivalent, possibly more practical, condition for Assumption 2.4 to hold. A convex function $\alpha(\cdot)$ is called proper if it does not take on the value $-\infty$ and has a finite value at some point. It is called coercive if it is bounded from below and $\alpha(u)/|u| \to \infty$ as $|u| \to \infty$. An equivalent condition for coercivity of a convex function $\alpha(\cdot)$ is that its conjugate function, given by $\alpha^*(q) = \sup_u\{q\cdot u - \alpha(u)\}$, is finite. We will say that a concave function $\beta(\cdot)$ is proper or coercive if $-\beta(\cdot)$ is proper or coercive in the sense just described. Recall that $\tilde{f}$ is the extension of f to $(-\infty,T]\times\mathbb{R}^k\times\mathbb{R}^l$, defined in (6).

PROPOSITION 2.7 (finiteness of the conjugate function). *Assumption* 2.4 *holds if and only if, for every* $t \in (-\infty,T]$, *the convex function* $\alpha(u) = \sup_v \tilde{f}(t,u,v)$ *and the concave function* $\beta(v) = \inf_u \tilde{f}(t,u,v)$ *are both proper and coercive.*

Proof. For simplicity of notation, fix $t \in (-\infty,T]$. Suppose that Assumption 2.4 holds. Then, for functions $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ from $\mathbb{R}^k\times\mathbb{R}^l$ to $\overline{\mathbb{R}}$ defined by $a(u,q) = \sup_p\{u\cdot p - f^*(t,p,q)\}$, $b(p,v) = \inf_q\{v\cdot q - f^*(t,p,q)\}$, the following conditions hold: for every q, the convex function $a(\cdot,q)$ is proper and coercive, and for every p, the concave function $b(p,\cdot)$ is proper and coercive. But it is also true that $a(u,q) = \sup_v\{\tilde{f}(t,u,v) - v\cdot q\}$, $b(p,v) = \inf_u\{\tilde{f}(t,u,v) - u\cdot p\}$. Justification of these formulas can be found in Rockafellar [8]. In particular, the convex function $a(\cdot,0) = \alpha(\cdot)$ and the concave function $b(0,\cdot) = \beta(\cdot)$ are both proper and coercive.

Now assume that $\alpha(\cdot)$ and $\beta(\cdot)$ are proper and coercive. Then, for the lower conjugate of $\tilde{f}(t,\cdot,\cdot)$, we have

$$\begin{aligned}\underline{\tilde{f}}^*(t,p,0) &= \sup_{u\in\mathbb{R}^k}\inf_{v\in\mathbb{R}^l}\{p\cdot u - \tilde{f}(t,u,v)\} = \sup_{u\in\mathbb{R}^k}\{p\cdot u - \sup_{v\in\mathbb{R}^l}\tilde{f}(t,u,v)\}\\ &= \sup_{u\in\mathbb{R}^k}\{p\cdot u - \alpha(u)\},\end{aligned}$$

and the last quantity is finite, for every p. Similarly, for the upper conjugate function, we have

$$\overline{\tilde{f}}^*(t,0,q) = \inf_{v\in\mathbb{R}^l}\sup_{u\in\mathbb{R}^k}\{q\cdot v - \tilde{f}(t,u,v)\} = \inf_{v\in\mathbb{R}^l}\{q\cdot v - \beta(v)\},$$

and the last quantity is finite, for every q. Then also $\underline{\tilde{f}}^*(t,0,q) < +\infty$ for every q, which, combined with the finiteness of $\underline{\tilde{f}}^*(t,p,0)$ for every p, implies that the saddle function $\underline{\tilde{f}}^*(t,\cdot,\cdot)$ is proper. We can now apply 34.3 in [6] to deduce that $\underline{\tilde{f}}^*(t,\cdot,\cdot)$ is actually finite. Then 34.2.1 in [6] implies that Assumption 2.4 holds. □

3. Necessary and sufficient saddle condition. To proceed with reducing the saddle point problem for an integral functional to the saddle point problem for the

integrand function, we need the notion of decomposable sets of functions, which is a slight modification of the notion of decomposable spaces. This and other notions used in this section, like normal integrands and measurability of set valued mappings, are discussed in Rockafellar and Wets [9, Chapter 14].

Let Z be a set of measurable functions $z : [\tau, T] \to \mathbb{R}^m$, and let $R(t) \subset \mathbb{R}^m$ be a nonempty set depending measurably on t. Define Z_R to be the set of all $z \in Z$ such that $z(t) \in R(t)$ almost everywhere on $[\tau, T]$. The set Z is called decomposable with respect to $R(\cdot)$ if, for every function $z_0 \in Z_R$, every measurable set $W \subset [\tau, T]$, and any bounded, measurable function $z_1 : W \to \mathbb{R}^m$ such that $z_1(t) \in R(t)$ for almost every $t \in W$, Z contains the function given by

$$z(t) = \begin{cases} z_0(t) & \text{for } t \in [\tau, T] \setminus W, \\ z_1(t) & \text{for } t \in W. \end{cases}$$

If $R(t) = \mathbb{R}^m$ for almost all $t \in [\tau, T]$, we call the set Z decomposable. Note that, in such a case, Z is also decomposable with respect to any other constraint set $R'(\cdot)$. Decomposable spaces, as defined in [9], are decomposable sets. An example of decomposable spaces is provided by L^p spaces. On the other extreme, discrete sets of functions consisting of more than one function cannot be decomposable if $R(t)$ is convex and is not a singleton for almost every t.

Recall that the auxiliary saddle function $S(t, u, v)$, given by (5), is finite only for $(u, v) \in P(t) \times Q(t)$. Therefore, the saddle points of $S(t, u, v)$ over $P(t) \times Q(t)$ are the same as over $\mathbb{R}^k \times \mathbb{R}^l$; see 36.3 in Rockafellar [6]. The statement (11) in the following theorem can be understood in either sense.

THEOREM 3.1 (saddle point condition). *Any pair of controls* $(\bar{u}(\cdot), \bar{v}(\cdot)) \in \mathcal{U}(\tau, \xi) \times \mathcal{V}(\tau, \xi)$ *satisfying*

$$(\bar{u}(t), \bar{v}(t)) \text{ is a saddle point of } S(t, u, v) \text{ for almost all } t \in [\tau, T] \tag{11}$$

is a saddle point of $\Phi(\tau, \xi, u(\cdot), v(\cdot))$. *Conversely, if* $\mathcal{U}(\tau, \xi)$ *and* $\mathcal{V}(\tau, \xi)$ *are decomposable with respect to* $P(\cdot)$ *and* $Q(\cdot)$ *and if a saddle point* $(\bar{u}(\cdot), \bar{v}(\cdot))$ *of* $\Phi(\tau, \xi, u(\cdot), v(\cdot))$ *exists, then* (11) *holds.*

COROLLARY 3.2. *If* $\mathcal{U}(\tau, \xi)$ *and* $\mathcal{V}(\tau, \xi)$ *are decomposable with respect to* $P(\cdot)$ *and* $Q(\cdot)$, *as is the case when* $\mathcal{U}(\tau, \xi)$ *and* $\mathcal{V}(\tau, \xi)$ *are* L^p *spaces, the following statements are equivalent:*

(a) *Controls* $\bar{u}(\cdot) \in \mathcal{U}(\tau, \xi)$, $\bar{v}(\cdot) \in \mathcal{V}(\tau, \xi)$ *are open-loop solutions of the game* $\mathcal{G}(\tau, \xi)$.

(b) $(\bar{u}(t), \bar{v}(t))$ *is a saddle point of* $S(t, u, v)$ *for almost all* $t \in [\tau, T]$.

The proof of Theorem 3.1 is an application of the following facts, which easily follow from 14.60 in Rockafellar and Wets [9].

LEMMA 3.3. *Let* $\gamma : [a, b] \times \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$ *be a function such that* $t \to \gamma(t, u(t), v(t))$ *is measurable for any* $u(\cdot) \in \mathcal{U}$, $v(\cdot) \in \mathcal{V}$, *where* $\mathcal{U}$ *and* $\mathcal{V}$ *are some sets of measurable functions. Define* $\Gamma(u(\cdot), v(\cdot)) = \int_a^b \gamma(t, u(t), v(t))dt$.

(a) *If* $\bar{u}(\cdot) \in \mathcal{U}$ *and* $\bar{v}(\cdot) \in \mathcal{V}$ *are such that* $(\bar{u}(t), \bar{v}(t))$ *is a saddle point of* $\gamma(t, \cdot, \cdot)$ *over* $\mathbb{R}^n \times \mathbb{R}^m$ *for almost all* $t \in [a, b]$, *then* $(\bar{u}(\cdot), \bar{v}(\cdot))$ *is a saddle point for* $\Gamma(\cdot, \cdot)$ *over* $\mathcal{U} \times \mathcal{V}$.

Assume additionally that $(t, u) \mapsto \gamma(t, u, v(t))$ *and* $(t, v) \mapsto -\gamma(t, u(t), v)$ *are normal integrands for any* $u(\cdot) \in \mathcal{U}$, $v(\cdot) \in \mathcal{V}$, *and that the following condition holds: for some sets* $U(t)$ *and* $V(t)$, *depending measurably on* t, *with the property that, for almost all* $t \in [a, b]$, $u(t) \in U(t)$ *and* $v(t) \in V(t)$ *whenever* $u(\cdot) \in \mathcal{U}$ *and* $v(\cdot) \in \mathcal{V}$, $\mathcal{U}$ *is*

decomposable with respect to $U(\cdot)$, and $\mathcal{V}$ is decomposable with respect to $V(\cdot)$. (This condition is automatically satisfied when $\mathcal{U}$ and $\mathcal{V}$ are decomposable.)

(b) *If $(\bar{u}(\cdot), \bar{v}(\cdot))$ is a saddle point for $\Gamma(\cdot,\cdot)$ over $\mathcal{U} \times \mathcal{V}$, and the saddle value is finite, then $(\bar{u}(t), \bar{v}(t))$ is a saddle point for $\gamma(t,\cdot,\cdot)$ over $\mathbb{R}^n \times \mathbb{R}^m$ for almost all $t \in [a,b]$.*

The additional assumption of normality of integrands preceding part (b) implies, in particular, that $t \mapsto \gamma(t, u(t), v(t))$ and $t \mapsto -\gamma(t, u(t), v(t))$ are measurable functions of t. An example of normal integrands is provided by Caratheodory integrands—functions $(t,z) \mapsto \eta(t,z)$ measurable in t and continuous in z. The other condition preceding (b) will later be invoked for the control sets $\mathcal{U} = \{u(\cdot) \in \mathcal{U}(\tau,\xi) \mid u(t) \in P(t)$ almost everywhere for $t \in [\tau,T]\}$ and $\mathcal{V} = \{v(\cdot) \in \mathcal{V}(\tau,\xi) \mid v(t) \in Q(t)$ almost everywhere for $t \in [\tau,T]\}$, with $P(\cdot)$ and $Q(\cdot)$ playing the role of $U(\cdot)$ and $V(\cdot)$.

PROPOSITION 3.4 (normal integrands). *The function $(t,u) \mapsto f(t,u,v(t))$ is a normal integrand for any measurable $v(\cdot)$ such that $v(t) \in Q(t)$ almost everywhere in $[\tau,T]$. Symmetrically, $(t,u) \mapsto -f(t,u(t),v)$ is a normal integrand for any measurable $u(\cdot)$ such that $u(t) \in P(t)$ almost everywhere in $[\tau,T]$.*

Proof. First, assume that $v(t) \in Q(t)$ almost everywhere in $[\tau,T]$. Then we have $f(\cdot,\cdot,v(\cdot)) = \tilde{f}(\cdot,\cdot,v(\cdot))$, where $\tilde{f}(t,u,v) = f(t,u,v)$ when $v \in Q(t)$ and $\tilde{f}(t,u,v) = +\infty$ elsewhere. We can view $\tilde{f}$ as a sum of a Caratheodory integrand $\hat{f}$ and an indicator of $P(t) \times Q(t)$; so, according to 14.32 in [9], $\tilde{f}$ is a normal integrand. The mentioned $\hat{f}$ can be, for example,

$$\hat{f}(t,u,v) = f\left(t, \Pi_{P(t)\times Q(t)}(u,v)\right),$$

where Π_S is the projection onto the set S. The expression $\Pi_{P(t)\times Q(t)}((u,v))$, by 14.17 in [9], is measurable in t, so $\hat{f}$ is also measurable in t for fixed (u,v). For a fixed time t, the projection is continuous in (u,v), so the same property holds for $\hat{f}$. Thus $\hat{f}$ is a Caratheodory integrand. The proof of the second part of the proposition is parallel. □

Proof of Theorem 3.1. Given the controls $u(\cdot)$ and $v(\cdot)$ and the initial condition (2), we obtain the trajectory

$$x(t) = \mathcal{A}(t,\tau)\xi + \int_\tau^t \mathcal{A}(t,s)\left(B(s)u(s) + C(s)v(s)\right) ds. \tag{12}$$

The cost (4) can be rewritten as

$$\int_\tau^T f(t,u(t),v(t))dt + d \cdot \left(\int_\tau^T \mathcal{A}(T,s)(B(s)u(s) + C(s)v(s))ds + \mathcal{A}(T,\tau)\xi \right)$$

$$= \int_\tau^T \left[f(t,u(t),v(t)) + d \cdot \mathcal{A}(T,t)(B(t)u(t) + C(t)v(t)) \right] dt + d \cdot \mathcal{A}(T,\tau)\xi. \tag{13}$$

The last term in the last expression is independent of the controls. To find the saddle point of (13), we can then concentrate on the integral part of this expression. Part (a) of Lemma 3.3 implies that condition (11) is sufficient. We now prove it is also necessary. The expression $d \cdot \mathcal{A}(T,t)(B(t)u + C(t)v)$ is continuous in (t,u,v). Thus it is a normal integrand in (t,u) for a fixed $v(\cdot)$, and its negative is a normal integrand in (t,v) for a fixed $u(\cdot)$. Similar properties hold for $f(t,u,v)$ by Lemma 3.4 and then also for the integrand in (13). The last statement follows from the fact that a

sum of normal integrands is a normal integrand. Let $(\bar{u}(\cdot), \bar{v}(\cdot))$ be a saddle point of $\Phi(\tau, \xi, u(\cdot), v(\cdot))$. Then $\bar{u}(\cdot)$ and $\bar{v}(\cdot)$ satisfy (3), and, by Assumption 2.3, the value of (13) for these controls is finite. The assumption of decomposability of $\mathcal{U}(\tau, \xi)$ with respect to $P(t)$ implies, in particular, that the set $\mathcal{U}$, as defined in the comments following Lemma 3.3, is decomposable with respect to $P(t)$. Symmetric statements can be made for $\mathcal{V}(\tau, \xi)$. Applying Lemma 3.3 to the control sets $\mathcal{U}$, $\mathcal{V}$ and the constraint sets $P(t)$, $Q(t)$ finishes the proof. □

4. Existence of saddle controls. The condition that $(\bar{u}(t), \bar{v}(t))$ is a saddle point of $S(t, u, v)$ is equivalent, by 37.4 in Rockafellar [6], to either of the following expressions:

$$(-B^*(t)\mathcal{A}^*(T,t)d, -C^*(t)\mathcal{A}^*(T,t)d) \in \partial f\,(t, \bar{u}(t), \bar{v}(t))\,, \tag{14}$$

$$(\bar{u}(t), \bar{v}(t)) \in \partial f^*(t, -B^*(t)\mathcal{A}^*(T,t)d, -C^*(t)\mathcal{A}^*(T,t)d). \tag{15}$$

In the above formulas and in what follows, $\mathcal{A}^*(T,t)$, $B^*(t)$, and $C^*(t)$ denote the transposes of the matrices $\mathcal{A}(T,t)$, $B(t)$, and $C(t)$.

THEOREM 4.1 (existence of saddle controls). *Measurable controls $u(\cdot)$ and $v(\cdot)$ satisfying* (15) *for $t \in (-\infty, T]$ exist.*

(a) *If, in addition, for every fixed (p, q), $f^*(t, p, q)$ is locally L^1 in t, the controls are locally L^1 functions.*

(b) *If, in addition, $f^*(t, p, q)$ is continuous in (t, p, q), then the controls are locally L^∞ functions.*

In the proof of the theorem, we will use following lemma, taken from Rockafellar and Wolenski [10]. We need a definition first: for a function $\gamma : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a point x with $\gamma(x)$ finite, the general subdifferential $\partial^g\gamma(x)$ is the set of all y, such that there exist sequences $x^\nu \to x$, with $\gamma(x^\nu) \to \gamma(x)$, and $y^\nu \to y$, with each pair x^ν, y^ν satisfying the condition $\gamma(x') \geq \gamma(x^\nu) + y^\nu \cdot (x' - x^\nu) + o(|x' - x^\nu|)$.

LEMMA 4.2. *Let $h(\cdot, \cdot)$ be a finite, convex-concave function. Then*

$$\partial h(p, q) = \operatorname{con} \partial^g h(p, q). \tag{16}$$

Proof. By 12.27 in [9], the mapping $T(\cdot, \cdot)$ defined by $(u, -v) \in T(p, q)$ whenever $(u, v) \in \partial h(p, q)$ is maximal monotone. By 35.8 in [6], $T(p, q)$ is single valued if and only if $h(\cdot, \cdot)$ is differentiable at (p, q), and this is the case for almost all (p, q), since $h(\cdot, \cdot)$ is locally Lipschitz continuous. Then the structure of monotone mappings, as described in 12.67 in [9] implies that

$$T(p, q) = \operatorname{con}\{(u, -v) \mid \exists (p^\nu, q^\nu) \to (p, q) \text{with } \nabla h(p^\nu, q^\nu) \to (u, v)\}. \tag{17}$$

We can now apply 9.61 in [9] to conclude that $\partial h(p, q) = \operatorname{con} \partial^g h(p, q)$. □

Proof of Theorem 4.1. Since $f^*(t, p, q)$ is finite, the right side of the inclusion (15) is a nonempty compact convex set. We now argue that it depends measurably on t. For fixed p and q, $f^*(t, p, q)$ is measurable in t. It follows from the fact that conjugacy in the convex sense preserves measurability in time; applying this twice to $f(\cdot, u, v)$ gives us measurability of $f^*(t, p, q)$. Measurability in t and continuity in (p, q) mean that $f^*(t, p, q)$ is a Caratheodory integrand and so also a normal integrand. By Lemma 4.2

$$\partial f^*(t, B^*(t)y(t), C^*(t)y(t)) = \operatorname{con} \partial^g f^*(t, B^*(t)y(t), C^*(t)y(t)).$$

The subdifferential on the right side depends measurably on time, by 14.56 in [9]. Taking the convex hull preserves measurability, by 14.12 in [9]. Therefore, the right side of the inclusion (15) is measurable in time with nonempty compact convex values. By 14.6 in [9], there exists a measurable selection, that is, a pair of measurable functions $u(\cdot)$ and $v(\cdot)$ satisfying (15). Applying Lemma 3 from Rockafellar [7] to the function $f^*(t,p,q)$ implies part (a). If $f^*(t,p,q)$ is continuous, then it is epi/hypocontinuous in t, which implies the graphical continuity of $\partial f^*(t,\cdot,\cdot)$—see Rockafellar [8]. In particular, the graph of $\partial f^*(t,p,q)$ is locally bounded. For $t \in [\tau, T]$,

$$(t, -B^*(t)\mathcal{A}^*(T,t)d, -C^*(t)\mathcal{A}^*(T,t)d) \in K$$

for some compact set K, so the right side of the inclusion (15) is bounded. This implies part (b). □

If any pair of controls $(\bar{u}(\cdot), \bar{v}(\cdot))$ satisfying (15) is such that $\bar{u}(\cdot)$, $\bar{v}(\cdot)$ are in the desired control spaces $\mathcal{U}(\tau,\xi)$, $\mathcal{V}(\tau,\xi)$, the game $\mathcal{G}(\tau,\xi)$ has open-loop solutions; this is guaranteed by Theorem 3.1. General conditions guaranteeing that this is the case are as follows.

COROLLARY 4.3 (existence of open-loop solutions). *Under either of the following two conditions, the game $\mathcal{G}(\tau,\xi)$ has open-loop solutions.*

(a) *The condition in part* (a) *of Theorem* 4.1 *holds, and the control sets $\mathcal{U}(\tau,\xi)$ and $\mathcal{V}(\tau,\xi)$ contain all locally integrable functions.*

(b) *The condition of part* (b) *of Theorem* 4.1 *holds, and the control sets $\mathcal{U}(\tau,\xi)$ and $\mathcal{V}(\tau,\xi)$ contain all essentially bounded functions,*

The solutions are independent of τ and ξ in the following sense: there exist functions $\bar{u}_\infty : (-\infty, T] \mapsto \mathbb{R}^k$, $\bar{v}_\infty : (-\infty, T] \mapsto \mathbb{R}^l$ such that, for any (τ,ξ), the truncations of $\bar{u}_\infty$, $\bar{v}_\infty$ to $[\tau, T]$ are open-loop solutions to $\mathcal{G}(\tau,\xi)$.

In the case where $f(t,u,v) = g(t,u) - h(t,v)$ for some functions $g(t,u)$ and $h(t,v)$ convex in u and v, the conjugate function is $f^*(t,p,q) = g^*(t,p) - h^*(t,q)$. Here $g^*(t,\cdot)$ and $h^*(t,\cdot)$ denote convex functions conjugate to $g(t,\cdot)$ and $h(t,\cdot)$. Condition (15) can be written as

$$\bar{u}(t) \in \partial g^*(t, -B^*(t)\mathcal{A}^*(T,t)d) \quad \text{and} \quad -\bar{v}(t) \in \partial h^*(t, -C^*(t)\mathcal{A}^*(T,t)d).$$

In this case, not only do the saddle controls $\bar{u}(\cdot)$ and $\bar{v}(\cdot)$ not depend on τ and ξ, but they can be chosen independently of each other. Indeed, to choose $\bar{u}(\cdot)$, Player One needs only to know the function $g(\cdot,\cdot)$ and matrices A and B. That player's choice does not depend on $h(\cdot,\cdot)$ or C.

We now comment on the special structure of the inclusion (15) for the game discussed in Example 2.6 ($P(t)$, $Q(t)$ polyhedral, $f(t,\cdot,\cdot)$ quadratic, given by (9)). Let $N_C(y)$ denote the normal cone to the set C at y. (If C is polyhedral, then so is $N_C(y)$.) We have, for any $(u,v) \in P(t) \times Q(t)$,

$$\partial f(t,u,v) = \left\{E(t)u + G^*(t)v + N_{P(t)}(u)\right\} \times \left\{-F(t)v + G(t)u - N_{Q(t)}(v)\right\}.$$

For points $(u,v) \notin P(t) \times Q(t)$, the subdifferential of $f(t,\cdot,\cdot)$ is empty. The graph of the mapping $\partial f(t,\cdot,\cdot)$ is thus piecewise polyhedral—it consists of a union of finitely many polyhedral sets. The same holds for $\partial f^*(t,\cdot,\cdot)$, as $\partial f^*(t,\cdot,\cdot) = (\partial f(t,\cdot,\cdot))^{-1}$, and so the graphs of the two mappings are invertible linear images of one another. In particular, the right-hand side of the inclusion (15) is always a polyhedral set (possibly a singleton).

If the matrices defining f depend continuously on t, we can conclude that the right-hand side of (15) is locally bounded. Indeed, such an assumption guarantees the so-called epi/hypocontinuity of f in t, which in turn is equivalent to the graphical continuity of ∂f, as well as ∂f^*, in t. As the latter subdifferential is always nonempty, it must be locally bounded (in all variables). This allows us to conclude that the game in Example 2.6 has open-loop solutions when the control sets $\mathcal{U}(\tau,\xi)$ and $\mathcal{V}(\tau,\xi)$ contain all essentially bounded functions (as is the case when the control sets are L^2 spaces—a natural choice for games with quadratic costs).

Exploring the structure of the subdifferential of a saddle function, as described in (17), we can say the following: when the matrices E, F, G do not depend on t, the right-hand side of (15) can be written as $\operatorname{con}\{r_1(t), r_2(t), \dots, r_j(t)\}$ for some piecewise continuous functions $r_i(\cdot)$ (with the possibility that all r_i agree for some t). In particular, the right-hand side of (15) depends "piecewise continuously" on t.

5. Hamiltonian system. We define the Hamiltonian, namely, the function $H : (-\infty, T] \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, as the saddle value of the concave-convex function $(u,v) \to y \cdot (A(t)x + B(t)u + C(t)v) - f(t,u,v)$; that is,

$$H(t,x,y) = \sup_u \inf_v \{y \cdot (A(t)x + B(t)u + C(t)v) - f(t,u,v)\}.$$

By the definition of f^*, we obtain

$$H(t,x,y) = y \cdot A(t)x + f^*(t, B^*(t)y, C^*(t)y). \tag{18}$$

We now give another characterization of saddle controls, in terms of Clarke subdifferential $\partial^c H$ of the Hamiltonian. For any locally Lipschitz function $\psi(\cdot)$,

$$\partial^c \psi(\bar{z}) = \operatorname{con}\{\lim \nabla \psi(z^\nu) \mid x^\nu \to x\}, \tag{19}$$

where the limits are taken over all sequences $\{x^\nu\}$ of points where ψ is differentiable (see Clarke [4] for details). Below, $\partial^c H(t,x,y)$ denotes the Clarke subdifferential of $H(t,\cdot,\cdot)$. Whenever $f^*(t,p,q)$ is differentiable in (p,q), the Hamiltonian is differentiable in (x,y), and $\partial^c H(t,x,y)$ reduces to $\nabla_x H(t,x,y) \times \nabla_y H(t,x,y)$. Note that, since $f(t,\cdot,\cdot)$ is a finite convex-concave function, the proof of Lemma 4.2 shows that $\partial f^*(t,p,q) = \partial^c f^*(t,p,q)$.

THEOREM 5.1 (generalized Hamiltonian system). *If the Hamiltonian inclusion*

$$(-\dot{y}(t), \dot{x}(t)) \in \partial^c H(t, x(t), y(t)) \tag{20}$$

holds for almost all $t \in [\tau, T]$ *and*

$$-y(T) = d, \tag{21}$$

then $x(\cdot)$ *satisfies the dynamics* (1) *for some controls* $u(\cdot)$ *and* $v(\cdot)$ *satisfying the saddle condition* (15). *If, for every* $t \in [\tau, T]$, *either* $f^*(t,p,\cdot)$ *is differentiable for every* p *or* $f^*(t,\cdot,q)$ *is differentiable for every* q, *then the reverse implication holds.*

Proof. Directly from the definition (19) we get that

$$\partial^c H(t,x,y) \subset (A^*(t)y, A(t)x + [B(t), C(t)]\, \partial^c f^*(t, B^*(t)y, C^*(t)y)), \tag{22}$$

where $\partial^c f^*(t,p,q)$ denotes the Clarke subdifferential of $f^*(t,p,q)$ in (p,q). The Hamiltonian condition (20) now reduces to

$$-\dot{y}(t) = A^*(t)y(t),$$

$$\dot{x}(t) \in A(t)x(t) + [B(t), C(t)]\, \partial^c f^*(t, B^*(t)y(t), C^*(t)y(t)).$$

The first equation, combined with the transversality condition (21) implies that $y(t) = -\mathcal{A}^*(T,t)d$. We now concentrate on the second inclusion. The Clarke subdifferential $\partial^c f^*(t,\cdot,\cdot)$ is equal to $\partial f^*(t,\cdot,\cdot)$. The inclusion becomes

$$\dot{x}(t) \in A(t)x(t) + [B(t), C(t)]\, \partial f^*(t, B^*(t)y(t), C^*(t)y(t)).$$

By remarks made in the proof of Theorem 4.1, $\partial f^*(t, B^*(t)y(t), C^*(t)y(t))$ is measurable in t. Let $E(t) = [B(t), C(t)]$. The mapping $(t,w) \to E(t)w$ is a Caratheodory mapping. For almost all $t \in [\tau, T]$, there exists a $w \in \partial f^*(t, B^*(t)y(t), C^*(t)y(t))$ such that $E(t)w \in \dot{x}(t) - A(t)x(t)$, and the mapping on the right side of the inclusion is single (so closed) valued and measurable. We can extend this mapping to the whole interval $[\tau, T]$ by assigning it an empty value whenever $\dot{x}(t)$ does not exist; this does not change the closed valuedness or measurability. Theorem 14.16 in [9] implies that there exists a measurable $w(\cdot)$ defined on a full measure subset $S \subset [\tau, T]$, with $w(t) \in \partial f^*(t, B^*(t)y(t), C^*(t)y(t))$, for all $t \in S$ such that

$$\dot{x}(t) = A(t)x(t) + E(t)w(t).$$

We can now write $w(t)$ as $(u(t), v(t))$, where $(u(\cdot), v(\cdot))$ satisfies (15), since $y(t) = -\mathcal{A}^*(T,t)d$. The first part of the theorem is proved.

Now assume that for a fixed t, $f^*(t,p,\cdot)$ is differentiable for every p. Then $f^*(t,\cdot,\cdot)$ is subdifferentially regular at $(B^*(t)y, C^*(t)y)$ in the sense of 7.25 in [9], and by 10.6 in [9], $[B(t), C(t)]\, \partial^c f^*(t, B^*(t)y, C^*(t)y)$ is the Clarke subdifferential of $f^*(t,\cdot,\cdot)$ at $(B^*(t)y, C^*(t)y)$ with respect to y. The inclusion (22) becomes an equation, and all of the above arguments can be reversed. If $f^*(t,\cdot,q)$ is differentiable, we can make an argument, similar to the one above, for the function $-f^*(t,\cdot,q)$. □

COROLLARY 5.2. *Assume that $x(\cdot)$ with $x(\tau) = \xi$ and $y(\cdot)$ with $-y(T) = d$ satisfy the Hamiltonian inclusion* (20) *for almost all $t \in [\tau, T]$. Then $x(\cdot)$ is an equilibrium trajectory of the game $\mathcal{G}(\tau, \xi)$.*

6. The value function. The value $W(\tau, \xi)$ of the game is defined to be the saddle value of the game $\mathcal{G}(\tau, \xi)$. Let $\mathcal{U}$ and $\mathcal{V}$ be some sets of controls on the interval $(-\infty, T]$, such that any measurable and locally integrable solution $(\bar{u}(\cdot), \bar{v}(\cdot))$ of (15) satisfies $\bar{u}(\cdot) \in \mathcal{U}$ and $\bar{v}(\cdot) \in \mathcal{V}$. Assume that, for every $(\tau, \xi) \in (-\infty, T] \times \mathbb{R}^n$, the control sets $\mathcal{U}(\tau, \xi)$ and $V(\tau, \xi)$ are the restrictions of $\mathcal{U}$ and $\mathcal{V}$ to the interval $[\tau, T]$. Then the value function $W(\cdot,\cdot)$ is well defined. In particular, there exist measurable and locally integrable functions $\bar{u}_\infty(\cdot)$ and $\bar{v}_\infty(\cdot)$ on $(-\infty, T]$ such that, for every (τ, ξ), the value function is given by $W(\tau, \xi) = \Phi(\tau, \xi, \bar{u}(\cdot), \bar{v}(\cdot))$, where $\bar{u}(\cdot)$ and $\bar{v}(\cdot)$ are the restrictions of $\bar{u}_\infty(\cdot)$ and $\bar{v}_\infty(\cdot)$ to $[\tau, T]$. Looking at the cost expression in (13), we get that, for almost all (τ, ξ),

$$\nabla_\xi W(\tau, \xi) = \mathcal{A}^*(T, \tau)d,$$

$$W_\tau(\tau, \xi) = -f(\tau, \bar{u}(\tau), \bar{v}(\tau)) - d \cdot \mathcal{A}(T, \tau)(A\xi + B\bar{u}(\tau) + C\bar{v}(\tau)).$$

Whenever both partial derivatives exist, the Hamilton–Jacobi equation holds:

$$-W_\tau(\tau, \xi) + H(\tau, \xi, -\nabla_\xi W(\tau, \xi)) = 0. \tag{23}$$

More can be said in the case where the functions $\bar{u}_\infty(\cdot)$ and $\bar{v}_\infty(\cdot)$ are continuous.

THEOREM 6.1. *Assume that $f^*(\cdot,\cdot,\cdot)$ is continuous in all three variables and, for all $t \in (-\infty, T]$, $f^*(t,\cdot,\cdot)$ is differentiable. Then the value function $W(\cdot,\cdot)$ is continuously differentiable and satisfies the Hamilton–Jacobi equation* (23).

The Hamilton–Jacobi equation allows us to rewrite the auxiliary saddle function (5) as

$$S(t,u,v) = f(t,u,v) + \nabla_\xi W(\tau,\xi)(B(t)u + C(t)v). \tag{24}$$

Theorem 11 states that any controls $\bar{u}(\cdot)$, $\bar{v}(\cdot)$ such that $(\bar{u}(t), \bar{v}(t))$ is a saddle point of (24) almost everywhere are saddle controls of the game. A similar result was obtained by Subbotin [12] under different assumptions and in the setting of closed-loop controls. A solution of (23) was used there to generate closed-loop saddle controls of the game, as saddle points of (24).

If, in addition to the assumptions of the above theorem, the Hamiltonian function is Lipschitz continuous in the y variable, the value function is the unique solution of (23) with the boundary condition

$$W(T,\xi) = d \cdot \xi \tag{25}$$

not only in the classical sense but in the minimax sense. For definitions and the proof, see Subbotin [12]. Note that the Hamiltonian is Lipschitz continuous in y, in particular when $f^*(t,\cdot,\cdot)$ has this property. We finish the paper with the characterization of this case, showing that $f^*(t,\cdot,\cdot)$ is globally Lipschitz continuous if and only if the control sets $P(t)$ and $Q(t)$ are bounded.

PROPOSITION 6.2 (Lipschitz continuity of the conjugate function). *Let $h(\cdot,\cdot)$ be a closed convex-concave function, and let K be the nonempty set where $h(\cdot,\cdot)$ is finite valued. The following statements are equivalent:*

(a) *The class of functions conjugate to $h(\cdot,\cdot)$ contains a globally Lipschitz continuous function $h^*(\cdot,\cdot)$. (Note that this actually implies that $h^*(\cdot,\cdot)$ is the unique function in the mentioned class.)*

(b) *The set K is bounded.*

Proof. Recall that the subdifferential $\partial h^*(\cdot,\cdot)$ of the saddle function $h^*(\cdot,\cdot)$ equals $\operatorname{con} \partial^g h^*(\cdot,\cdot)$, where $\partial^g h(\cdot,\cdot)$ is the generalized subdifferential in the sense of [9]. Also recall that if $h^*(\cdot,\cdot)$ is finite, then $\partial h^*(\cdot,\cdot)$, so also $\partial^g h^*(\cdot,\cdot)$, is locally bounded. This fact will allow us to use Theorem 9.13 of Rockafellar and Wets [9]. Note also that the boundedness of K implies that the class of functions conjugate to $h(\cdot,\cdot)$ consists of a unique finite function $h^*(\cdot,\cdot)$ and that the global Lipschitz continuity of the latter function entails its finiteness.

The boundedness of K is equivalent to the boundedness of $\operatorname{rge} \partial h^*(\cdot,\cdot)$, the range of the subdifferential mapping $\partial h^*(\cdot,\cdot)$, by Theorems 37.4 and 37.5 in Rockafellar [6]. This is equivalent to $\partial^g h^*(\cdot,\cdot)$ being globally bounded and, by Theorem 9.13 in [9], to the local Lipschitz modulus of $h^*(\cdot,\cdot)$ being globally bounded. This, in turn, is equivalent to $h^*(\cdot,\cdot)$ being globally Lipschitz continuous; see Theorem 9.2 in [9]. □

Acknowledgments. The author wishes to express his gratitude to Arkadii Kryazhimskii and Alexander Tarasiev for introducing him to the field of differential games during the author's stay at the International Institute for Applied Systems Analysis in Laxenburg, Austria.

REFERENCES

[1] T. Basar and G. J. Olsder, *Dynamic Noncooperative Game Theory*, Math. Sci. Engrg. 160, Academic Press, New York, 1982.

[2] L. D. Berkovitz, *A variational approach to differential games*, in Advances in Game Theory, Princeton University Press, Princeton, NJ, 1964, pp. 127–174.

[3] L. D. Berkovitz, *Lectures on differential games*, in Differential Games and Related Topics, H. W. Kuhn and G. P. Sego, eds., North–Holland, Amsterdam, 1971, pp. 3–45.

[4] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Canad. Math. Soc. Ser. Monogr. Adv. Texts, John Wiley, New York, 1983.

[5] G. Leitmann and H. Stalford, *Sufficiency for optimal strategies in Nash equilibrium games*, in Techniques of Optimization, A. V. Balakrishnan, ed., Academic Press, New York, 1972.

[6] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[7] R. T. Rockafellar, *Generalized Hamiltonian equations for convex problems of Lagrange*, Pacific J. Math., 33 (1970), pp. 411–427.

[8] R. T. Rockafellar, *Generalized second derivatives of convex functions and saddle functions*, Trans. Amer. Math. Soc., 322 (1990), pp. 51–77.

[9] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1997.

[10] R. T. Rockafellar and P. R. Wolenski, *Convexity in Hamilton-Jacobi theory* I: *Dynamics and duality*, SIAM J. Control Optim., 39 (2000), pp. 1323–1350.

[11] R. C. Scalzo, *Existence of equilibrium points in N-person differential games*, in Differential Games and Control Theory, O. Roxin, P. T. Liu, and R. L. Sternberg, eds., Marcel Dekker, New York, 1974, pp. 125–140.

[12] A. I. Subbotin, *Generalized Solutions of First-Order PDEs: The Dynamical Optimization Perspective*, Birkhäuser Boston, Boston, 1995.

[13] B. Tolwinski, *On the existence of Nash equilibrium points for differential games with linear and nonlinear dynamics*, Control Cybernet., 7 (1978), pp. 57–69.

[14] P. Varaiya, *N-person nonzero sum differential games with linear dynamics*, SIAM J. Control, 8 (1970), pp. 441–449.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 5, pp. 1505–1516

DISSIPATIVE CONTROL OF MECHANICAL SYSTEMS: A GEOMETRIC APPROACH*

MIGUEL C. MUÑOZ-LECANDA† AND F. JAVIER YÁNIZ-FERNÁNDEZ†

Abstract. Dissipative and passive mechanical systems are studied from a geometric point of view. Since the natural geometric background is a Riemannian manifold, we begin by generalizing La Salle theorems about the stability of equilibrium points of dynamical systems to a complete Riemannian manifold. The stability of dissipative mechanical systems is studied using the particular geometric properties of the tangent bundle, and passivity based controls are designed to stabilize equilibrium points. The case of partially dissipative systems is formulated and used with a dynamical extension to design controls for bringing the system to a desired point of the phase space.

Key words. dissipative control, mechanical systems, passive systems, stability of equilibrium points

AMS subject classifications. 53C20, 70G05, 93B29, 93B52, 93D20

PII. S0363012900374804

1. Introduction.

Although the specific properties of stability of the equilibrium points of dissipative systems have been well known for years, the systematic study of such systems is fairly recent. Indeed, the study of passivity control, based on these properties, has received wide attention, as is evident from the extensive scientific and technical literature on the subject.

The results and the usual methods, however, are only local; they use coordinate expressions and conceal the geometric properties of the state spaces and the vector fields defining the systems.

Furthermore, since the 1970s the use of geometric intrinsic methods has proven to be powerful for describing the interesting properties of control systems, shedding light on the relation between them, and improving the methods used to design the control. A look at the papers by Brockett, Millman, and Sussmann [5], Isidori [7], Nijmeijer and van der Schaft [9], or Lewis and Murray [8], and the references quoted there is enough to ensure that these methods are worthy of being studied and used.

In this paper we give a geometric intrinsic formulation of some aspects of mechanical dissipative systems and the use of these properties to design controls by passivity.

The natural geometric background for describing these systems is a Riemannian manifold, where dissipative vector fields are naturally formulated, including as a particular case those coming from a Rayleigh function. This Riemannian manifold is the *configuration space* of the system.

A mechanical system in a Riemannian manifold, (M, g), is given by a vector field, which is the external force field. The most simple case is that in which this vector field is the gradient of a function, the potential function, but the force often has additional components. We will suppose that one of these components is dissipative and the other contains the control forces. With these data we can write the equation of motion for

*Received by the editors July 3, 2000; accepted for publication (in revised form) July 5, 2001; published electronically January 18, 2002. This work was supported by the Spanish CICYT PB98-0920 project.

http://www.siam.org/journals/sicon/40-5/37480.html

†Departamento de Matemática Aplicada IV, Universidad Politécnica de Cataluña, Edificio C–3, C/Jordi Girona 1, E-08034 Barcelona, Spain (matmcml@mat.upc.es, yaniz@mat.upc.es).

a mechanical system with controls, the commonly called *Newton equation*, which is equivalent to *Euler–Lagrange equations*. If, however, we wish to write this equation of motion as a dynamical system—that is, as the equation for the integral curves of a vector field in a manifold, since it is a second order differential equation—we must go to the tangent bundle of the initial configuration manifold. This tangent bundle is the *phase space* of coordinate-velocities of the mechanical system.

Taking this as the natural starting point, the aim of this paper is twofold: first, to describe geometrically the stability properties of equilibrium points of a dissipative vector field in a Riemannian manifold, and second, to use these properties to design passivity controls on mechanical systems in order to stabilize an equilibrium point or to bring the system to a desired working point. We have used [10], [12], [14], and the corresponding chapters of [11] and [13] as a natural guide to the problem in a classical context. They contain enough applications of the results of our study to assure the interest of our approximation to the problem.

The paper is organized as follows. Sections 2 and 3 are devoted to stating the systems we are going to use and the suitable notations for mechanical systems, dissipative and passive systems. In section 4, La Salle's theorem is generalized to a complete vector field in a complete Riemannian manifold, and in section 5 it is applied to passive systems. In section 6, we study the stability of equilibrium points in dissipative mechanical systems, including those which are partially dissipative. Finally, section 7 is devoted to stabilizing equilibrium points of mechanical systems using passivity controls, and to bringing the system to a desired point by means of a dynamical dissipative extension.

Throughout this paper we suppose that the manifolds and maps are differentiable, and differentiability means infinite differentiability. The manifolds are Hausdorff. As a reference for differential geometry, notation employed, and concepts, see [1] and [3].

2. Dynamical and control systems.

2.1. Dissipative mechanical systems. Let M be a differentiable manifold. A *dynamical system* on M is a vector field Y in M. M is called the *phase space* of the system. The differential equation associated to Y, that is, the equation of the integral curves of Y, is given by $Y \circ \gamma(t) = \dot{\gamma}(t)$, where γ is a curve in M. It is called the *evolution equation* of the system.

The vector field $Y \in \mathfrak{X}(M)$ defines a *dissipative system* if there exists a nonnegative function $S \in C^\infty(M)$ which is *decreasing* along the integral curves of Y; that is, $\mathcal{L}_Y S \leq 0$, where $\mathcal{L}_Y$ is the Lie derivative with respect to the vector field Y. S is called the *dissipation function* for Y. If $\mathcal{L}_Y S < 0$, that is, S is *strictly decreasing* along the integral curves of Y, then we call the system *strictly dissipative*.

2.2. Control systems and passive systems. A *control system* on a manifold M is given by a vector field X, called the *drift*, and the *control vector fields* $X_1, \ldots, X_m$. The vector field associated to the system is $Y = X + \sum_{i=1}^m X_i u_i$, where $u_i : M \longrightarrow \mathbb{R}$ are the *controls* or *inputs* of the system.

Usually, for a control system, in addition to the evolution equation, we have the *outputs* of the system, functions $y_1 = h_1, \ldots, y_m = h_m$, $h_i \in C^\infty(M)$, related to the observable variables of the system.

Following [4], we introduce the idea of *passivity* as the counterpart for control systems of a dissipative system. Consider a control system given by $Y = X + \sum_{i=1}^m X_i u_i$ on the manifold M with outputs $y_1, \ldots, y_m$. It is *passive* if there exists a nonnegative

function $S \in C^\infty(M)$ such that

$$S(\gamma(T)) - S(\gamma(0)) \leq \int_0^T \sum_{i=1}^m y_i(\gamma(s))u_i(\gamma(s))\,ds$$

for every solution $\gamma(t)$ with the given inputs $u_1, \ldots, u_m$. The function S is called the *storage function*. This last expression can be understood as an inequality in the evolution of the energy of the system and will be clarified when we apply these ideas to mechanical systems.

3. Mechanical systems. Let (Q, g) be a Riemannian manifold, $\dim Q = n$, and let ∇ be the Levi–Civita connection associated to the Riemannian metric g. A differentiable curve in Q, $\gamma\colon I \longrightarrow Q$, is a *trajectory*, or a *solution*, of the mechanical system associated to the vector field $F \in \mathfrak{X}(Q)$ if γ satisfies the differential equation

$$\nabla_{\dot\gamma}\dot\gamma = F \circ \gamma, \tag{1}$$

where $\dot\gamma(t)$ is the tangent vector of the curve γ at the point $\gamma(t)$ and $\nabla_{\dot\gamma}\dot\gamma$ is the covariant derivative of $\dot\gamma(t)$ with respect to the tangent vector $\dot\gamma(t)$. F is called the *external force* vector field. In local coordinates (q^i) on Q, the components $(\gamma^1, \ldots, \gamma^n)$ of γ satisfy the differential equations

$$\ddot\gamma^k + \sum_{i,j=1}^n \Gamma_{ij}^k \dot\gamma^i \dot\gamma^j = F^k(\gamma^1, \ldots, \gamma^n), \qquad k = 1, \ldots, n, \tag{2}$$

where $\{\Gamma_{ij}^k\}$ are the Christoffel symbols of the connection ∇ in the given coordinates. Equations (1) and (2) are the classical *Newton equations of movement*.

If there exists a differentiable function $U\colon Q \to \mathbb{R}$ such that the force field $F = -\operatorname{grad} U$, we call U the *potential function*. A *simple mechanical system* is given by a Riemannian manifold, (Q, g), the *configuration space*, and a force field $F = -\operatorname{grad} U$. We write such systems as (Q, g, U).

Let $\mathrm{T}Q$ be the tangent bundle of Q, with $\tau_Q\colon \mathrm{T}Q \mapsto Q$ the natural projection; the *kinetic energy* is the function $T\colon \mathrm{T}Q \longrightarrow \mathbb{R}$ given by $T(v_q) = \frac{1}{2} g_q(v_q, v_q)$. In local coordinates, its expression is $T(q^i, v^i) = \frac{1}{2} g_{ij}(q) v^i v^j$, where (q^i, v^i) are the natural local coordinates on $\mathrm{T}Q$ associated to the coordinates (q^i) on Q.

For simple mechanical systems, the *Lagrangian function* is defined by $L = T - \tau_Q^* U$. For simplicity we will write it as $L = T - U$.

Frequently, the force vector field, F, splits into two vector fields, and only one of these fields comes from a potential function. Then if $F = -\operatorname{grad} U + R$, the dynamical equations are written as

$$\nabla_{\dot\gamma}\dot\gamma = -\operatorname{grad} U \circ \gamma + R \circ \gamma, \tag{3}$$

and, using the Lagrangian function, $L = T - U$, the equivalent (see [1]) Euler–Lagrange equations are

$$\frac{d}{dt}\left.\frac{\partial L}{\partial v^i}\right|_{\dot\gamma} - \left.\frac{\partial L}{\partial q^i}\right|_{\dot\gamma} = \sum_{j=1}^n g_{ij} R^j \circ \gamma, \qquad i = 1, \ldots, n. \tag{4}$$

In this situation, the mechanical system is given by $\Sigma = (Q, g, U, R)$.

It is well known that the dynamical equations of mechanics are second order differential equations in the configuration manifold Q, so there is a corresponding first order equation in the tangent bundle $\mathrm{T}Q$, and we have a vector field $Y \in \mathfrak{X}(\mathrm{T}Q)$ associated to Newton equation (3) or to the equivalent Lagrange equations (4). In local coordinates (q^i, v^i) of $\mathrm{T}Q$, this vector field Y has the expression

$$\begin{aligned} Y &= \sum_{i=1}^{n} \left(v^i \frac{\partial}{\partial q^i} + \left(F^i - \sum_{j,k=1}^{n} \Gamma^i_{jk} v^j v^k \right) \frac{\partial}{\partial v^i} \right) \\ &= \sum_{i=1}^{n} \left(v^i \frac{\partial}{\partial q^i} - \sum_{j,k=1}^{n} \Gamma^i_{jk} v^j v^k \frac{\partial}{\partial v^i} \right) + \sum_{i=1}^{n} F^i \frac{\partial}{\partial v^i} = X_g + F^v, \end{aligned}$$

where X_g is the *geodesic field* of the Riemannian metric g, F^v is the vertical lift from Q to $\mathrm{T}Q$ of the force vector field F, $F = -\mathrm{grad}\, U + R$, and we have that $\sum_{j=1}^{n} g_{ij} F^j = -\frac{\partial U}{\partial q^i} + \sum_{j=1}^{n} g_{ij} R^j$. Notice that the vector field $Y \in \mathfrak{X}(\mathrm{T}Q)$ satisfies the second order condition. It is a holonomic vector field in TQ; hence its integral curves in $\mathrm{T}Q$ are canonical liftings of curves in the manifold Q. This tangent bundle, $\mathrm{T}Q$, is called the *phase space* of the mechanical system. See [1] for details on $\mathrm{T}Q$ and properties of the second order vector fields.

When working with mechanical systems, it is usual to consider forces which depend on the velocities. These are vector fields on the manifold Q along the projection $\tau_Q \colon \mathrm{T}Q \longrightarrow Q$; that is, mappings $R \colon \mathrm{T}Q \longrightarrow \mathrm{T}Q$ such that $\tau_Q \circ R = \tau_Q$. We denote by $\mathfrak{X}(Q, \tau_Q)$ the set of such fields.

The dynamical equations of a mechanical system with the force vector field given by $F = -\mathrm{grad}\, \mathrm{U} + R$ with $R \in \mathfrak{X}(Q, \tau_Q)$ are $\nabla_{\dot\gamma} \dot\gamma = -(\mathrm{grad}\, \mathrm{U}) \circ \gamma + R \circ \dot\gamma$. Observe that in this equation we have $R \circ \dot\gamma$ instead of $R \circ \gamma$, because R depends on the velocities.

3.1. Dissipative mechanical systems. Given $R \in \mathfrak{X}(Q, \tau_Q)$, we call it *dissipative* if it satisfies $g_p(R(v_q), v_q) \leq 0$ for every $v_q \in \mathrm{T}Q$. A *dissipative mechanical system* is a mechanical system, $\Sigma = (Q, g, U, R)$ with R dissipative. If $R \in \mathfrak{X}(Q, \tau_Q)$ verifies $g_q(R(v_q), v_q) \leq -\alpha g_q(v_q, v_q)$ for every $v_q \in \mathrm{T}Q$, with α a real number, $\alpha > 0$, then R is called *strictly dissipative* and Σ is a *strictly dissipative system.*

Recall that the energy associated to a mechanical system given by $\Sigma = (Q, g, U, R)$ is the function $E = T + \tau_Q^* U \in C^\infty(\mathrm{T}Q)$. The vector field R is called dissipative because of the following proposition.

PROPOSITION 3.1. *Let $\Sigma = (Q, g, U, R)$ be a mechanical system. If the vector field R is dissipative, then the energy decreases along the solutions. Moreover, if R is strictly dissipative and γ satisfies the condition that $\dot\gamma(t) \neq 0$ for all t, then E is strictly decreasing.*

Proof. Let $\gamma \colon I \longrightarrow Q$ be a trajectory of the mechanical system. Then we have that

$$\begin{aligned} \tfrac{d(E \circ \dot\gamma)}{dt} &= \nabla_{\dot\gamma}(E \circ \dot\gamma) = \nabla_{\dot\gamma}(T \circ \dot\gamma + (\tau_Q^\star U) \circ \dot\gamma) \\ &= \nabla_{\dot\gamma}(\tfrac{1}{2} g(\dot\gamma, \dot\gamma)) + \nabla_{\dot\gamma}(U \circ \gamma) = g(\nabla_{\dot\gamma}\dot\gamma, \dot\gamma) + \nabla_{\dot\gamma}(U \circ \gamma) \\ &= g(-(\mathrm{grad}\, U) \circ \gamma, \dot\gamma) + g(R(\dot\gamma), \dot\gamma) + \nabla_{\dot\gamma}(U \circ \gamma) \\ &= g(R \circ \dot\gamma, \dot\gamma) \leq 0. \qquad \square \end{aligned}$$

So if $\Sigma = (Q, g, U, R)$ is a mechanical system and R is a dissipative force, then the associated vector field $Y = X_g - (\mathrm{grad}\, U)^v + R^v$ defines a dissipative system

in the phase space $\mathrm{T}Q$, provided that $U(q) \geq 0$ for every $q \in Q$. In this case, the dissipation function is the total energy E. Observe that condition $U \geq 0$ is equivalent to supposing that the potential function is bounded below.

3.2. Mechanical systems with controls and passive mechanical systems. In *mechanical systems with controls*, the vector fields of control are forces, and the inputs are coefficients to modulate the strength of these forces. Then the equation of motion is

$$\nabla_{\dot\gamma}\dot\gamma = -(\operatorname{grad} U)\circ\gamma + (F_1u_1 + \cdots + F_mu_m)\circ\dot\gamma,$$

where $F_1, \ldots, F_m$ are the control forces, which often depend on the position and the velocity, and $u_i\colon \mathrm{T}Q \longrightarrow \mathbb{R}$ are the inputs.

The *natural outputs* (see [9]) of a mechanical system with controls are the functions $h_i(v_q) = g(F_i(v_q), v_q)$; that is, $h_i = \mathrm{i}(F_i)g$ for $i = 1, \ldots, m$, where $\mathrm{i}(F_i)g$ is the natural tensor contraction.

From now on, a simple mechanical control system with natural outputs will be denoted by $\Sigma = (Q, g, U, F_1, \ldots, F_m)$.

Consider now a mechanical system with a dissipative component R of the force and with controls denoted by $\Sigma = (Q, g, U, R, F_1, \ldots, F_m)$. According to the above definitions and comments, it can be regarded as a control system with the following characteristics:

–Phase space: $M = \mathrm{T}Q$.

–Vector field of the system: $Y = X_g + F^v + \sum_{i=1}^m u_iX_i$, with $F = -\operatorname{grad} U + R$ and $X_i = F_i^v$.

–Natural outputs: $y_i = h_i = \mathrm{i}(F_i)g \in C^\infty(\mathrm{T}Q)$, $j = 1, \ldots, m$.

As pointed out above, the vector field $Y \in \mathfrak{X}(\mathrm{T}Q)$ satisfies the second order condition on $\mathrm{T}Q$.

PROPOSITION 3.2. *Suppose that $U(q) \geq 0$ for every $q \in Q$, or equivalently, that U is bounded from below; then the system Σ is a passive system with the mechanical energy $E = T + U$ as a storage function.*

Proof. From Proposition 3.1 we have

$$\frac{d(E\circ\dot\gamma)}{dt} = g(R\circ\dot\gamma, \dot\gamma) + \sum_{i=1}^m (u_i\circ\dot\gamma)g(F_i, \dot\gamma).$$

By integration of this expression from 0 to t, it follows that

$$E(\dot\gamma(t)) - E(\dot\gamma(0)) = \int_0^t g(R\circ\dot\gamma, \dot\gamma) + \int_0^t \sum_{i=1}^m y_i(\dot\gamma(s))u_i(\dot\gamma(s)) \leq \int_0^t \sum_{i=1}^m y_i(\dot\gamma(s))u_i(\dot\gamma(s)),$$

which is the passivity condition for E. ∎

Observe that the last inequality is the balance of energy referred to in the above section. The right-hand side is the work of the external forces, and the left-hand side is the stored energy.

4. Stability of equilibrium points in complete Riemannian manifolds.

4.1. Statement of the problem. Let (M, g) be a Riemannian manifold. It is known that from the metric g we can define in M a distance d in such a way that the induced metric topology from d coincides with the original one in M. Then (M, g) is a complete manifold if the induced distance is complete. Throughout this section we

suppose that (M, g) is a complete Riemannian manifold. For equivalent conditions on the completeness of a Riemannian manifold, see [6].

Let $X \in \mathfrak{X}(M)$ and let $x_0 \in M$ be an equilibrium point of X. Consider the associated dynamical system $X \circ \gamma = \dot{\gamma}$, where $\gamma\colon (a, b) \longrightarrow M$ is a differentiable curve. We denote by $\gamma(t; x)$ the solution with initial condition $\gamma(0; x) = x \in M$. We say that a solution $\gamma(t; x)$ is *bounded* if its image is bounded as a subset of the metric space M.

The goal of this section is to find conditions for the stability of the equilibrium point x_0, in the case where the Riemann manifold (M, g) is complete and the vector field X is also complete. In order to achieve this, we shall generalize the results of La Salle in $\mathbb{R}^n$ to this situation. See [16] for more details concerning the results in $\mathbb{R}^n$.

Recall that an equilibrium point x_0 is *stable* if for every $\epsilon > 0$ there exists a $\delta > 0$ such that if $\mathrm{d}(x, x_0) \leq \delta$, then $\mathrm{d}(\gamma(t; x); x_0) \leq \epsilon$, and *asymptotically stable* if it is stable and there exists a neighborhood O of x_0 such that $\lim_{t\to\infty} \mathrm{d}(\gamma(t; x), x_0) = 0$ for all $x \in O$. If $O = M$, we say that x_0 is *globally asymptotically stable*.

4.2. Liapunov functions. If $x_0 \in M$ is an equilibrium point of $X \in \mathfrak{X}(M)$, we say that $V \in C^\infty(M)$ is a *Liapunov function* of X at x_0 if $V(x_0) = 0$, $V(x) > 0$ for all $x \neq x_0$, and $\mathcal{L}_X V(x) \leq 0$ for all $x \in M$.

PROPOSITION 4.1. *If there exists a Liapunov function $V \in C^\infty(M)$ of X at x_0, then x_0 is stable.*

Proof. Let $\partial B(x_0, r) = \{x \in M | \mathrm{d}(x, x_0) = r\}$. Obviously, it is a closed and bounded set. Since (M, g) is complete, the Hopf–Rinow theorem ensures that it is compact. Then let β be the minimum value of the Liapunov function V on this compact set.

Consider $K = \{x \in B(x_0, r) \mid V(x) \leq \alpha\}$, where $\alpha \in (0, \beta)$. Let us show that K is invariant under the flow of X. Take $x \in K$ and consider $\gamma(t; x)$. Since $V(\gamma(t; x))$ is a decreasing function, it satisfies $V(\gamma(t; x)) \leq V(\gamma(0; x)) \leq \alpha$, and as a consequence $\gamma(t; x) \in B(x_0, r)$ for $t \geq 0$ because $\alpha < \beta$. Then K is invariant, and therefore x_0 is stable. □

Note. From this proposition, we have that if X is a dissipative field and the dissipative function S has a strict absolute minimum at the equilibrium point x_0, then x_0 is stable, since $V = S - S(x_0)$ is a suitable Liapunov function.

4.3. La Salle's theorem. Let $X \in \mathfrak{X}(M)$ be a complete vector field on the complete Riemannian manifold (M, g). For $x \in M$, consider the set

$$W(x) = \left\{ y \in M \mid \exists\, (t_k) \subset \mathbb{R},\, t_k > 0,\, t_k \to \infty,\, \lim_{k\to\infty} \gamma(t_k; x) = y \right\}.$$

This is the set of limit points of the integral curve $\gamma(t; x)$.

LEMMA 4.2. *If the solution $\gamma(t; x)$ is bounded, then $W(x)$ is an invariant and compact subset of M and $\lim_{t\to\infty} \mathrm{D}(\gamma(t; x), W(x)) = 0$ (where* D *is the point-set distance).*

Proof. Let $\hat{x} \in W(x)$; then there exists (t_k) such that $\lim_{k\to\infty} \gamma(t_k; x) = \hat{x}$. In order to show that $W(x)$ is invariant under X, we must prove that $\gamma(t; \hat{x}) \in W(x)$ for all $t \in \mathbb{R}$. For every $t \in \mathbb{R}$, take the sequence $(t + t_k)$ from the existence and uniqueness of solutions of ODE $\gamma(t + t_k; x) = \gamma(t; \gamma(t_k; x))$. The flow box theorem implies that $\lim_{k\to\infty} \gamma(t; \gamma(t_k; x)) = \gamma(t; \hat{x})$; therefore $W(x)$ is invariant.

Since $\gamma(t; x)$ is bounded, $W(x)$ is also bounded. Thus, we need only to prove that it is closed because of the Hopf–Rinow theorem. Let $(x_n) \subset W(x)$ and assume

that $\lim_{n\to\infty} x_n = \hat{x}$. Since $x_n \in W(x)$ for all $n \in \mathbb{N}$, there exist (t_k^n) such that $\lim_{k\to\infty} \gamma(t_k^n; x) = x_n$ for all $n > 0$. Then we can find a sequence $(\hat{t}_m)$ of (t_k^n) such that $\lim_{m\to\infty} \mathrm{d}(\gamma(\hat{t}_m; x), \hat{x}) = 0$. Hence, $\lim_{m\to\infty} \gamma(\hat{t}_m; x) = \hat{x}$, and the result follows.

Suppose that $\lim_{t\to\infty} \mathrm{D}(\gamma(t; x), W(x)) \neq 0$; then there exists $\epsilon > 0$ and a sequence (t_m), with $t_m \longrightarrow \infty$, such that $\mathrm{d}(\gamma(t_m; x), y) \geq \epsilon$ for all $y \in W(x)$. However, the curve $\gamma(t; x)$ is contained in a compact set, since this solution is bounded and (M, g) is a complete manifold. Thus we can find a convergent partial sequence $(\gamma(\hat{t}_k; x))$ such that $\lim_{k\to\infty} \gamma(\hat{t}_k; x) = \hat{y}$ for some $\hat{y} \in W(x)$. However, this is not possible because $\hat{y} \in W(x)$ and $\mathrm{d}(\hat{y}, y) \geq \epsilon$ for all $y \in W(x)$. □

It is essential for the solution to be bounded. It is well known that there are dynamical systems in $\mathbb{R}^3$, a complete Riemannian manifold, which have unbounded solutions with chaotic behavior.

PROPOSITION 4.3 (La Salle's theorem). *Let $X \in \mathfrak{X}(M)$ be a complete vector field on a complete Riemannian manifold (M, g). Let x_0 be an equilibrium point of X and $V \in C^\infty(M)$ a Liapunov function of X at x_0. Suppose that $x \in M$ is such that $\gamma(t; x)$ is bounded. Then $W(x) \subseteq \{y \in M | \mathcal{L}_X V(y) = 0\}$.*

Proof. The function $v(t) = V(\gamma(t; x))$ is positive and decreasing; then $\lim_{t\to\infty} v(t) = \hat{v} \geq 0$. Take $y \in W(x)$; there exists (t_m) such that $\lim_{m\to\infty} \gamma(t_m; x) = y$. Since V is continuous, it follows that $V(y) = \lim_{m\to\infty} V(\gamma(t_m; x)) = \hat{v}$. As $W(x)$ is invariant, we have $V(\gamma(t; y)) = V(y)$. Therefore $\mathcal{L}_X V(y) = 0$. □

PROPOSITION 4.4. *Under the same conditions of Proposition 4.3, if $K \subseteq M$ is the maximal invariant set under the flow of X such that $K \subseteq \{y \in M | \mathcal{L}_X V(y) = 0\}$, we have that $\lim_{t\to\infty} \mathrm{D}(\gamma(t; x), K) = 0$.*

Proof. Suppose that $\lim_{t\to\infty} \mathrm{D}(\gamma(t; x), K) \neq 0$; then there exists $\epsilon > 0$ and a sequence (t_m) such that $\mathrm{d}(\gamma(t_m; x), y) \geq \epsilon$ for all $y \in K$. Since $\gamma(t; x)$ is contained in a compact set, a convergent partial subsequence $\gamma(\hat{t}_k; x)$ can be found such that $\lim_{k\to\infty} \gamma(\hat{t}_k; x) = \hat{y} \in W(x) \subseteq K$. But this is a contradiction, since $\mathrm{d}(\hat{y}, y) \geq \epsilon$ for all $y \in K$. □

COROLLARY 4.5. *Under the same conditions as in Proposition 4.3, if the integral curves of X are bounded and the Liapunov function V of X at x_0 satisfies $\mathcal{L}_X V(x) < 0$ for $x \neq x_0$, then x_0 is globally asymptotically stable.*

Proof. Since V is a Liapunov function, x_0 is a stable equilibrium point. On the other hand, the conditions that we have assumed imply that $K = \{x_0\}$, and from the previous proposition, $\lim_{t\to\infty} \mathrm{D}(\gamma(t; x), x_0) = 0$ for all $x \in M$. □

Note. If X is strictly dissipative, the integral curves are bounded, and the dissipation function has a strict absolute minimum in x_0, then x_0 is globally asymptotically stable (see the note following Proposition 4.1).

In order to assure that the integral curves are bounded, if M is not compact, it is enough for the Liapunov function to satisfy that $\lim_{d(q,x_0)\to\infty} V(q) = \infty$.

5. Applications to control systems. Consider now the dynamical control system on the complete Riemannian manifold (M, g) given by $Y = X + \sum_{i=1}^m X_i u_i$ with outputs $y_i = h_i \in C^\infty(M)$, $i = 1, \ldots, m$. Let x_0 be an equilibrium point of the drift field X.

PROPOSITION 5.1. *Suppose that the system is passive with storage function S with a strict absolute minimum at x_0. If there exists a function $\phi = (\phi_1, \ldots, \phi_m) \in C^\infty(\mathbb{R}^m, \mathbb{R}^m)$ such that $\sum_{i=1}^m y_i \phi_i(y_1, \ldots, y_m) \geq 0$ and $\phi_i(y_1(x_0), \ldots, y_m(x_0)) = 0$, $i = 1, \ldots, m$, then we can stabilize the equilibrium point x_0 by the feedback $u_i(v_q) = -\phi_i(y_1(v_q), \ldots, y_m(v_q))$.*

Proof. It is enough to prove that the storage function S is a Liapunov function of $Z = X - \sum_{i=1}^{m} X_i\phi_i(y_1, \dots, y_m)$. Without lost of generality, we can suppose that $S(x_0) = 0$. Since the control system is passive and satisfies $\sum_{i=1}^{m} y_i\phi_i(y_1, \dots, y_m) \geq 0$, the next inequalities hold:

$$S(\gamma(T)) - S(\gamma(0)) \leq -\int_0^T \sum_{i=1}^{m} y_i(\gamma(s))\phi_i(y_1(\gamma(s)), \dots, y_m(\gamma(s)))\, ds \leq 0,$$

where γ is an integral curve of the field Z. Hence S is decreasing along the integral curves of Z. Moreover, $S(x) > 0$ for $x \neq x_0$ and $S(x_0) = 0$. □

Remember that a dynamical control system is called *zero-stable detectable* at x_0 if for every $x \in M$ such that $h(\gamma(t; x)) = 0$ we have $\lim_{t\to\infty} \gamma(t; x) = x_0$.

PROPOSITION 5.2. *Suppose now that the system is zero-stable detectable, is passive where the storage function S has a strict absolute minimum at x_0, and the sets $S^{-1}([0, a])$ are bounded. If there exists a function $\phi = (\phi_1, \dots, \phi_m)$ such that $\sum_{i=1}^{m} y_i\phi_i(y_1, \dots, y_m) > 0$, $(y_1, \dots, y_m) \neq 0$, and $\phi_i(y_1(x_0), \dots, y_m(x_0)) = 0$, $i = 1, \dots, m$, then the feedback $u_i = -\phi_i(y_1(v_q), \dots, y_m(v_q))$ makes x_0 globally asymptotically stable.*

Note. The zero-stable detectable property assures that the outputs faithfully transmit the behavior of the system. See [10] and [11] for details on this notion and examples.

Proof. From the above proposition, we have that x_0 is a stable equilibrium point. Notice that $\gamma(t)$ is defined for all $t \geq 0$ because the energy is decreasing along the integral curves of $Z = X - \sum_{i=1}^{m} X_i\phi_i(y_1, \dots, y_m)$ and $\gamma(t) \subseteq S^{-1}([0, \alpha])$, where $S^{-1}([0, \alpha])$ are bounded and closed sets.

Consider an integral curve $\gamma(t; x)$ of Z. It follows that $\lim_{t\to\infty} S(\gamma(t; x)) = a_0 \geq 0$. Then $S(\hat{x}) = a_0$ for all $\hat{x} \in W(x)$ because of the continuity of the function S. Let $\hat{x} \in W(x)$; since $W(x)$ is an invariant set and the control system is passive, the next inequalities hold:

$$0 = S(\gamma(T; \hat{x})) - S(\gamma(0; \hat{x})) \leq -\int_0^T \sum_{i=1}^{m} y_i\phi_i(y_1, \dots, y_m)\, dt < 0.$$

Hence $\sum_{i=1}^{m} y_i(\gamma(t, \hat{x}))\phi_i(y_1(\gamma(t, \hat{x})), \dots, \gamma(t, \hat{x})) = 0$, and in consequence $y_i(\gamma(t, \hat{x})) = 0$, $t \geq 0$. However, as the system is zero-stable detectable, $\lim_{t\to\infty} \gamma(t; \hat{x}) = x_0$ and $S(\lim_{t\to\infty} \gamma(t, \hat{x})) = a_0 = S(x_0) = 0$. Then $\lim_{t\to\infty} S(\gamma(t; x)) = 0$ and $\lim_{t\to\infty} \gamma(t; x) = x_0$. □

Example. Consider the dynamical control system $\dot{x}_1 = x_2$, $\dot{x}_2 = -x_1$, $\dot{x}_3 = u$, $y = x_3$. It is passive with storage function $S = \frac{1}{2}x_3^2$, and the equilibrium point $(0, 0, 0)$ is stable, but not asymptotically stable, since it is not zero-stable detectable.

6. Stability of equilibrium points of mechanical dissipative systems. This section is devoted to the study of the stability of mechanical systems. Since vector fields on TQ correspond to mechanical systems, according to the foregoing sections, the tangent bundle of a complete Riemannian manifold Q must be provided with a complete Riemannian metric in order to establish the results in an overall context and to determine the conditions for one equilibrium point to be globally asymptotically stable. This is always possible if we take as a metric on TQ the Sasaki metric g^T. It is known that the topology induced by g^T coincides with the natural one in TQ. Moreover, g^T is Riemannian, and if g is complete, then g^T is also complete.

Therefore, $(\mathrm{T}Q, g^T)$ is a complete Riemannian manifold, and we can apply the results of the previous section. See [2] and [15] for more details. Obviously, $\mathrm{T}Q$ can be provided with other different metrics. The only problem is to select one which makes $\mathrm{T}Q$ a complete Riemannian manifold.

6.1. Dissipative systems. Consider a mechanical system $\Sigma = (Q, g, U, R)$, where (Q, g) is a complete Riemannian manifold and $R \in \mathfrak{X}(Q, \tau_Q)$ is dissipative. The equilibrium points of the system are of the form $0_{q^o} \in \mathrm{T}_{q^o}Q$, since the vector field associated to the system, $Y = X_g - (\operatorname{grad} U)^v + R^v$, satisfies the second order condition. We denote by 0_{q^o} the zero vector in $\mathrm{T}_{q^o}Q$. Moreover, if $R \in \mathfrak{X}(Q, \tau_Q)$ verifies that $R(0_q) = 0$ for every $q \in Q$, for example if R comes from a Rayleigh function, then q^o is a zero of the gradient of U.

It is known that if $U \in C^\infty(Q)$ is bounded from below and $R \in \mathfrak{X}(Q, \tau_Q)$ is dissipative, then $Y \in \mathfrak{X}(\mathrm{T}Q)$ is positive complete; see [1] for more details. Therefore, we can study the asymptotic behavior of any solution.

PROPOSITION 6.1. *Let (Q, g) be a complete Riemannian manifold, and consider a mechanical system $\Sigma = (Q, g, U, R)$ with $R \in \mathfrak{X}(Q, \tau_Q)$ dissipative. Suppose that q^o is a strict absolute minimum of U and $R(0_q) = 0$ for every $q \in Q$. Then we have that 0_{q^o} is stable.*

Moreover, if 0_{q^o} is the only equilibrium point, $R \in \mathfrak{X}(Q, \tau_Q)$ is strictly dissipative, and the solutions are bounded, then 0_{q^o} is globally asymptotically stable.

Proof. In order to prove the stability, it is enough to show that $\tilde{E} = T + \tau_Q^* U - U(q^o)$ is a Liapunov function; see Proposition 4.1. That is, $\tilde{E}(v_q) > 0$ for all $v_q \neq 0_{q^o}$ and $\mathcal{L}_Y \tilde{E}(v_q) \leq 0$. To prove the last assertion, let $\dot{\gamma}(t; v_q)$ be a solution of the system. Then, since R is dissipative, we have $\mathcal{L}_Y \tilde{E}(v_q) = g(R \circ \dot{\gamma}(t; v_q), \dot{\gamma}(t; v_q))|_{t=0} \leq 0$, as in Proposition 3.1.

Asymptotic stability follows from Proposition 4.3. Consider a solution of the system $\dot{\gamma}(t; v_q)$; then $\mathcal{L}_Y \tilde{E}(v_q) = g(R \circ \dot{\gamma}(t; v_q), \dot{\gamma}(t; v_q))|_{t=0} \leq -\alpha g(\dot{\gamma}(t; v_q), \dot{\gamma}(t; v_q))|_{t=0} = -\alpha g_q(v_q, v_q)$ for some $\alpha > 0$. So the maximal invariant set K, defined in Proposition 4.4, verifies $K \subseteq \{y \in \mathrm{T}Q | \mathcal{L}_X \tilde{E}(y) = 0\} \subseteq \{0_q | q \in Q\}$, and since there is only one equilibrium point, then $K = \{0_{q^o}\}$. Therefore, 0_{q^o} is globally asymptotically stable. □

6.2. Partially dissipative systems. Let (Q_1, g_1), (Q_2, g_2) be Riemannian manifolds with Levi–Civita connections ∇_1 and ∇_2. We can provide the manifold $Q = Q_1 \times Q_2$ with a natural Riemannian metric $g = g_1 \oplus g_2$, which is complete if g_1 and g_2 are also complete. For (Q, g) the Levi–Civita connection is $\nabla = \nabla^1 \oplus \nabla^2$; that is, $\nabla_{\dot{\gamma}} \dot{\gamma} = \nabla^1_{\dot{\gamma}_1} \dot{\gamma}_1 + \nabla^2_{\dot{\gamma}_2} \dot{\gamma}_2$, where $\gamma = (\gamma_1, \gamma_2)$ is a curve in Q.

From the above, the dynamical equations of the system $\Sigma = (Q, g, U, R)$ can be written as

$$\begin{aligned} \nabla^1_{\dot{\gamma}_1} \dot{\gamma}_1 &= -(\operatorname{grad}_1 U) \circ (\gamma_1, \gamma_2) + R_1 \circ (\dot{\gamma}_1, \dot{\gamma}_2), \\ \nabla^2_{\dot{\gamma}_2} \dot{\gamma}_2 &= -(\operatorname{grad}_2 U) \circ (\gamma_1, \gamma_2) + R_2 \circ (\dot{\gamma}_1, \dot{\gamma}_2), \end{aligned} \tag{5}$$

where $\mathrm{i}(\operatorname{grad}_i U) g_i = -\mathrm{d}_i U$ and d_i is the exterior differential operator in Q_i, $i = 1, 2$. Recall that if $X \in \mathfrak{X}(Q)$, then $X = X_1 + X_2$, where $X_i \in (Q, \pi_i)$ and $\pi_i : Q \to Q_i$ is the natural projection.

PROPOSITION 6.2. *Given a mechanical system $\Sigma = (Q = Q_1 \times Q_2, g = g_1 \oplus g_2, U, R)$ with $R = (0, R_2)$, $R_2 \in \mathfrak{X}(Q_2, \tau_{Q_2})$ a dissipative force, if $U \in C^\infty(Q)$ has a strict absolute minimum at $q^o \in Q$, then the equilibrium point 0_{q^o} is stable.*

Proof. The function $\tilde{E} = T + \tau_Q^* U - U(q^o)$ is a Liapunov function at 0_{q^o} because $\tilde{E} > 0$ for $v_q \neq 0_{q^o}$ and $\mathcal{L}_Y \tilde{E}(v_q) = g(R \circ \dot{\gamma}, \dot{\gamma}) = g_2(R_2 \circ \dot{\gamma}_2, \dot{\gamma}_2) \leq 0$. Therefore, 0_{q^o} is stable. □

The vector field $R \in \mathfrak{X}(Q, \tau_Q)$ is called *strictly partially dissipative* if it verifies $g(R \circ v_q, v_q) \leq -\alpha g_2(v_{q_2}, v_{q_2})$, with α a real number, $\alpha > 0$, for all $v_q \in \mathrm{T}_q Q$. This condition means that R is only strictly dissipative on (Q_2, g_2).

PROPOSITION 6.3. *Let $\Sigma = (Q = Q_1 \times Q_2, g = g_1 \oplus g_2, U, R)$ be a mechanical system with $R \in \mathfrak{X}(Q, \tau_Q)$ strictly partially dissipative and such that if $v_{q_2} = 0$, then $R(v_q) = 0$.*

Suppose that $U \in C^\infty(Q)$ has a strict absolute minimum at q^o and every solution $\gamma(t) = (\gamma_1, \gamma_2)$ such that $\dot{\gamma}_2 = 0$ and $(\mathrm{grad}_2 U) \circ \gamma = 0$ satisfies $\lim_{t\to\infty} \dot{\gamma}(t) = 0_{q^o}$. Then 0_{q^o} is globally asymptotically stable.

Proof. Notice that the solutions are defined for all $t \geq 0$, because the vector field associated to the system is dissipative, and then it is positive complete.

Consider a solution of the system Σ, $\dot{\gamma}(t; v_q)$, and take the function $\tilde{E} = T + U - U(q^o)$. Since $\tilde{E} \geq 0$ and it verifies

$$\frac{d(\tilde{E} \circ \dot{\gamma})}{dt} = g(R \circ \dot{\gamma}, \dot{\gamma}) \leq -\alpha g(\dot{\gamma}_2, \dot{\gamma}_2) \leq 0, \tag{6}$$

then $\lim_{t\to\infty} \tilde{E}(\dot{\gamma}(t; v_q)) = a_0 \geq 0$. By the continuity of $\tilde{E}$, $\tilde{E}(\hat{x}) = a_0$ for all $\hat{x} \in W(v_q)$, with the same notation as in section 4.3.

Consider now the solution $\dot{\gamma}(t; \hat{x})$, where $\hat{x} \in W(v_q)$. From inequality (6) and the invariance of $W(v_q)$, it is clear that $g_2(\dot{\gamma}_2(t; \hat{x}), \dot{\gamma}_2(t; \hat{x})) = 0$. Hence $\dot{\gamma}_2(t; \hat{x}) = 0, t \geq 0$.

Substitution of $\dot{\gamma}_2(t; \hat{x}) = 0$ into (5) gives that $(\mathrm{grad}_2 U)(\gamma(t; \hat{x})) = 0$ and then $\lim_{t\to\infty}(\dot{\gamma}(t; \hat{x})) = 0_{q^o}$. Hence $a_0 = E(\dot{\gamma}(t; \hat{x})) = 0$ and therefore $\lim_{t\to\infty} \dot{\gamma}(t; v_q) = 0_{q^o}$. □

7. Dissipative control of mechanical systems.

7.1. Stabilization on equilibrium points. Let (Q, g) be a complete Riemannian manifold and 0_{q^o} an equilibrium point of the simple mechanical system (Q, g, U). Consider the mechanical control system $\Sigma = (Q, g, U, F_1, \ldots, F_m)$ with the natural outputs. Then we have the following proposition.

PROPOSITION 7.1. *If the potential function U has a strict absolute minimum at q^o, then 0_{q^o} is a stable equilibrium point of the system $\widetilde{\Sigma} = (Q, g, U, R)$, where $R(v_q) = -\beta \sum_{i=1}^m F_i(v_q) g_q(F_i(v_q), v_q)$, $\beta > 0$.*

Note. The system $\widetilde{\Sigma}$ is obtained from Σ by the feedback $u_i(v_q) = -\beta h_i(v_q) = -\beta g_q(F_i(v_q), v_q)$, with β a real number, $\beta > 0$.

Proof. The function $\tilde{E} = T + U - U(q^o)$ is a Liapunov function for $\widetilde{\Sigma} = (Q, g, U, R)$, because $\tilde{E}(v_q) > 0$ for every $v_q \neq 0_{q^o}$ and $\mathcal{L}_Z \tilde{E}(v_q) = -\beta \sum (g(F_i(v_q), v_q))^2$, where $Z \in \mathfrak{X}(\mathrm{T}Q)$ is the dynamical vector field associated to the system $\widetilde{\Sigma} = (Q, g, U, R)$. □

COROLLARY 7.2. *Under the same condition of the above proposition, suppose there exist functions $\phi_1, \ldots, \phi_m \in C^\infty(\mathbb{R}^m, \mathbb{R})$ such that $\phi_i(0) = 0$, $i = 1, \ldots, m$, and $\sum_{i=1}^m y_i \phi_i(y_1, \ldots, y_m) \geq 0$. Then 0_{q^o} is a stable equilibrium point of $\widetilde{\Sigma} = (Q, g, U, R = -\sum_{i=1}^m F_i \phi_i(y_1, \ldots, y_m))$.*

7.2. Stabilization at an arbitrary point.

7.2.1. Dynamic extensions of a control system. Let Σ_p be a control system on the manifold M_p; that is, $\Sigma_p = (M_p, Y_p = X_p^0 + \sum_{i=1}^m F_i^p u_i^p)$, with controls $u_i^p \in C^\infty(M_p)$. Denote by Σ_c the couple made by a manifold M_c and a vector field $Y_c \in \mathfrak{X}(M_c, \pi_c)$, where $\pi_c \colon M_p \times M_c \longrightarrow M_c$. Then $\Sigma_c = (M_c, Y_c)$ is a system on M_c which depends on the states of M_p.

With Σ_p and Σ_c, we can define a new control system $\Sigma = (M, Y)$, which is called a *dynamic extension* of Σ_p, in the following way. As phase space consider the manifold $M = M_p \times M_c$ and the vector field on M given by $Y = (X_p^0, Y_c) + \sum_{i=1}^m (F_i^p, 0) u_i$, where the new controls $u_i \in C^\infty(M)$. Note that the controls u_i depend on the state variables of M_c and M_p. Moreover, if Σ_c has good properties, for instance if it is a dissipative system, then these properties can be used to stabilize Σ_p.

The notation comes from that usually employed in control theory, M_p for the "plant system," the system under study, and M_c for the "controller" (see [10]), and we say that $\Sigma = (M, Y)$ is a *dynamic extension* of Σ_p.

7.2.2. Mechanical systems. Consider the mechanical system with controls $\Sigma_p = (Q_p, g_p, U_p, F_1^p, \dots, F_{m_p}^p)$. We are interested in dynamic extensions of Σ_p.

Suppose that we have a Riemannian manifold (Q_c, g_c), a function $U_c \colon Q_p \times Q_c \longrightarrow \mathbb{R}$, and a vector field $R_c \in \mathfrak{X}(Q_c, \tau_c)$, where $\tau_c \colon \mathrm{T}Q_c \longrightarrow Q_c$ is the natural projection. Denote by Σ_c the set (Q_c, g_c, U_c, R_c). With Σ_p and Σ_c we can define Σ, a dynamic extension of Σ_p, given by $Q = Q_p \times Q_c$, $g = g_p \oplus g_c$, and $Y = (X_g^p, X_g^c) - (\mathrm{grad}_p U_p, \mathrm{grad}_c U_c)^v + R^v + \sum_{i=1}^{m_p} F_i^v u_i$, where $R = (0, R_c)$ and $F_i = (F_i^p, 0)$, with controls $u_i \in C^\infty(Q)$.

Once again the new controls depend on the states of Q_p and Q_c, not only on Q_p. Notice that Σ is not a mechanical system with controls, because we have neither a potential function $U \in C^\infty(Q)$ nor $\mathrm{grad}\, U$ as a force field.

7.2.3. Stabilization by dynamic extensions. Let the mechanical control system be $\Sigma_p = (Q_p, g_p, U_p, F_1^p, \dots, F_{m_p}^p)$. We will try to design controls to stabilize the system at a desired point q_p^o. The idea is to construct a suitable dynamic extension Σ such that the whole system given by the dynamic extension makes q_p^o stable by an appropriate feedback.

With this aim, consider that Σ is a dynamic extension of Σ_p given by $\Sigma_c = (Q_c, g_c, U_c, R_c)$, and suppose that $\mathrm{grad}_p U_c \in \langle F_1^p, \dots, F_{m_p}^p \rangle$. We have the following results.

PROPOSITION 7.3. *Suppose that $q^0 = (q_p^0, q_c^0)$ is a strict absolute minimum of $U = U_p + U_c$ and that $R_c \in \mathfrak{X}(Q_c, \tau_c)$ is a dissipative vector field such that $R_c(0_{q_c}) = 0$ for every $q_c \in Q_c$. Then the feedback $u_i \in C^\infty(Q)$, such that $\sum_{i=1}^{m_p} F_i u_i = \mathrm{grad}_{\mathrm{p}} U_c$, makes 0_{q^o} stable.*

Note. Observe that the system which results from Σ by the feedback previously described is the mechanical system $\overline{\Sigma} = (Q, g, U, R)$.

Proof. Note that if $R_c \in \mathfrak{X}(Q_c, \tau_{Q_c})$ is dissipative, that is, if it verifies that $g_c(R_c(v_c), v_c) \leq 0$, then $R = (0, R_c)$ is also dissipative with respect to the metric g, since $g(R(v_p, v_c), (v_p, v_c)) = g_p(0, v_p) + g_c(R_c(v_c), v_c) = g_c(R_c(v_c), v_c) \leq 0$.

Then the function $\tilde{E} = T + \tau_Q^* U - U(q^o)$ is a Liapunov function of $\overline{\Sigma}$ because $\tilde{E}(v_p, v_c) > 0$ for every $(v_p, v_c) \neq 0_{q^o}$, and $\mathcal{L}_Z \tilde{E}((v_p, v_c)) = g(R((v_p, v_c)), (v_p, v_c)) \leq 0$ with Z the dynamical vector field of the whole system. □

PROPOSITION 7.4. *Under the same conditions as in the above proposition, let us suppose that $R_c \in \mathfrak{X}(Q_c, \tau_c)$ is strictly dissipative and every solution $\gamma = (\gamma_p, \gamma_c)$*

of the system $\overline{\Sigma} = (Q, g, U, R)$ *such that* $\dot{\gamma}_c(t) = 0$ *and* $grad_c\, U_c(\gamma_p, \gamma_c) = 0$ *satisfies* $\lim_{t\to\infty} \dot{\gamma}(t) = 0_{p^o}$. *Then* 0_{p^o} *is a globally asymptotically stable point of the system* $\overline{\Sigma}$.

Comment. The last condition is related to the property of being zero-stable detectable. See section 5 or [10].

Proof. Since $R \in \mathfrak{X}(Q, \tau_Q)$ is partially dissipative, we have the conditions of Proposition 6.3. □

Note. Given the mechanical control system Σ_p, suppose that $q_p^0 \in Q_p$ is the point where we want to stabilize the system. According to the above propositions, it is enough to find a system Σ_c with the condition that the new potential function $U = U_p + U_c$ has a strict absolute minimum at (q_p^0, q_c^0), where q_c^0 can be any point of the manifold Q_c.

Observe that the process has two different parts:

1. to find the potential function U_c, with the purpose of obtaining a new one $U = U_p + U_c$ that has a strict absolute minimum at (q_p^0, q_c^0);
2. to design controls u_i with the aim of obtaining a dynamical extension which can stabilize the system in q_p^0 using the dissipative force associated to Σ_c.

See [10] for a local point of view, examples, and applications of all these results. Compare the above proposition with Theorem 3.1 in [10].

Acknowledgments. We wish to thank the anonymous referees for their careful reading of the manuscript. Their helpful comments are greatly appreciated. Thanks also to Mr. Jeff Palmer for his assistance in preparing the English version.

REFERENCES

[1] R. Abraham and J. E. Marsden, *Foundations of Mechanics*, 2nd ed., Addison–Wesley, Reading, MA, 1978.
[2] R. Abraham, J. E. Marsden, and T. Ratiu, *Manifolds, Tensor Analysis, and Applications*, Springer-Verlag, New York, 1988.
[3] V. Arnold, *Mathematical Methods of Classical Mechanics*, Springer-Verlag, New York, 1989.
[4] C. Byrnes, A. Isidori, and J. Willems, *Passivity, feedback equivalence, and the global stabilization of minimum phase nonlinear systems*, IEEE Trans. Automat. Control, 36 (1976), pp. 1228–1240.
[5] R. Brockett, R. Millman, and H. Sussmann, *Differential Geometric Control Theory*, Birkhäuser Boston, Cambridge, MA, 1983.
[6] M. P. do Carmo, *Riemannian Geometry*, Birkhäuser Boston, Cambridge, MA, 1992.
[7] A. Isidori, *Nonlinear Control Systems*, Springer-Verlag, Berlin, 1989.
[8] A. D. Lewis and R. M. Murray, *Configuration controllability of simple mechanical control systems*, SIAM Rev., 41 (1999), pp. 555–574.
[9] H. Nijmeijer and A. van der Schaft, *Nonlinear Dynamical Control Systems*, Springer-Verlag, New York, 1990.
[10] R. Ortega, A. Loria, R. Kelly, and L. Praly, *On passivity based output feedback and global stabilization of Euler-Lagrange systems*, Internat. J. Robust Nonlinear Control, 5 (1995), pp. 313–323.
[11] R. Ortega, A. Loria, P. J. Nicklasson, and H. Sira-Ramírez, *Passivity-Base Control of Euler–Lagrange Systems*, Springer-Verlag, London, 1998.
[12] G. K. Pozharitskii, *On asymptotic stability of equilibria and stationary motions of mechanical systems with partial dissipation*, J. Appl. Math. Mech., 25 (1962), pp. 979–993.
[13] R. Sepulchre, M. Janhović, and P. Kokotović, *Constructive Nonlinear Control*, Springer-Verlag, London, 1997.
[14] M. Takegaki and S. Arimoto, *A new feedback method for dynamic control of manipulators*, Trans. ASME J. Dyn. Syst. Meas. Control, 103 (1981), pp. 119–125.
[15] K. Yano and S. Ishihara, *Tangent and Cotangent Bundles*, Marcel Dekker, New York, 1973.
[16] J. Zabczyk, *Mathematical Control Theory*, Birkhäuser Boston, Cambridge, MA, 1992.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 5, pp. 1517–1539

STATE-CONSTRAINED OPTIMAL CONTROL GOVERNED BY NON–WELL-POSED PARABOLIC DIFFERENTIAL EQUATIONS*

GENGSHENG WANG† AND LIJUAN WANG†

Abstract. This paper is concerned with a maximum principle of optimal control problems governed by some parabolic differential equations which could be non–well-posed. Both an integral-type state constraint and a two-point boundary (time variable) state constraint are considered.

Key words. optimal control, non–well-posed parabolic equation, state constraint

AMS subject classifications. 49K20, 35J65

PII. S0363012900377006

1. Introduction.

In this paper we shall study optimal control problems governed by some semilinear parabolic differential equations which, in particular, could be singular, i.e., have local solution only. We shall call such systems non–well-posed systems and call the optimal control problems governed by such systems non–well-posed optimal control problems.

Throughout this paper, we denote by $\Omega \subset R^n$, $n \geq 3$, a bounded open subset with smooth boundary $\partial\Omega$ (say, for instance, the class of C^2). Let $Q = \Omega \times (0,T)$ with $T > 0$ and $\sum = \partial\Omega \times (0,T)$. Let $a_{ij}(x) \in C^2(\bar{\Omega})$ with $a_{ij}(x) = a_{ji}(x)$ for all $x \in \bar{\Omega}$, satisfying $\sum_{i,j=1}^n a_{ij}(x)\xi_i\xi_j \geq \Lambda \sum_{i=1}^n \xi_i^2$ for all $\xi_i \in R, i = 1, \ldots, n$, and $x \in \bar{\Omega}$, where $\Lambda > 0$. Set $Ay(x,t) = -\sum_{i,j=1}^n \frac{\partial}{\partial x_i}(a_{ij}(x)\frac{\partial y(x,t)}{\partial x_j})$. We set

$$Y = \left\{ y \in L^2(0,T;H_0^1(\Omega)) \mid \frac{\partial y}{\partial t} + Ay \in L^2(Q) \right\}$$

and

$$\|y\|_Y = \left(\|y\|^2_{L^2(0,T;H_0^1(\Omega))} + \left\| \frac{\partial y}{\partial t} + Ay \right\|^2_{L^2(Q)} \right)^{\frac{1}{2}}.$$

Then, as we know (cf. [15]), Y endowed with the norm $\|\cdot\|_Y$ is a Hilbert space, and $Y = H^{2,1}(Q) \cap L^2(0,T;H_0^1(\Omega))$, where

$$H^{2,1}(Q) = \left\{ y \mid y, \frac{\partial y}{\partial t}, \frac{\partial y}{\partial x_i}, \frac{\partial^2 y}{\partial x_i \partial x_j} \in L^2(Q),\ i,j = 1, \ldots, n \right\}.$$

We set

$$H^{2,1}_{\frac{2n}{n+2}}(Q) = \left\{ y \in L^{\frac{2n}{n+2}}(Q) \mid \frac{\partial y}{\partial t}, \frac{\partial y}{\partial x_i}, \frac{\partial^2 y}{\partial x_i \partial x_j} \in L^{\frac{2n}{n+2}}(Q),\ i,j = 1, \ldots, n \right\}$$

and

$$W_0^{1,\frac{2n}{n+2}}(\Omega) = \left\{ w \in L^{\frac{2n}{n+2}}(\Omega) \mid \frac{\partial w}{\partial x_i} \in L^{\frac{2n}{n+2}}(\Omega),\ i = 1, \ldots, n,\ w = 0,\ \text{on } \partial\Omega \right\}.$$

*Received by the editors August 18, 2000; accepted for publication (in revised form) April 1, 2001; published electronically January 18, 2002. This work was supported by the National Natural Science Foundation of China, grant 10071028.
http://www.siam.org/journals/sicon/40-5/37700.html
†Department of Mathematics, Huazhong Normal University, Wuhan, Hubei, 430079, P.R. of China (wanggs@ccnu.edu.cn, kyc@ccnu.edu.cn).

The first problem (P_1) we study in this paper is as follows.

Inf $L(y,u) \equiv$ Inf $\int_0^T [g(t,y)+h(u)]dt$ over all $(y,u) \in Y \times L^2(Q)$ such that

$$\begin{cases} \frac{\partial y(x,t)}{\partial t} + Ay(x,t) + f(x,t,y(x,t)) = u(x,t), & \text{in } Q, \\ y(x,t) = 0, & \text{on } \sum, \\ y(x,0) = y_0(x), & \text{in } \Omega \end{cases} \tag{1.1}$$

and

$$F(y) \in W. \tag{1.2}$$

Here $y_0(x) \in L^2(\Omega)$, and we assume the following:

$\boldsymbol{(H_1)}$ $g : [0,T] \times L^2(\Omega) \to R^+$ is measurable in t, and for every $\delta > 0$ there exists $L_\delta > 0$ independent of t such that $g(t,0) \in L^\infty(0,T)$ and

$$|g(t,y_1) - g(t,y_2)| \le L_\delta \|y_1 - y_2\|_{L^2(\Omega)} \text{ for all } t \in [0,T], \quad \|y_1\|_{L^2(\Omega)} + \|y_2\|_{L^2(\Omega)} \le \delta.$$

$h : L^2(\Omega) \to \bar{R} = (-\infty, +\infty]$ is lower semicontinuous and convex with the following growth property:

$$h(u) \ge c_1 \|u\|^2_{L^2(\Omega)} + c_2,$$

where $c_1 > 0$ and $c_2 \in R$.

$\boldsymbol{(H_2)}$ $f : \bar{\Omega} \times [0,T] \times R \to R$ is continuous, and $f'_y(x,t,\cdot)$ is continuous. Moreover,

$$|f(x,t,y)| \le a_1(x,t) + b_1|y|^{r_1}$$

and

$$|f'_y(x,t,y)| \le \tilde{a}_1(x,t) + \tilde{b}_1|y|^{r_1 - 1},$$

where $a_1(x,t) \in L^2(Q), \tilde{a}_1(x,t) \in L^n(Q)$ with $a_1(x,t) \ge 0$ a.e. in Q, $\tilde{a}_1(x,t) \ge 0$ a.e. in Q; $b_1, \tilde{b}_1 \ge 0$ are two constants; and $r_1 \in R$ with $1 \le r_1 \le \frac{n}{n-2}$.

$\boldsymbol{(H_3)}$ X is a Banach space with X^* strictly convex, and $F : L^2(Q) \to X$ is in the class of C^1. $W \subset X$ is a convex and closed subset.

Let (y_1^*, u_1^*) be an optimal pair for problem (P_1), i.e., $(y_1^*, u_1^*) \in Y \times L^2(Q)$ and satisfies (1.1) and (1.2); moreover, $L(y_1^*, u_1^*) \le L(y,u)$ for all $(y,u) \in Y \times L^2(Q)$ satisfying (1.1) and (1.2). In addition to (H_1), (H_2), and (H_3), we assume the following:

$\boldsymbol{(H_4)}$ $F'(y_1^*)D_r - W$ has finite codimensionality in X for some $r > 0$, where $D_r = \{\, z \in Y \,:\, \|z\|_Y \le r \text{ and } z(x,0) = 0 \,\}$.

For the definition of finite codimensionality of a set and related results, we refer the reader to [13]. Note that $z(x,0)$ makes sense and belongs to $L^2(Q)$ because $z \in Y$ (cf. [15]). It is clear that for each $r > 0$, $D_r \ne \emptyset$ because $0 \in D_r$.

The second problem (P_2) we shall study in this paper is as follows.

Inf $L(y,u) \equiv$ Inf $\int_0^T [g(t,y)+h(u)]dt$ over all $(y,u) \in Y \times L^2(Q)$ such that

$$\begin{cases} \frac{\partial y(x,t)}{\partial t} + Ay(x,t) + f(x,t,y(x,t)) = u(x,t), & \text{in } Q, \\ y(x,t) = 0, & \text{on } \sum \end{cases} \tag{1.3}$$

and

$$(y(x,0), y(x,T)) \in S, \tag{1.4}$$

where the functionals g and h satisfy (H_1) and the function f satisfies (H_2). Moreover we assume the following:

(H_5) $S \subset L^2(\Omega) \times L^2(\Omega)$ is a closed convex subset.

Let (y_2^*, u_2^*) be an optimal pair for problem (P_2), i.e., $(y_2^*, u_2^*) \in Y \times L^2(Q)$ and satisfies (1.3) and (1.4); moreover, $L(y_2^*, u_2^*) \le L(y, u)$ for all $(y, u) \in Y \times L^2(Q)$ satisfying (1.3) and (1.4). We define, for $r_1, r_2, r_3 > 0$,

$$\begin{aligned} B_{r_1,r_2,r_3} = \{(z_0, z_1) \in L^2(\Omega) \times L^2(\Omega) \mid &\text{for all } y \in Y \text{ with } \|y - y_2^*\|_Y \le r_1, \\ &\exists\, (z, v) \in Y \times L^2(Q) \text{ with } \|z\|_Y \le r_2, \|v\|_{L^2(Q)} \le r_3 \\ &\text{such that } z(x,0) = z_0, z(x,T) = z_1 \text{ and} \\ &\tfrac{\partial z}{\partial t} + Az + f_y'(x,t,y)z = v \text{ in } Q \,\}. \end{aligned} \tag{1.5}$$

Observe that for all $r_1, r_2, r_3 > 0$, $B_{r_1,r_2,r_3} \neq \emptyset$, because $(0,0) \in B_{r_1,r_2,r_3}$. In addition to (H_1), (H_2), and (H_5), we assume this:

(H_6) There exist $r_1, r_2, r_3 > 0$ such that the set $(B_{r_1,r_2,r_3} - S)$ has finite codimensionality in $L^2(\Omega) \times L^2(\Omega)$.

By (H_1), (H_2), (H_3), and (H_4), we may get the necessary conditions for (y_1^*, u_1^*) to be optimal for problem (P_1), which is given by Theorem 1.1 below, and by (H_1), (H_2), (H_5), and (H_6) we may obtain the necessary conditions for (y_2^*, u_2^*) to be optimal for problem (P_2), which is presented by Theorem 1.2.

In order to get the existence of optimal pairs for problems (P_1) and (P_2), we need the additional assumptions that follow.

(H_7) There exist $\tilde{c}_1 > 0$ and $\tilde{c}_2 \ge 0$ such that

$$g(t,y) \ge \tilde{c}_1 |y|^{2\tilde{r}} + \tilde{c}_2 \quad \text{for all } (t,y) \in [0,T] \times R,$$

where $\tilde{r} = \frac{n}{n-2}$.

(H_8) $D_{ad} \neq \emptyset$, where

$$D_{ad} = \{(y,u) \in Y \times L^2(Q) : (y,u) \text{ satisfies (1.1) and (1.2)}\}.$$

(H_9) $\tilde{D}_{ad} \neq \emptyset$, where

$$\tilde{D}_{ad} = \{(y,u) \in Y \times L^2(Q) : (y,u) \text{ satisfies (1.3) and (1.4)}\}.$$

(H_{10}) The set $\{x_0 \in L^2(\Omega) : (x_0, x_1) \in S \text{ for some } x_1 \in L^2(\Omega)\}$ is bounded in $L^2(\Omega)$.

The following notation will be in effect throughout this paper. $[F'(y)]^*$ denotes the adjoint operator of operator $F'(y)$, where $y \in L^2(Q)$; $\partial g(t,y)$ and $\partial h(u)$ denote the subdifferential of g to the second variable at y in the sense of Clarke (cf. [9]) and the subdifferential of convex functional h at u (cf. [2]), respectively, where $y \in L^2(Q)$ and $u \in L^2(Q)$; $d_S(\cdot,\cdot)$ denotes the distance of $(\cdot,\cdot)$ to S in $L^2(Q) \times L^2(Q)$; and $\langle \cdot, \cdot \rangle$ and $\langle \cdot, \cdot \rangle_{X^*,X}$ denote the inner product in $L^2(\Omega)$ and the pairing between X^* and X, respectively.

From now on, we shall omit x, t in all functions of x, t if there is no ambiguity. The main results we obtain in this paper are presented as follows.

THEOREM 1.1. *Suppose that (H_1), (H_2), (H_3) hold. Let (y_1^*, u_1^*) be optimal for problem (P_1). We suppose further that (H_4) holds. Then there exist a triplet $(\lambda_0, \xi_0, p) \in R \times X^* \times H^{2,1}_{\frac{2n}{n+2}}(Q)$ with $(\lambda_0, \xi_0) \neq 0$ and a function $\alpha \in \partial g(t, y_1^*)$ a.e. in Q such that*

$$\begin{cases} -\frac{\partial p}{\partial t} + Ap + f_y'(x,t,y_1^*)p + \lambda_0 \alpha + [F'(y_1^*)]^* \xi_0 = 0, & \text{in } Q, \\ p(x,t) = 0, & \text{on } \sum, \\ p(x,T) = 0, & \text{in } \Omega, \end{cases} \tag{1.6}$$

$$p \in \lambda_0 \partial h(u_1^*) \ \ a.e.\ in\, Q, \tag{1.7}$$

$$\langle \xi_0, \psi - F(y_1^*)\rangle_{X^*,X} \le 0 \ \ for\ all\, \psi \in W. \tag{1.8}$$

Moreover, if $[F'(y_1^*)]^*$ *is injective, then* $\lambda_0 \neq 0$.

THEOREM 1.2. *Suppose that* (H_1), (H_2), *and* (H_5) *hold. Let* (y_2^*, u_2^*) *be optimal for problem* (P_2). *We further suppose that* (H_6) *holds. Then there exist a function* $p \in L^2(Q) \cap L^{\frac{2n}{n+2}}(0,T; W_0^{1,\frac{2n}{n+2}}(\Omega))$, *a function* $\alpha \in \partial g(t, y_2^*)$ *a.e. in* Q, *and* $\lambda_0 \in R$ *with* $\lambda_0 \neq 0$ *such that*

$$-\frac{\partial p}{\partial t} + Ap \in L^{\frac{2n}{n+2}}(Q), \tag{1.9}$$

$$\begin{cases} -\frac{\partial p}{\partial t} + Ap + f_y'(x,t,y_2^*)p + \lambda_0 \alpha = 0, & in\, Q, \\ p(x,t) = 0, & on\ \sum, \end{cases} \tag{1.10}$$

$$\langle p(x,0), x_0 - y_2^*(x,0)\rangle - \langle p(x,T), x_1 - y_2^*(x,T)\rangle \le 0 \ \ for\ all\, (x_0,x_1) \in S, \tag{1.11}$$

and

$$p \in \lambda_0 \partial h(u_2^*) \ \ a.e.\ in\, Q. \tag{1.12}$$

THEOREM 1.3. *Let* (H_1), (H_2), (H_3), (H_7), *and* (H_8) *hold. Then problem* (P_1) *has at least one solution.*

THEOREM 1.4. *Let* (H_1), (H_2), (H_5), (H_7), (H_9), *and* (H_{10}) *hold. Then problem* (P_2) *has at least one solution.*

Now we shall point out some special cases of state constraints covered by (1.2) and (1.4) and a special state system covered by (1.1), which is non–well-posed. We stress that the state constraint (1.2) is of the type of an integral. We do not deal with the pointwise state constraint in this paper, for which we refer the reader to [7] and [8].

Example 1.5. Let $X = R^m$ and $h_i \in L^2(Q)$ with $1 \le i \le m$, which are linearly independent in $L^2(Q)$.

Define $F(y) = (\int_Q y(x,t)h_1(x,t)dxdt, \ldots, \int_Q y(x,t)h_m(x,t)dxdt)$. It is clear that F is a linear and bounded operator from $L^2(Q)$ to R^m.

Let $W = ([a_1,b_1], \ldots, [a_m,b_m]) \subset R^m$, $a_i < b_i$, $i = 1, \ldots, m$; then W is convex and closed with finite codimensionality. Thus by Proposition 3.4 of Chapter 4 of [13], the set $(F_y'(y)D_r - W)$ has finite codimensionality in X for any $y \in L^2(Q)$ and $r > 0$.

Consider the state constraint of the form:

$$a_i \le \int_Q y(x,t)h_i(x,t)dxdt \le b_i, \qquad i = 1, \ldots, m. \tag{1.13}$$

It is clear that (1.13) is equivalent to (1.2).

On the other hand, $[F'(y)]^*\xi : R^m \to L^2(Q)$ can be defined by

$$[F'(y)]^*\xi = \sum_{i=1}^m \xi_i h_i \ \text{ for all } \xi = (\xi_1, \ldots, \xi_m) \in R^m.$$

Since $h_1, \ldots, h_m$ are linearly independent, $[F'(y)]^*$ is injective for each $y \in L^2(Q)$.

Example 1.6. Let $S = \{x_0\} \times S_1$, where $x_0 \in H_0^1(\Omega)$ and $S_1 \subset L^2(\Omega)$ has finite codimensionality in $L^2(\Omega)$. We assume that there exists a constant $r > 0$ such that for each $z_0 \in H_0^1(\Omega)$ with $\|z_0\|_{H_0^1(\Omega)} \le r$, and $y \in Y$ with $\|y - y_2^*\|_Y \le r$, the variational system

$$\begin{cases} \frac{\partial z}{\partial t} + Az + f_y'(x,t,y)z = v, & \text{in } Q, \\ z(x,t) = 0, & \text{on } \sum, \\ z(x,0) = z_0, & \text{in } \Omega \end{cases} \tag{1.14}$$

is null controllable, i.e., there exists $(z, v) \in Y \times L^2(Q)$ which satisfies (1.14) and $z(x,T) = 0$.

By Proposition 3.4 in Chapter 4 of [13], one can easily check that (H_6) holds in this case. In other words, state constraint (1.4) covers this case.

From this example, we may see that hypothesis (H_6) mixes some geometric assumptions on set S and some observability assumptions on the variational system corresponding to the state system around (y_2^*, u_2^*).

Example 1.7. Consider the following system:

$$\begin{cases} \frac{\partial y(x,t)}{\partial t} + Ay(x,t) = y^3(x,t) + u(x,t), & \text{in } Q, \\ y(x,t) = 0, & \text{on } \sum, \\ y(x,0) = 0, & \text{in } \Omega, \end{cases} \tag{1.15}$$

where $Q = \Omega \times (0,T)$, $\Omega \subset R^3$, is a bounded domain with smooth boundary. We consider the control set to be $L^2(Q)$.

Let $f(x,t,y) \equiv -y^3$. One can easily check that f satisfies the conditions in (H_2). As we know (cf. [14]), for each $u \in L^2(Q)$, system (1.15) has in general no global solution. This is an unstable or non–well-posed system.

More generally, we may consider the following system:

$$\begin{cases} \frac{\partial y(x,t)}{\partial t} + Ay(x,t) = y(x,t)|y(x,t)|^{q-1} + u(x,t), & \text{in } Q, \\ y(x,t) = 0, & \text{on } \sum, \\ y(x,0) = 0, & \text{in } \Omega, \end{cases}$$

where $1 \le q \le \frac{n}{n-2}$, $n \ge 3$.

Let $f(x,t,y) = -y|y|^{q-1}$; one can check that f satisfies all conditions in (H_2).

Due to their practical applications, there has been a great deal of work on optimal control problems governed by parabolic differential equations; see, for example, [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 16, 17, 18]. In all of this work, the state equations handled are well-posed. In [14], Lions studied a non–well-posed optimal control problem where the cost functional is given by

$$J(y,u) = \frac{1}{6}\|y - y_d\|_{L^6(Q)}^6 + \frac{N}{2}\|u\|_{L^2(Q)}^2 \tag{1.16}$$

and the state system reads

$$\begin{cases} \frac{\partial y}{\partial t} - \Delta y - y^3 = u, & \text{in } Q, \\ y(x,t) = 0, & \text{on } \sum, \\ y(x,0) = 0, & \text{in } \Omega. \end{cases} \tag{1.17}$$

Stimulated by [14], we study optimal control problems (P_1) and (P_2), where the state equations are more general than (1.17) and the cost functionals are more general than (1.16)—in particular, may be nonsmooth. The major novelty of this paper is the presence of state constraints under "finite codimensionality hypotheses" in the non–well-posed optimal control problems. To the best of our knowledge, the maximum principles for such problems have not been studied. The main purpose of this paper is to derive the maximum principles for problems (P_1) and (P_2). So the existence of optimal pairs for problems (P_1) and (P_2) are assumed when we study the maximum principles.

Since the state equations handled in this paper may be non–well-posed, we can not expect the same results for the sensitivity of the states to the controls as those in [1, 2, 8, and 11]. Thus we cannot apply the techniques in [1, 2, 8, and 11] to derive the maximum principles for problems (P_1) and (P_2). Because of the involvement of the state constraints in the non–well-posed optimal control problems and because the cost functionals may be nonsmooth, we cannot apply the approximate method, which was employed in [14], to obtain the maximum principles for problems (P_1) and (P_2). Our main idea for overcoming these difficulties is to transform the original control problems to the optimization problems of two variables y and u by considering the state system as another constraint composed of the state variable y and the control variable u. (There already exist a state constraint (1.2) for problem (P_1) and a state constraint (1.4) for problem (P_2).) More precisely, we introduce two kinds of penalty functionals, via which we may transform the original problems (P_1) and (P_2) into optimization problems (P_1^ε) and (P_2^ε) which have smooth cost functionals and have no constraints. We find optimal solutions for problems (P_1^ε) and (P_2^ε), which turn out to be close to the original optimal pairs of problems (P_1) and (P_2), respectively. Then we derive the necessary conditions for the optimal solutions of the approximate problems. Finally, by passing to the limits for $\varepsilon \to 0$, we may obtain the maximum principles for problems (P_1) and (P_2).

This paper is organized as follows. In the next section, we prove the maximum principle for problem (P_1), i.e., Theorem 1.1. In section 3, we show Theorem 1.2. In section 4, we prove the existence of optimal pairs for problems (P_1) and (P_2), i.e., Theorem 1.3 and Theorem 1.4.

2. Maximum principle for problem (P_1). Throughout this section, we shall assume that (H_1)–(H_4) hold. We start this section with an introduction to the approximations g^ε of g and h_ε of h. For the details, we refer the reader to [2]. Let

$$g^\varepsilon(t,y) = \int_{R^N} g(t, P_N y - \varepsilon \Lambda_N s)\rho(s)ds, \qquad \varepsilon > 0,$$

where ρ is a mollifier in R^N, $N = [\varepsilon^{-1}]$, $P_N : L^2(\Omega) \to X_N$ is the projection of $L^2(\Omega)$ on X_N, which is the finite dimensional space generated by $\{e_i\}_{i=1}^N$, where $\{e_i\}_{i=1}^\infty$ is an orthonormal basis in $L^2(\Omega)$. $\Lambda_N : R^N \to X_N$ is the operator defined by $\Lambda_N(s) = \sum_{i=1}^N s_i e_i$, $s = (s_1, \ldots, s_N)$.

Let $h_\varepsilon : L^2(\Omega) \to R$ be defined by

$$h_\varepsilon(u) = \inf \left\{ \frac{\|u - v\|^2_{L^2(\Omega)}}{2\varepsilon} + h(v) \ : \ v \in L^2(\Omega) \right\}, \ \ \varepsilon > 0.$$

We define penalty functional L_ε on $Y \times L^2(Q)$ by

$$
\begin{aligned}
L_\varepsilon(y,u) = \quad & \int_0^T [g^\varepsilon(t,y) + h_\varepsilon(u)]dt + \frac{1}{2}\int_Q |u-u_1^*|^2 dxdt + \frac{1}{2\tilde{r}}\int_Q |y-y_1^*|^{2\tilde{r}} dxdt \\
& + \tfrac{1}{2\varepsilon}\int_\Omega |y(x,0)-y_0(x)|^2 dx + \frac{1}{2\varepsilon}[\varepsilon + d_W(F(y))]^2 \\
& + \tfrac{1}{2\varepsilon}\int_Q \left|\frac{\partial y}{\partial t} + Ay + f(x,t,y) - u\right|^2 dxdt,
\end{aligned} \tag{2.1}
$$

where $\tilde{r} = \frac{n}{n-2}$. Note that as $y \in Y$, $y(x,0)$ makes sense and belongs to $L^2(\Omega)$ (cf. [15]). In addition, by Sobolev's imbedding theorem, $y \in L^{\frac{2n}{n-2}}(Q)$. Thus, for each $\varepsilon > 0$, L_ε is well defined. Then we may introduce an approximate problem (P_1^ε), for each $\varepsilon > 0$, as follows.

(P_1^ε) Inf $L_\varepsilon(y,u)$ over all $(y,u) \in Y \times L^2(Q)$.

The following two lemmas provide the existence of an optimal solution for approximate problem (P_1^ε) and the convergence of such optimal solutions.

LEMMA 2.1. *For each $\varepsilon > 0$, problem (P_1^ε) has at least one solution.*

Proof. Let $d = \mathrm{Inf}_{(y,u)\in Y\times L^2(Q)} L_\varepsilon(y,u)$. By (H_1), it is obvious that $d > -\infty$. Let $(y_m, u_m) \in Y \times L^2(Q)$ be such that

$$
d \le L_\varepsilon(y_m,u_m) \le d + \frac{1}{m}. \tag{2.2}
$$

By virtue of (2.1) and (2.2), $\{(y_m,u_m)\}$ is bounded in $L^{2\tilde{r}}(Q)\times L^2(Q)$, $\{y_m(x,0)\}$ is bounded in $L^2(\Omega)$, and $\{\frac{\partial y_m}{\partial t} + Ay_m + f(x,t,y_m) - u_m\}$ is bounded in $L^2(Q)$.

On the other hand, by (H_2) and by Sobolev's imbedding theorem, we get that

$$
\{f(x,t,y_m)\} \quad \text{is bounded in } L^2(Q). \tag{2.3}
$$

Thus $\{\frac{\partial y_m}{\partial t} + Ay_m\}$ is bounded in $L^2(Q)$. Then by the same argument as in [15], $\{y_m\}$ is bounded in Y, and so we may extract subsequences of $\{y_m\}$ and $\{u_m\}$, still denoted by themselves, such that

$$
\begin{cases}
u_m \to \tilde{u} \ \text{ weakly in } L^2(Q) \ \text{ as } m \to \infty, \\
y_m \to \tilde{y} \ \text{ weakly in } Y, \text{ strongly in } L^2(Q) \ \text{ as } m \to \infty, \\
y_m(x,t) \to \tilde{y}(x,t) \ \text{ a.e. in } Q \ \text{ as } m \to \infty, \\
y_m(x,0) \to \tilde{y}(x,0) \ \text{ weakly in } L^2(\Omega) \ \text{ as } m \to \infty.
\end{cases} \tag{2.4}
$$

Since f is continuous, we infer that

$$
f(x,t,y_m(x,t)) \to f(x,t,\tilde{y}(x,t)) \ \text{ a.e. in } Q. \tag{2.5}
$$

By (2.3) and (2.5), there exists a subsequence of $\{y_m\}$, still denoted by itself, such that

$$
f(x,t,y_m) \to f(x,t,\tilde{y}) \ \text{ weakly in } L^2(Q) \ \text{ as } m \to \infty, \tag{2.6}
$$

which together with (2.4) shows that

$$
\frac{\partial y_m}{\partial t} + Ay_m + f(x,t,y_m) - u_m \to \frac{\partial \tilde{y}}{\partial t} + A\tilde{y} + f(x,t,\tilde{y}) - \tilde{u} \ \text{ weakly in } L^2(Q). \tag{2.7}
$$

Then it follows from (2.4) and (2.7) that

(2.8)
$$\underline{\lim}_{m\to\infty}\left\|\frac{\partial y_m}{\partial t}+Ay_m+f(x,t,y_m)-u_m\right\|_{L^2(Q)}\geq\left\|\frac{\partial \tilde{y}}{\partial t}+A\tilde{y}+f(x,t,\tilde{y})-\tilde{u}\right\|_{L^2(Q)}$$

and

$$\underline{\lim}_{m\to\infty}\|y_m(x,0)-y_0(x)\|_{L^2(\Omega)}\geq\|\tilde{y}(x,0)-y_0(x)\|_{L^2(\Omega)}. \tag{2.9}$$

Since $y_m\to\tilde{y}$ strongly in $L^2(Q)$, and because of (H_3), we conclude that

$$\frac{1}{2\varepsilon}[\varepsilon+d_W(F(y_m))]^2\to\frac{1}{2\varepsilon}[\varepsilon+d_W(F(\tilde{y}))]^2\ \text{ as } m\to\infty. \tag{2.10}$$

On the other hand, by the same argument as in Chapter 5 of [2], we get that

(2.11)
$$\begin{aligned}&\underline{\lim}_{m\to\infty}\left\{\int_0^T[g^\varepsilon(t,y_m)+h_\varepsilon(u_m)]dt+\frac{1}{2}\int_Q|u_m-u_1^*|^2dxdt+\frac{1}{2\tilde{r}}\int_Q|y_m-y_1^*|^{2\tilde{r}}dxdt\right\}\\&\geq\int_0^T[g^\varepsilon(t,\tilde{y})+h_\varepsilon(\tilde{u})]dt+\frac{1}{2}\int_Q|\tilde{u}-u_1^*|^2dxdt+\frac{1}{2\tilde{r}}\int_Q|\tilde{y}-y_1^*|^{2\tilde{r}}dxdt.\end{aligned}$$

It follows immediately from (2.2) and (2.8)–(2.11) that $L_\varepsilon(\tilde{y},\tilde{u})=d$. This completes the proof. □

LEMMA 2.2. *Let $(y_\varepsilon,u_\varepsilon)$ be optimal for problem (P_1^ε). Then there exists a generalized subsequence of ε, still denoted by itself, such that*

$$y_\varepsilon\to y_1^*\ \textit{strongly in}\ Y\ \ \textit{as}\ \varepsilon\to 0$$

and

$$u_\varepsilon\to u_1^*\ \textit{strongly in}\ L^2(Q)\ \ \textit{as}\ \varepsilon\to 0.$$

Proof. It is clear that

$$L_\varepsilon(y_\varepsilon,u_\varepsilon)\leq L_\varepsilon(y_1^*,u_1^*)=\int_0^T[g^\varepsilon(t,y_1^*)+h_\varepsilon(u_1^*)]dt+\frac{\varepsilon}{2}.$$

By a standard argument as in [2], this implies that

$$\overline{\lim}_{\varepsilon\to 0}L_\varepsilon(y_\varepsilon,u_\varepsilon)\leq\int_0^T[g(t,y_1^*)+h(u_1^*)]dt=L(y_1^*,u_1^*). \tag{2.12}$$

By virtue of (2.1) and by (H_1), $\{y_\varepsilon\}$ is bounded in $L^{2\tilde{r}}(Q)$, $\{u_\varepsilon\}$ is bounded in $L^2(Q)$, and

$$\int_\Omega|y_\varepsilon(x,0)-y_0(x)|^2dx\leq C\varepsilon, \tag{2.13}$$

$$\int_Q\left|\frac{\partial y_\varepsilon}{\partial t}+Ay_\varepsilon+f(x,t,y_\varepsilon)-u_\varepsilon\right|^2dxdt\leq C\varepsilon \tag{2.14}$$

for some constant C independent of ε.

By the same argument as in the proof of Lemma 2.1, there exists a generalized subsequence of ε, still denoted by itself, such that

$$y_\varepsilon \to \tilde{y} \text{ weakly in } Y \text{ and strongly in } L^2(Q) \text{ as } \varepsilon \to 0, \tag{2.15}$$

$$y_\varepsilon(x,0) \to \tilde{y}(x,0) \text{ weakly in } L^2(\Omega) \text{ as } \varepsilon \to 0, \tag{2.16}$$

$$u_\varepsilon \to \tilde{u} \text{ weakly in } L^2(Q) \text{ as } \varepsilon \to 0, \tag{2.17}$$

and

$$\frac{\partial y_\varepsilon}{\partial t} + Ay_\varepsilon + f(x,t,y_\varepsilon) - u_\varepsilon \to \frac{\partial \tilde{y}}{\partial t} + A\tilde{y} + f(x,t,\tilde{y}) - \tilde{u} \text{ weakly in } L^2(Q). \tag{2.18}$$

It follows from (2.13), (2.14), (2.16), and (2.18) that

$$\begin{cases} \frac{\partial \tilde{y}}{\partial t} + A\tilde{y} + f(x,t,\tilde{y}) - \tilde{u} = 0, & \text{in } Q, \\ \tilde{y}(x,t) = 0, & \text{on } \sum, \\ \tilde{y}(x,0) = y_0(x), & \text{in } \Omega. \end{cases} \tag{2.19}$$

On the other hand, by (2.1) and (2.12), we know that $\frac{1}{2\varepsilon}[\varepsilon + d_W(F(y_\varepsilon))]^2 \le C$, which shows that $d_W(F(y_\varepsilon)) \to 0$ as $\varepsilon \to 0$. This together with (2.15) and (H_3) yields that

$$F(\tilde{y}) \in W. \tag{2.20}$$

Since (y_1^*, u_1^*) is an optimal pair of problem (P_1), it follows from (2.19) and (2.20) that

$$L(y_1^*, u_1^*) \le L(\tilde{y}, \tilde{u}). \tag{2.21}$$

By the same argument as in [2], we get that

$$\underline{\lim}_{\varepsilon \to 0} \int_0^T [g^\varepsilon(t, y_\varepsilon) + h_\varepsilon(u_\varepsilon)]dt \ge \int_0^T [g(t,\tilde{y}) + h(\tilde{u})]dt. \tag{2.22}$$

By (2.1),(2.21), and (2.22), we deduce that

$$\underline{\lim}_{\varepsilon \to 0} L_\varepsilon(y_\varepsilon, u_\varepsilon) \ge L(y_1^*, u_1^*). \tag{2.23}$$

Then by (2.1), (2.12), and (2.23), there exists a generalized subsequence of ε, still denoted by itself, such that

$$y_\varepsilon \to y_1^* \text{ strongly in } L^{2\tilde{r}}(Q), \text{ weakly in } Y \text{ as } \varepsilon \to 0, \tag{2.24}$$

$$y_\varepsilon(x,0) \to y_1^*(x,0) \text{ strongly in } L^2(\Omega) \text{ as } \varepsilon \to 0, \tag{2.25}$$

$$u_\varepsilon \to u_1^* \text{ strongly in } L^2(Q) \text{ as } \varepsilon \to 0. \tag{2.26}$$

Now we claim that

$$f(x,t,y_\varepsilon) \to f(x,t,y_1^*) \text{ strongly in } L^2(Q) \text{ as } \varepsilon \to 0. \tag{2.27}$$

Here is the argument. We observe first that

$$|f(x,t,y_\varepsilon) - f(x,t,y_1^*)| = |y_\varepsilon - y_1^*| \cdot |a_\varepsilon(x,t)|,$$

where $a_\varepsilon(x,t) = \int_0^1 f_y'(x,t,y_1^* + \theta(y_\varepsilon - y_1^*))d\theta$.

By (H_2), we yield that $a_\varepsilon^2(x,t) \le \{\tilde{a}_1(x,t) + \tilde{b}_1[|y_1^*| + |y_\varepsilon|]^{r_1 - 1}\}^2$, which, combined with Sobolev's imbedding theorem, indicates that $\{a_\varepsilon^2(x,t)\}$ is bounded in $L^{\frac{n}{2}}(Q)$. Then by Holder's inequality, we infer that

$$\int_Q |f(x,t,y_\varepsilon) - f(x,t,y_1^*)|^2 dxdt \le \left[\int_Q |y_\varepsilon - y_1^*|^{2\tilde{r}} dxdt\right]^{\frac{n-2}{n}} \cdot \left[\int_Q |a_\varepsilon|^n dxdt\right]^{\frac{2}{n}},$$

which immediately implies (2.27).

Now by (2.14), (2.26), and (2.27), we infer that

$$\frac{\partial(y_\varepsilon - y_1^*)}{\partial t} + A(y_\varepsilon - y_1^*) \to 0 \text{ strongly in } L^2(Q) \text{ as } \varepsilon \to 0.$$

This together with (2.25) shows that (cf. [15]) $y_\varepsilon \to y_1^*$ strongly in Y as $\varepsilon \to 0$, which completes the proof. □

Now we turn to proving Theorem 1.1.

Proof of Theorem 1.1. Let $Z = \{z \in Y \mid z(x,0) = 0\}$. As we know (cf. [15]), the space Z is a Hilbert space endowed with the norm of Y. Let $(z,v) \in Z \times L^2(Q)$ be arbitrary but fixed, and set $y_\varepsilon^\rho = y_\varepsilon + \rho z$, $u_\varepsilon^\rho = u_\varepsilon + \rho v$, where $\rho > 0$. It is clear that $(y_\varepsilon^\rho, u_\varepsilon^\rho) \in Y \times L^2(Q)$, and so

$$\frac{L_\varepsilon(y_\varepsilon^\rho, u_\varepsilon^\rho) - L_\varepsilon(y_\varepsilon, u_\varepsilon)}{\rho} \ge 0 \text{ for all } \rho > 0. \tag{2.28}$$

After some simple calculations, we obtain that

$$\begin{aligned} &\lim_{\rho\to 0} \int_0^T \frac{g^\varepsilon(t,y_\varepsilon^\rho) - g^\varepsilon(t,y_\varepsilon)}{\rho} dt = \int_Q \nabla g^\varepsilon(t,y_\varepsilon) z dxdt, \\ &\lim_{\rho\to 0} \int_0^T \frac{h_\varepsilon(u_\varepsilon^\rho) - h_\varepsilon(u_\varepsilon)}{\rho} dt = \int_Q \nabla h_\varepsilon(u_\varepsilon) v dxdt, \end{aligned} \tag{2.29}$$

$$\begin{aligned} &\lim_{\rho\to 0} \tfrac{1}{2\rho} \int_Q \left[|u_\varepsilon^\rho - u_1^*|^2 - |u_\varepsilon - u_1^*|^2\right] dxdt = \int_Q (u_\varepsilon - u_1^*) v dxdt, \\ &\lim_{\rho\to 0} \tfrac{1}{2\tilde{r}\rho} \int_Q \left[|y_\varepsilon^\rho - y_1^*|^{2\tilde{r}} - |y_\varepsilon - y_1^*|^{2\tilde{r}}\right] dxdt = \int_Q (y_\varepsilon - y_1^*)^{2\tilde{r}-1} z dxdt. \end{aligned} \tag{2.30}$$

Since $y_\varepsilon^\rho(x,0) = y_\varepsilon(x,0)$, we have

$$\frac{1}{2\varepsilon\rho} \int_\Omega \{[y_\varepsilon^\rho(x,0) - y_0(x)]^2 - [y_\varepsilon(x,0) - y_0(x)]^2\} dx = 0 \text{ for all } \rho > 0. \tag{2.31}$$

Now we claim that there is a generalized subsequence of ρ, denoted by itself again, such that

(2.32)
$$\lim_{\rho\to 0}\frac{1}{2\varepsilon\rho}\int_Q\left[\left|\frac{\partial y_\varepsilon^\rho}{\partial t}+Ay_\varepsilon^\rho+f(x,t,y_\varepsilon^\rho)-u_\varepsilon^\rho\right|^2-\left|\frac{\partial y_\varepsilon}{\partial t}+Ay_\varepsilon+f(x,t,y_\varepsilon)-u_\varepsilon\right|^2\right]dxdt$$
$$=\frac{1}{\varepsilon}\int_Q\left[\frac{\partial y_\varepsilon}{\partial t}+Ay_\varepsilon+f(x,t,y_\varepsilon)-u_\varepsilon\right]\cdot\left[\frac{\partial z}{\partial t}+Az+f_y'(x,t,y_\varepsilon)z-v\right]dxdt.$$

Here is the argument. We observe first that

$$\frac{f(x,t,y_\varepsilon^\rho)-f(x,t,y_\varepsilon)}{\rho}=a_\varepsilon^\rho(x,t)z(x,t),$$

where $a_\varepsilon^\rho(x,t)=\int_0^1 f_y'(x,t,y_\varepsilon(x,t)+\theta\rho z(x,t))d\theta$. Since $f_y'(x,t,\cdot)$ is continuous and $(y_\varepsilon^\rho,u_\varepsilon^\rho)\to(y_\varepsilon,u_\varepsilon)$ strongly in $Y\times L^2(Q)$ as $\rho\to 0$, there exists a generalized subsequence of ρ, still denoted by itself, such that

$$[a_\varepsilon^\rho(x,t)z(x,t)-f_y'(x,t,y_\varepsilon(x,t))z(x,t)]^2\to 0\ \text{ a.e. in } Q\ \text{ as } \rho\to 0.$$

It follows from (H_2) that

$$[a_\varepsilon^\rho(x,t)z-f_y'(x,t,y_\varepsilon(x,t))z]^2\le 4\{\tilde a_1(x,t)+\tilde b_1[|y_\varepsilon|+|z|]^{r_1-1}\}^2z^2\in L^1(Q);$$

here we have used Sobolev's imbedding theorem and Holder's inequality. Then by the Lebesgue dominated convergence theorem, we obtain that

$$a_\varepsilon^\rho(x,t)z\to f_y'(x,t,y_\varepsilon)z\ \text{ strongly in } L^2(Q)\ \text{ as } \rho\to 0, \tag{2.33}$$

which implies that

$$f(x,t,y_\varepsilon^\rho)\to f(x,t,y_\varepsilon)\ \text{ strongly in } L^2(Q)\ \text{ as } \rho\to 0. \tag{2.34}$$

Thus (2.32) follows immediately from (2.33) and (2.34).

By the same argument as in [18], we get that

$$\begin{aligned}&\lim_{\rho\to 0}\frac{1}{2\varepsilon\rho}\{[\varepsilon+d_W(F(y_\varepsilon^\rho))]^2-[\varepsilon+d_W(F(y_\varepsilon))]^2\}\\&=\frac{\varepsilon+d_W(F(y_\varepsilon))}{\varepsilon}\langle\xi_\varepsilon,F'(y_\varepsilon)z\rangle_{X^*,X},\end{aligned} \tag{2.35}$$

where

$$\xi_\varepsilon\in\partial d_W(F(y_\varepsilon))=\begin{cases}\nabla d_W(F(y_\varepsilon)), & \text{if } F(y_\varepsilon)\notin W,\\ 0, & \text{if } F(y_\varepsilon)\in W\end{cases}$$

and

$$\|\xi_\varepsilon\|_{X^*}=\begin{cases}1, & \text{if } F(y_\varepsilon)\notin W,\\ 0, & \text{if } F(y_\varepsilon)\in W.\end{cases} \tag{2.36}$$

Let

$$\lambda_\varepsilon=\frac{\varepsilon}{\varepsilon+d_W(F(y_\varepsilon))}\ \text{ and }\ \mu_\varepsilon=\frac{1}{\varepsilon+d_W(F(y_\varepsilon))}. \tag{2.37}$$

It is obvious that

$$1 \le \lambda_\varepsilon^2 + \|\xi_\varepsilon\|_{X^*}^2 \le 2. \tag{2.38}$$

Now it follows from (2.28)–(2.37) that

$$\begin{aligned} &\lambda_\varepsilon \left\{ \int_Q \nabla g^\varepsilon(t, y_\varepsilon) z dx dt + \int_Q \nabla h_\varepsilon(u_\varepsilon) v dx dt + \int_Q (u_\varepsilon - u_1^*) v dx dt \right. \\ &\left. + \int_Q (y_\varepsilon - y_1^*)^{2\tilde{r}-1} z dx dt \right\} + \int_Q [F'(y_\varepsilon)]^* \xi_\varepsilon z dx dt \\ &+ \mu_\varepsilon \int_Q \left(\frac{\partial y_\varepsilon}{\partial t} + A y_\varepsilon + f(x,t,y_\varepsilon) - u_\varepsilon \right) \left(\frac{\partial z}{\partial t} + Az + f_y'(x,t,y_\varepsilon) z - v \right) dx dt \ge 0 \end{aligned} \tag{2.39}$$

for all $(z, v) \in Z \times L^2(Q)$.

Let

$$p_\varepsilon = \mu_\varepsilon \left(\frac{\partial y_\varepsilon}{\partial t} + A y_\varepsilon + f(x,t,y_\varepsilon) - u_\varepsilon \right). \tag{2.40}$$

It is clear that $p_\varepsilon \in L^2(Q)$. By taking $z = 0$ in (2.39), we obtain that

$$\lambda_\varepsilon \left\{ \int_Q \nabla h_\varepsilon(u_\varepsilon) v dx dt + \int_Q (u_\varepsilon - u_1^*) v dx dt \right\} - \int_Q p_\varepsilon v dx dt \ge 0 \tag{2.41}$$

for all $v \in L^2(Q)$. This implies that

$$p_\varepsilon = \lambda_\varepsilon [\, \nabla h_\varepsilon(u_\varepsilon) + u_\varepsilon - u_1^* \,] \text{ a.e. in } Q. \tag{2.42}$$

By Lemma 2.2, $u_\varepsilon \to u_1^*$ strongly in $L^2(Q)$ as $\varepsilon \to 0$. Then by the same argument as in [2], $\{\nabla h_\varepsilon(u_\varepsilon)\}$ is bounded in $L^2(Q)$. Thus it follows from (2.42) that $\{p_\varepsilon\}$ is bounded in $L^2(Q)$.

By taking $v = 0$ in (2.39), we obtain that

$$\begin{aligned} &\int_Q \left\{ \lambda_\varepsilon [\, \nabla g^\varepsilon(t, y_\varepsilon) + (y_\varepsilon - y_1^*)^{2\tilde{r}-1} \,] + [F'(y_\varepsilon)]^* \xi_\varepsilon \right\} z dx dt \\ &+ \int_Q p_\varepsilon \left(\frac{\partial z}{\partial t} + Az + f_y'(x,t,y_\varepsilon) \right) z dx dt = 0 \text{ for all } z \in Z. \end{aligned} \tag{2.43}$$

This shows that (cf. [14])

$$\begin{cases} \frac{\partial p_\varepsilon}{\partial t} - A p_\varepsilon - f_y'(x,t,y_\varepsilon) p_\varepsilon = \lambda_\varepsilon [\nabla g^\varepsilon(t, y_\varepsilon) + (y_\varepsilon - y_1^*)^{2\tilde{r}-1}] + [F'(y_\varepsilon)]^* \xi_\varepsilon, & \text{in } Q, \\ p_\varepsilon(x,t) = 0, & \text{on } \sum, \\ p_\varepsilon(x,T) = 0, & \text{in } \Omega. \end{cases} \tag{2.44}$$

Note that (2.42) and (2.44) can be regarded as necessary conditions for $(y_\varepsilon, u_\varepsilon)$ to be optimal for problem (P_1^ε).

Now we are going to pass to the limit for $\varepsilon \to 0$ in (2.42) and (2.44) (or (2.43)) and derive the necessary conditions for (y_1^*, u_1^*) to be optimal for problem (P_1).

First we deal with (2.42). By (2.38), there exists a generalized subsequence of $\{\lambda_\varepsilon\}$, still denoted by itself, such that

$$\lambda_\varepsilon \to \lambda_0 \text{ as } \varepsilon \to 0. \tag{2.45}$$

By Lemma 2.2 and by the same argument as in [2], there exist $\alpha \in \partial g(t, y_1^*)$, $\beta \in \partial h(u_1^*)$ a.e. in Q such that on a generalized subsequence of ε, still denoted by itself,

$$\nabla g^{\varepsilon}(t, y_{\varepsilon}) \to \alpha \text{ weakly in } L^2(Q) \text{ as } \varepsilon \to 0 \tag{2.46}$$

and

$$\nabla h_{\varepsilon}(u_{\varepsilon}) \to \beta \text{ weakly in } L^2(Q) \text{ as } \varepsilon \to 0. \tag{2.47}$$

Since $\{p_{\varepsilon}\}$ is bounded in $L^2(Q)$, there exists a generalized subsequence of ε, still denoted by itself, such that

$$p_{\varepsilon} \to p \text{ weakly in } L^2(Q) \text{ as } \varepsilon \to 0. \tag{2.48}$$

By (2.45), (2.47), and (2.48), we may pass to the limit for $\varepsilon \to 0$ in (2.42) and obtain (1.7).

Next we deal with (2.44). We claim the following inequality:

$$\|p_{\varepsilon} f_y'(x, t, y_{\varepsilon})\|_{L^{\frac{2n}{n+2}}(Q)} \le C \text{ for some positive constant } C \text{ independent of } \varepsilon. \tag{2.49}$$

Here is the argument. By Lemma 2.2 and by Sobolev's imbedding theorem, $\{|y_{\varepsilon}|^{r_1 - 1}\}$ is bounded in $L^n(Q)$. Thus it follows from (H_2) that $\{f_y'(x, t, y_{\varepsilon})\}$ is bounded in $L^n(Q)$. Then by Holder's inequality, we obtain that

$$\int_Q |p_{\varepsilon} f_y'(x, t, y_{\varepsilon})|^{\frac{2n}{n+2}} dx dt \le \left[\int_Q |p_{\varepsilon}|^2 dx dt\right]^{\frac{n}{n+2}} \left[\int_Q |f_y'(x, t, y_{\varepsilon})|^n dx dt\right]^{\frac{2}{n+2}},$$

which implies (2.49), because $\{p_{\varepsilon}\}$ is bounded in $L^2(Q)$.

On the other hand, by (2.38), Lemma 2.2, and by Sobolev's imbedding theorem, we get that there exists a generalized subsequence of ε, still denoted by itself, such that

$$\lambda_{\varepsilon}(y_{\varepsilon} - y_1^*)^{2\tilde{r} - 1} \to 0 \text{ strongly in } L^{\frac{2n}{n+2}}(Q) \text{ as } \varepsilon \to 0 \tag{2.50}$$

and

$$\xi_{\varepsilon} \to \xi_0 \text{ weakly star in } X^* \text{ as } \varepsilon \to 0,$$

which, combined with (H_3) and Lemma 2.2, indicate that

$$[F'(y_{\varepsilon})]^* \xi_{\varepsilon} \to [F'(y_1^*)]^* \xi_0 \text{ weakly in } L^2(Q) \text{ as } \varepsilon \to 0. \tag{2.51}$$

Now it follows from $(2.44)_1$ (the first equality of (2.44)), (2.49), (2.50), and (2.51) that $\{-\frac{\partial p_{\varepsilon}}{\partial t} + A p_{\varepsilon}\}$ is bounded in $L^{\frac{2n}{n+2}}(Q)$. Then by $(2.44)_2$ and $(2.44)_3$, $\{p_{\varepsilon}\}$ is bounded in $H^{2,1}_{\frac{2n}{n+2}}(Q)$ (cf. [14]). So there exist $p \in H^{2,1}_{\frac{2n}{n+2}}(Q)$ and a generalized subsequence of $\{p_{\varepsilon}\}$, still denoted by itself, such that

$$p_{\varepsilon} \to p \text{ weakly in } H^{2,1}_{\frac{2n}{n+2}}(Q) \text{ as } \varepsilon \to 0 \tag{2.52}$$

and

$$p_{\varepsilon}(x, t) \to p(x, t) \text{ a.e. in } Q \text{ as } \varepsilon \to 0. \tag{2.53}$$

Since $f'_y(x,t,\cdot)$ is continuous, by Lemma 2.2 and by (2.53), there exists a generalized subsequence of ε, still denoted by itself, such that

$$p_\varepsilon(x,t)f'_y(x,t,y_\varepsilon(x,t)) \to p(x,t)f'_y(x,t,y_1^*(x,t)) \text{ a.e. in } Q \text{ as } \varepsilon \to 0. \tag{2.54}$$

By (2.49) and (2.54), we obtain that

$$p_\varepsilon f'_y(x,t,y_\varepsilon) \to p f'_y(x,t,y_1^*) \text{ weakly in } L^{\frac{2n}{n+2}}(Q) \text{ as } \varepsilon \to 0. \tag{2.55}$$

From (2.45), (2.46), (2.50), (2.51), (2.52), and (2.55), we may pass to the limit for $\varepsilon \to 0$ in (2.44) to yield (1.6) (cf. [14]).

Now we turn to proving (1.8). By the definition of the subdifferential of the distance function $d_W(\cdot)$, we have

$$\langle \xi_\varepsilon, \psi - F(y_\varepsilon)\rangle_{X^*,X} \le 0 \text{ for all } \psi \in W.$$

It follows from (H_3) that $F(y_\varepsilon) \to F(y_1^*)$ strongly in X as $\varepsilon \to 0$. Thus we deduce that

$$\langle \xi_\varepsilon, \psi - F(y_1^*)\rangle_{X^*,X} \le \langle \xi_\varepsilon, F(y_\varepsilon) - F(y_1^*)\rangle_{X^*,X} \text{ for all } \psi \in W. \tag{2.56}$$

By taking the limit for $\varepsilon \to 0$ in (2.56), we get (1.8).

It remains to show that $(\lambda_0, \xi_0) \neq 0$. If $\lambda_0 = 0$, then it follows from (2.38) that there exist constants $\delta > 0$ and $\varepsilon_1 > 0$ such that

$$\|\xi_\varepsilon\|_{X^*} \ge \delta > 0 \text{ for all } \varepsilon < \varepsilon_1. \tag{2.57}$$

By (2.39) and (2.56), we obtain that

(2.58)

$$\langle \xi_\varepsilon, F'(y_1^*)z - \psi + F(y_1^*)\rangle_{X^*,X} + \int_Q p_\varepsilon \left[\frac{\partial z}{\partial t} + Az + f'_y(x,t,y_\varepsilon)z - v\right] dxdt \ge -\eta_\varepsilon(z,v)$$

for all $(z,v) \in Z \times L^2(Q)$, where

$$\begin{aligned}\eta_\varepsilon(z,v) = \lambda_\varepsilon \Big\{ &\int_Q \nabla g^\varepsilon(t,y_\varepsilon) z dxdt + \int_Q \nabla h_\varepsilon(u_\varepsilon) v dxdt + \int_Q (u_\varepsilon - u_1^*) v dxdt \\ &+ \int_Q (y_\varepsilon - y_1^*)^{2\tilde{r}-1} z dxdt \Big\} + \langle \xi_\varepsilon, [F'(y_\varepsilon) - F'(y_1^*)]z + F(y_\varepsilon) - F(y_1^*)\rangle_{X^*,X}.\end{aligned}$$

For any $z \in D_r$, which was defined in section 1, and $\varepsilon > 0$, by taking $v = v_\varepsilon(z) \equiv \frac{\partial z}{\partial t} + Az + f'_y(x,t,y_\varepsilon)z$ in (2.58), we obtain that

$$\langle \xi_\varepsilon, F'(y_1^*)z - \psi + F(y_1^*)\rangle_{X^*,X} \ge -\eta_\varepsilon(z, v_\varepsilon(z)) \text{ for all } z \in D_r \text{ and } \varepsilon > 0. \tag{2.59}$$

On the other hand, by (H_2) and Lemma 2.2, there exists a positive constant, denoted by ε_1 again, such that $\|f'_y(x,t,y_\varepsilon)\|_{L^n(Q)} \le C$ for all $\varepsilon < \varepsilon_1$, where C is a constant independent of ε. Then by Holder's inequality, we infer that

$$\begin{aligned}&\int_Q |f'_y(x,t,y_\varepsilon)z|^2 dxdt \\ &\le \left[\int_Q |z|^{\frac{2n}{n-2}} dxdt\right]^{\frac{n-2}{n}} \cdot \left[\int_Q |f'_y(x,t,y_\varepsilon)|^n dxdt\right]^{\frac{2}{n}} \\ &\le Cr^2\end{aligned}$$

for all $z \in D_r$ and $\varepsilon < \varepsilon_1$, where C is a constant independent of ε and z. This shows that

$$\|v_\varepsilon(z)\|_{L^2(Q)} \le C \ \text{ for all } z \in D_r \ \text{ and } \varepsilon < \varepsilon_1. \tag{2.60}$$

By (H_3) and by (2.38), we infer that

$$\langle \xi_\varepsilon, [F'(y_\varepsilon) - F'(y_1^*)]z\rangle_{X^*,X} \to 0 \ \text{ as } \varepsilon \to 0, \ \text{ uniformly in } z \in D_r. \tag{2.61}$$

It follows from (2.60) and (2.61) that

$$\eta_\varepsilon(z, v_\varepsilon(z)) \to 0 \ \text{ as } \varepsilon \to 0, \ \text{ uniformly in } z \in D_r. \tag{2.62}$$

By (H_4), $F'(y_1^*)D_r - W + \{F(y_1^*)\}$ has finite codimensionality in X. Thanks to Lemma 3.6 of Chapter 4 of [13], we conclude from (2.57), (2.59), and (2.62) that $\xi_0 \neq 0$. Hence

$$(\lambda_0, \xi_0) \neq 0. \tag{2.63}$$

Finally, in the case in which $[F'(y_1^*)]$ is injective, if $\lambda_0 = 0$, then by (1.7) we obtain that $p = 0$. Hence, by (1.6), $[F'(y_1^*)]^*\xi_0 = 0$, which shows that $\xi_0 = 0$. This contradicts (2.63). So $\lambda_0 \neq 0$ in this case. This completes the proof. ∎

3. Maximum principle for problem (P_2). In this section, we shall prove Theorem 1.2. The main steps here are similar to those in section 2. However, the penalty functional is different from that in section 2 because we consider a two-point boundary constraint (time variable) here. For this reason we shall prove all results in detail.

First, for each $\varepsilon > 0$, we define a penalty functional L_ε on $Y \times L^2(Q)$ by

$$\begin{aligned} L_\varepsilon(y,u) = &\int_0^T [g^\varepsilon(t,y) + h_\varepsilon(u)]dt + \frac{1}{2}\int_Q |u - u_2^*|^2 dxdt + \frac{1}{2\tilde{r}}\int_Q |y - y_2^*|^{2\tilde{r}} dxdt \\ &+ \tfrac{1}{2}\int_\Omega [\ |y(x,0) - y_2^*(x,0)|^2 + |y(x,T) - y_2^*(x,T)|^2\ ]dx \\ &+ \tfrac{1}{2\varepsilon}[d_S(y(x,0), y(x,T)) + \varepsilon]^2 + \tfrac{1}{2\varepsilon}\int_Q \left| \frac{\partial y}{\partial t} + Ay + f(x,t,y) - u \right|^2 dxdt, \end{aligned} \tag{3.1}$$

where $\tilde{r} = \frac{n}{n-2}$.

Because $y \in Y$, we have that $y \in L^{2\tilde{r}}(Q)$ and $y(x,0), y(x,T) \in L^2(\Omega)$ (cf. [15]). Thus L_ε is well defined. So we may consider an approximate problem (P_2^ε) as follows.

(P_2^ε) Inf $L_\varepsilon(y,u)$ over all $(y,u) \in Y \times L^2(Q)$.

The following two lemmas provide the existence of optimal solutions of approximate problem (P_2^ε) and the convergence of such optimal solutions.

LEMMA 3.1. *For each $\varepsilon > 0$, problem (P_2^ε) has at least one solution.*

Proof. Let $(y_m, u_m) \in Y \times L^2(Q)$ be such that

$$d \le L_\varepsilon(y,u) \le d + \frac{1}{m}, \tag{3.2}$$

where $d = \mathrm{Inf}_{(y,u)\in Y\times L^2(Q)}\ L_\varepsilon(y,u)$. It is clear that $d > -\infty$.

By virtue of (3.1), using the same argument as in the proof of Lemma 2.1, one can show that there exist subsequences of $\{y_m\}$ and $\{u_m\}$, still denoted by themselves, such that

$$u_m \to \tilde{u} \ \text{ weakly in } L^2(Q) \text{ as } m \to \infty, \tag{3.3}$$

$$y_m \to \tilde{y} \ \text{ weakly in } Y, \text{ strongly in } L^2(Q) \text{ as } m \to \infty, \tag{3.4}$$

$$y_m(x,t) \to \tilde{y}(x,t) \ \text{ a.e. in } Q \ \text{ as } m \to \infty, \tag{3.5}$$

$$(y_m(x,0), y_m(x,T)) \to (\tilde{y}(x,0), \tilde{y}(x,T)) \ \text{ weakly in } L^2(\Omega) \times L^2(\Omega) \ \text{ as } m \to \infty. \tag{3.6}$$

By (3.6), we obtain that

$$\begin{aligned} &\varliminf_{m\to\infty} [\, \|y_m(x,0) - y_2^*(x,0)\|^2_{L^2(\Omega)} + \|y_m(x,T) - y_2^*(x,T)\|^2_{L^2(\Omega)} \,] \\ &\geq \|\tilde{y}(x,0) - y_2^*(x,0)\|^2_{L^2(\Omega)} + \|\tilde{y}(x,T) - y_2^*(x,T)\|^2_{L^2(\Omega)}. \end{aligned} \tag{3.7}$$

Since S is convex and closed in $L^2(\Omega) \times L^2(\Omega)$, $d_S(\cdot,\cdot)$ is weakly lower semicontinuous. Thus it follows from (3.6) that

$$\varliminf_{m\to\infty} d_S(y_m(x,0), y_m(x,T)) \geq d_S(\tilde{y}(x,0), \tilde{y}(x,T)). \tag{3.8}$$

By (3.7) and (3.8), using the same argument as in the proof of Lemma 2.1, we find that

$$\varliminf_{m\to\infty} L_\varepsilon(y_m, u_m) \geq L_\varepsilon(\tilde{y}, \tilde{u}),$$

which together with (3.2) shows that $L_\varepsilon(\tilde{y}, \tilde{u}) = d$. This completes the proof. □

LEMMA 3.2. *Let $(y_\varepsilon, u_\varepsilon)$ be optimal for problem (P_2^ε). Then there exists a generalized subsequence of ε, still denoted by itself, such that*

$$y_\varepsilon \to y_2^* \ \textit{strongly in } Y, \ \ u_\varepsilon \to u_2^* \ \textit{strongly in } L^2(Q) \ \textit{ as } \varepsilon \to 0.$$

Proof. It is clear that

$$L_\varepsilon(y_\varepsilon, u_\varepsilon) \leq L_\varepsilon(y_2^*, u_2^*) = \int_0^T [g^\varepsilon(t, y_2^*) + h_\varepsilon(u_2^*)]dt + \frac{\varepsilon}{2},$$

which implies that

$$\overline{\varlimsup_{\varepsilon\to 0}} L_\varepsilon(y_\varepsilon, u_\varepsilon) \leq L(y_2^*, u_2^*). \tag{3.9}$$

On the other hand, by virtue of (3.1) and by the same argument as in the proof of Lemma 2.2, we obtain that there exist $\tilde{y} \in Y$ and a generalized subsequence of ε, still denoted by itself, such that

$$y_\varepsilon \to \tilde{y} \ \text{ weakly in } Y, \text{ strongly in } L^2(Q) \text{ as } \varepsilon \to 0, \tag{3.10}$$

$$(y_\varepsilon(x,0), y_\varepsilon(x,T)) \to (\tilde{y}(x,0), \tilde{y}(x,T)) \ \text{ weakly in } L^2(\Omega) \times L^2(\Omega) \text{ as } \varepsilon \to 0, \tag{3.11}$$

$$u_\varepsilon \to \tilde{u} \ \text{ weakly in } L^2(Q) \text{ as } \varepsilon \to 0, \tag{3.12}$$

and

$$\frac{\partial y_\varepsilon}{\partial t} + Ay_\varepsilon + f(x,t,y_\varepsilon) - u_\varepsilon \to \frac{\partial \tilde{y}}{\partial t} + A\tilde{y} + f(x,t,\tilde{y}) - \tilde{u} \ \text{ weakly in } L^2(Q). \tag{3.13}$$

Moreover, $\tilde{y}$ satisfies the following equation:

$$\begin{cases} \frac{\partial \tilde{y}}{\partial t} + A\tilde{y} + f(x,t,\tilde{y}) = \tilde{u}, & \text{in } Q, \\ \tilde{y}(x,t) = 0, & \text{on } \sum. \end{cases} \tag{3.14}$$

It follows from (3.1) and (3.9) that

$$d_S(y_\varepsilon(x,0), y_\varepsilon(x,T)) \le C\sqrt{\varepsilon}, \tag{3.15}$$

where C is a positive constant independent of ε. Since S is convex and closed, it follows from (3.11) and (3.15) that

$$d_S(\tilde{y}(x,0), \tilde{y}(x,T)) \le \varliminf_{\varepsilon\to 0} d_S(y_\varepsilon(x,0), y_\varepsilon(x,T)) = 0,$$

which implies that

$$(\tilde{y}(x,0), \tilde{y}(x,T)) \in S. \tag{3.16}$$

By (3.14) and (3.16), we deduce that

$$L(\tilde{y}, \tilde{u}) \ge L(y_2^*, u_2^*). \tag{3.17}$$

Then by (3.10), (3.11), (3.12), (3.13), and (3.17), using the same argument as in the proof of Lemma 2.2, we obtain that

$$\varliminf_{\varepsilon\to 0} L_\varepsilon(y_\varepsilon, u_\varepsilon) \ge L(y_2^*, u_2^*). \tag{3.18}$$

Now by (3.1), (3.9), and (3.18), there exists a generalized subsequence of ε, still denoted by itself, such that

$$y_\varepsilon \to y_2^* \text{ strongly in } L^{2\tilde{r}}(Q) \text{ as } \varepsilon \to 0, \tag{3.19}$$

$$(y_\varepsilon(x,0), y_\varepsilon(x,T)) \to (y_2^*(x,0), y_2^*(x,T)) \text{ strongly in } L^2(\Omega) \text{ as } \varepsilon \to 0, \tag{3.20}$$

$$u_\varepsilon \to u_2^* \text{ strongly in } L^2(Q) \text{ as } \varepsilon \to 0. \tag{3.21}$$

From (3.19), (3.20), and (3.21), using the same argument as in the proof of Lemma 2.2, we infer that $y_\varepsilon \to y_2^*$ strongly in Y as $\varepsilon \to 0$. This completes the proof. □

Now we are ready to prove Theorem 1.2.

Proof of Theorem 1.2. Let $(z,v) \in Y \times L^2(Q)$ be arbitrary but fixed. For each $\rho > 0$, set $y_\varepsilon^\rho = y_\varepsilon + \rho z$, $u_\varepsilon^\rho = u_\varepsilon + \rho v$. It is clear that $(y_\varepsilon^\rho, u_\varepsilon^\rho) \in Y \times L^2(Q)$ and $y_\varepsilon^\rho \to y_\varepsilon$ strongly in Y, $u_\varepsilon^\rho \to u_\varepsilon$ strongly in $L^2(Q)$ as $\rho \to 0$. Moreover, we have that

$$\frac{L_\varepsilon(y_\varepsilon^\rho, u_\varepsilon^\rho) - L_\varepsilon(y_\varepsilon, u_\varepsilon)}{\rho} \ge 0 \text{ for all } \rho > 0. \tag{3.22}$$

By (H_5) and by the same argument as in [16], we obtain that

$$\begin{aligned} &\lim_{\rho\to 0} \tfrac{1}{2\varepsilon\rho}\{[\varepsilon + d_S(y_\varepsilon^\rho(x,0), y_\varepsilon^\rho(x,T))]^2 - [\varepsilon + d_S(y_\varepsilon(x,0), y_\varepsilon(x,T))]^2\} \\ &= \frac{\varepsilon + d_S(y_\varepsilon(x,0), y_\varepsilon(x,T))}{\varepsilon} \left[\int_\Omega a_\varepsilon z(x,0)dx + \int_\Omega b_\varepsilon z(x,T)dx\right], \end{aligned} \tag{3.23}$$

where $(a_\varepsilon, b_\varepsilon) \in d_S(y_\varepsilon(x,0), y_\varepsilon(x,T))$ satisfying

$$\|a_\varepsilon\|^2_{L^2(\Omega)} + \|b_\varepsilon\|^2_{L^2(\Omega)} = \begin{cases} 1 & \text{if } (y_\varepsilon(x,0), y_\varepsilon(x,T)) \notin S, \\ 0 & \text{if } (y_\varepsilon(x,0), y_\varepsilon(x,T)) \in S. \end{cases} \tag{3.24}$$

By (3.22) and (3.23), using the same argument as in the proof of Theorem 1.1, we yield that for all $(z,v) \in Y \times L^2(Q)$ and $\varepsilon > 0$,

$$\begin{aligned} 0 \le \lambda_\varepsilon \Big\{ & \int_Q \nabla g^\varepsilon(t, y_\varepsilon) z dxdt + \int_Q \nabla h_\varepsilon(u_\varepsilon) v dxdt + \int_Q (u_\varepsilon - u_2^*) v dxdt \\ & + \int_Q (y_\varepsilon - y_2^*)^{2\tilde{r}-1} z dxdt + \int_\Omega [y_\varepsilon(x,0) - y_2^*(x,0)] z(x,0) dx \\ & + \int_\Omega [y_\varepsilon(x,T) - y_2^*(x,T)] z(x,T) dx \Big\} + \int_\Omega a_\varepsilon z(x,0) dx + \int_\Omega b_\varepsilon z(x,T) dx \\ & + \int_Q p_\varepsilon \left(\frac{\partial z}{\partial t} + Az + f_y'(x,t,y_\varepsilon) z - v \right) dxdt, \end{aligned} \tag{3.25}$$

where

$$\lambda_\varepsilon = \frac{\varepsilon}{\varepsilon + d_S(y_\varepsilon(x,0), y_\varepsilon(x,T))}, \qquad \mu_\varepsilon = \frac{1}{\varepsilon + d_S(y_\varepsilon(x,0), y_\varepsilon(x,T))},$$

and

$$p_\varepsilon = \mu_\varepsilon \left(\frac{\partial y_\varepsilon}{\partial t} + Ay_\varepsilon + f(x,t,y_\varepsilon) - u_\varepsilon \right).$$

It is clear that $p_\varepsilon \in L^2(Q)$ and

$$1 \le \lambda_\varepsilon^2 + \|a_\varepsilon\|^2_{L^2(\Omega)} + \|b_\varepsilon\|^2_{L^2(\Omega)} \le 2. \tag{3.26}$$

By taking $z = 0$ in (3.25), we yield that

$$\lambda_\varepsilon \left\{ \int_Q \nabla h_\varepsilon(u_\varepsilon) v dxdt + \int_Q (u_\varepsilon - u_2^*) v dxdt \right\} - \int_Q p_\varepsilon v dxdt \le 0 \text{ for all } v \in L^2(Q),$$

which implies that

$$p_\varepsilon = \lambda_\varepsilon \nabla h_\varepsilon(u_\varepsilon) + \lambda_\varepsilon (u_\varepsilon - u_2^*) \text{ a.e. in } Q. \tag{3.27}$$

By (3.27) and by the same argument as in the proof of Theorem 1.1, we infer that $\{p_\varepsilon\}$ is bounded in $L^2(Q)$.

By taking $v = 0$ in (3.25), we obtain that

$$\begin{aligned} 0 \le & \int_Q \lambda_\varepsilon \nabla g^\varepsilon(t, y_\varepsilon) z dxdt + \int_\Omega a_\varepsilon z(x,0) dx + \int_\Omega b_\varepsilon z(x,T) dx \\ & + \int_Q p_\varepsilon \left[\frac{\partial z}{\partial t} + Az + f_y'(x,t,y_\varepsilon) z \right] dxdt + \lambda_\varepsilon \Big\{ \int_\Omega [y_\varepsilon(x,0) - y_2^*(x,0)] z(x,0) dx \\ & + \int_\Omega [y_\varepsilon(x,T) - y_2^*(x,T)] z(x,T) dx + \int_Q (y_\varepsilon - y_2^*)^{2\tilde{r}-1} z dxdt \Big\} \end{aligned} \tag{3.28}$$

for all $z \in Y$.

By (3.28) and by the same argument as in [14], we obtain the following:

(3.29)
$$\begin{cases} -\frac{\partial p_\varepsilon}{\partial t} + Ap_\varepsilon + f_y'(x,t,y_\varepsilon)p_\varepsilon + \lambda_\varepsilon \nabla g^\varepsilon(t,y_\varepsilon) + \lambda_\varepsilon (y_\varepsilon - y_2^*)^{2\tilde{r}-1} = 0, & \text{in } Q, \\ p_\varepsilon(x,t) = 0, & \text{on } \sum, \\ p_\varepsilon(x,T) = -b_\varepsilon - \lambda_\varepsilon[y_\varepsilon(x,T) - y_2^*(x,T)], & \text{in } \Omega, \\ p_\varepsilon(x,0) = a_\varepsilon + \lambda_\varepsilon[y_\varepsilon(x,0) - y_2^*(x,0)], & \text{in } \Omega. \end{cases}$$

Next we are going to pass to the limit for $\varepsilon \to 0$ in (3.27) and (3.28). To this end, we observe first that, by (3.26), there exists a generalized subsequence of ε, still denoted by itself, such that

$$(a_\varepsilon, b_\varepsilon) \to (a_0, b_0) \text{ weakly in } L^2(\Omega) \times L^2(\Omega) \text{ and } \lambda_\varepsilon \to \lambda_0 \text{ as } \varepsilon \to 0. \tag{3.30}$$

Thus, by (3.30), $(3.29)_3$, and $(3.29)_4$ and by Lemma 3.2, we find that

$$p_\varepsilon(x,T) \to -b_0,\ p_\varepsilon(x,0) \to a_0 \text{ weakly in } L^2(\Omega) \text{ as } \varepsilon \to 0. \tag{3.31}$$

On the other hand, by the same argument as in the proof of Theorem 1.1, we infer that

$$\|f_y'(x,t,y_\varepsilon)p_\varepsilon\|_{L^{\frac{2n}{n+2}}(Q)} \le C, \tag{3.32}$$

$$\lambda_\varepsilon (y_\varepsilon - y_2^*)^{2\tilde{r}-1} \to 0 \text{ strongly in } L^{\frac{2n}{n+2}}(Q) \text{ as } \varepsilon \to 0, \tag{3.33}$$

and

$$\lambda_\varepsilon \nabla g^\varepsilon(t,y_\varepsilon) \to \lambda_0 \alpha \text{ weakly in } L^2(0,T;L^2(\Omega)) \text{ as } \varepsilon \to 0, \tag{3.34}$$

where $\alpha \in \partial g(t, y_2^*)$ a.e. in Q.

By $(3.29)_1$, (3.32), (3.33), (3.34) and by Lemma 3.2, we infer that

$$\left\{ -\frac{\partial p_\varepsilon}{\partial t} + Ap_\varepsilon \right\} \text{ is bounded in } L^{\frac{2n}{n+2}}(Q). \tag{3.35}$$

By $(3.29)_2$, (3.31), and (3.35), using argument similar to those in [15], we obtain that $\{p_\varepsilon\}$ is bounded in $L^{\frac{2n}{n+2}}(0,T;W_0^{1,\frac{2n}{n+2}}(\Omega))$.

Then by the same argument as in [2], we find that there exists a generalized subsequence of ε, still denoted by itself, such that

$$\begin{aligned} p_\varepsilon \to p \quad & \text{weakly in } L^{\frac{2n}{n+2}}(0,T;W_0^{1,\frac{2n}{n+2}}(\Omega)) \cap L^2(Q), \\ & \text{strongly in } L^{\frac{2n}{n+2}}(Q) \text{ as } \varepsilon \to 0, \end{aligned} \tag{3.36}$$

$$p_\varepsilon(x,t) \to p(x,t) \text{ a.e. in } Q \text{ as } \varepsilon \to 0. \tag{3.37}$$

Here we have used the fact that $\{p_\varepsilon\}$ is bounded in $L^2(Q)$. Since f_y' is continuous, by Lemma 3.2 and (3.37), we obtain that

$$f_y'(x,t,y_\varepsilon(x,t))p_\varepsilon(x,t) \to f_y'(x,t,y_2^*(x,t))p(x,t) \text{ a.e. in } Q \text{ as } \varepsilon \to 0. \tag{3.38}$$

This together with (3.32) implies that

$$p_\varepsilon f_y'(x,t,y_\varepsilon) \to p f_y'(x,t,y_2^*) \text{ weakly in } L^{\frac{2n}{n+2}}(Q) \text{ as } \varepsilon \to 0,$$

and so we have that

$$\int_Q p_\varepsilon f_y'(x,t,y_\varepsilon)z\,dxdt \to \int_Q p f_y'(x,t,y_2^*)z\,dxdt \ \text{ as } \varepsilon \to 0, \tag{3.39}$$

because $z \in Y \subset L^{\frac{2n}{n-2}}(Q)$. Thus by (3.31), (3.33), (3.34), (3.36), and (3.39), we may take the limit for $\varepsilon \to 0$ in (3.28) to obtain that

$$\begin{aligned} 0 = &\int_Q \lambda_0 \alpha z\,dxdt + \int_\Omega a_0 z(x,0)dx + \int_\Omega b_0 z(x,T)dx \\ &+ \int_Q p\left(\frac{\partial z}{\partial t} + Az + f_y'(x,t,y_2^*)z\right)dxdt \ \text{ for all } z \in Y. \end{aligned} \tag{3.40}$$

By (3.40) and by the same argument as in [14], we infer that $p \in L^{\frac{2n}{n+2}}(0,T; W_0^{1,\frac{2n}{n+2}}(\Omega)) \cap L^2(Q)$ and satisfies (1.9) and (1.10). By (3.30), (3.36), and Lemma 3.2, we may pass to the limit for $\varepsilon \to 0$ in (3.27) to obtain (1.12).

On the other hand, since $(a_\varepsilon, b_\varepsilon) \in \partial d_S(y_\varepsilon(x,0), y_\varepsilon(x,T))$, we have that

$$\langle a_\varepsilon, x_0 - y_\varepsilon(x,0)\rangle + \langle b_\varepsilon, x_1 - y_\varepsilon(x,T)\rangle \le 0 \ \text{ for all } (x_0,x_1) \in S. \tag{3.41}$$

By (3.30) and by Lemma 3.2 again, we may take the limit for $\varepsilon \to 0$ in (3.41) to obtain (1.11).

It remains to show that $\lambda_0 \neq 0$. As a contradiction, we assume that $\lambda_0 = 0$. Then by (1.12), we get that $p = 0$. Because $\lambda_\varepsilon \to \lambda_0 = 0$, it follows from (3.26) that there exist $\delta > 0$ and $\varepsilon_1 > 0$ such that

$$2 \ge \|a_\varepsilon\|^2_{L^2(\Omega)} + \|b_\varepsilon\|^2_{L^2(\Omega)} \ge \delta > 0 \qquad \text{for all } \varepsilon < \varepsilon_1. \tag{3.42}$$

By (3.25) and (3.41), we deduce that

$$\begin{aligned} &\int_Q p_\varepsilon\left(\frac{\partial z}{\partial t} + Az + f_y'(x,t,y_\varepsilon)z - v\right)dxdt \\ &+ \int_\Omega a_\varepsilon[z(x,0) - x_0 + y_2^*(x,0)]dx + \int_\Omega b_\varepsilon[z(x,T) - x_1 + y_2^*(x,T)]dx \ge -\zeta_\varepsilon(z,v) \end{aligned} \tag{3.43}$$

for all $(z,v) \in Y \times L^2(Q)$, where

$$\begin{aligned} \zeta_\varepsilon(z,v) = \lambda_\varepsilon \Big\{ &\int_Q \nabla g^\varepsilon(t,y_\varepsilon)z\,dxdt + \int_Q \nabla h_\varepsilon(u_\varepsilon)v\,dxdt \\ &+ \int_Q (u_\varepsilon - u_2^*)v\,dxdt + \int_Q (y_\varepsilon - y_2^*)^{2\tilde{r}-1}z\,dxdt \\ &+ \int_\Omega [y_\varepsilon(x,0) - y_2^*(x,0)]z(x,0)dx + \int_\Omega [y_\varepsilon(x,T) - y_2^*(x,T)]z(x,T)dx \Big\} \\ &+ \int_\Omega a_\varepsilon[y_\varepsilon(x,0) - y_2^*(x,0)]dx + \int_\Omega b_\varepsilon[y_\varepsilon(x,T) - y_2^*(x,T)]dx. \end{aligned}$$

Observe that $\zeta_\varepsilon(z,v) \to 0$ as $\varepsilon \to 0$ uniformly in $(z,v) \in \{(z,v) \in Y \times L^2(Q) : \|z\|_Y \le r_2, \|v\|_{L^2(Q)} \le r_3\}$. By Lemma 3.2, there exists a positive constant, denoted by ε_1 again, such that $\|y_\varepsilon - y_2^*\|_Y \le r_1$ for all $\varepsilon < \varepsilon_1$. Then by (H_6), for any $(z_0, z_T) \in B_{r_1,r_2,r_3}$ and $\varepsilon < \varepsilon_1$, there exists $(z_\varepsilon, v_\varepsilon) \in Y \times L^2(Q)$ with $\|z_\varepsilon\|_Y \le r_2$, $\|v_\varepsilon\|_{L^2(Q)} \le r_3$ such that $z_\varepsilon(x,0) = z_0(0)$, $z_\varepsilon(x,0) = z_1(x)$, and

$$\frac{\partial z_\varepsilon}{\partial t} + Az_\varepsilon + f_y'(x,t,y_\varepsilon)z_\varepsilon = v_\varepsilon \ \text{ in } Q.$$

By taking $z = z_\varepsilon$ and $v = v_\varepsilon$ for $\varepsilon < \varepsilon_1$ in (3.43), we obtain that

$$\langle a_\varepsilon, z_0 - x_0 + y_2^*(x,0)\rangle + \langle b_\varepsilon, z_1 - x_1 + y_2^*(x,T)\rangle \geq -\zeta_\varepsilon(z_\varepsilon, v_\varepsilon) \tag{3.44}$$

and

$$\zeta_\varepsilon(z_\varepsilon, v_\varepsilon) \to 0 \ \text{ as } \varepsilon \to 0, \ \text{ uniformly in } (z_0, z_T) \in B_{r_1,r_2,r_3}. \tag{3.45}$$

Thanks to Lemma 3.6 of Chapter 4 of [13], we conclude by (3.42), (3.44), and (3.45) that $(a_0, b_0) \neq 0$, which yields that $p \neq 0$. This contradiction shows that $\lambda_0 \neq 0$ and completes the proof. □

4. Existence of optimal pairs for (P_1) and (P_2). In this section we shall prove Theorem 1.3 and Theorem 1.4.

Proof of Theorem 1.3. Since $D_{ad} \neq \emptyset$, we may set $d = \mathrm{Inf}_{(y,u)\in D_{ad}} L(y,u)$. It is clear that $d > -\infty$. We assume that $(y_m, u_m) \in D_{ad}$ satisfies

$$d \leq L(y_m, u_m) < d + \frac{1}{m}. \tag{4.1}$$

This, together with (H_1) and (H_7), implies that $\{u_m\}$ is bounded in $L^2(Q)$ and that $\{y_m\}$ is bounded in $L^{2\tilde{r}}(Q)$. By (H_2), one can easily check that $\{f(x,t,y_m)\}$ is bounded in $L^2(Q)$. Hence

$$\left\{\frac{\partial y_m}{\partial t} + Ay_m\right\} \ \text{ is bounded in } \ L^2(Q). \tag{4.2}$$

Since $y_m(x,t) = 0$ on $\sum$ and $y_m(x,0) = y_0(x)$ in Ω, it follows from (4.2) that $\{y_m\}$ is bounded in Y (cf. [15]). Thus we may extract a subsequence of $\{y_m\}$, denoted by itself again, such that

$$\begin{array}{ll} y_m \to y^* & \text{weakly in } Y, \ \text{strongly in } L^2(Q) \ \text{ as } m \to \infty, \\ u_m \to u^* & \text{weakly in } L^2(Q) \ \text{ as } m \to \infty. \end{array} \tag{4.3}$$

By the same argument as in the proof of Lemma 2.1, we have that

$$f(x,t,y_m) \to f(x,t,y^*) \ \text{ weakly in } L^2(Q) \ \text{ as } m \to \infty. \tag{4.4}$$

It follows from (4.3) and (4.4) that (y^*, u^*) satisfies (1.1). By (4.3) and by (H_3), we obtain that

$$F(y_m) \to F(y^*) \ \text{ strongly in } X \ \text{ as } m \to \infty,$$

which implies that $F(y^*) \in W$, because W is closed. Thus we have

$$(y^*, u^*) \in D_{ad}. \tag{4.5}$$

On the other hand, by (4.3) and by the same argument as in [2], we get that

$$\varliminf_{m\to\infty} \int_0^T [g(t,y_m) + h(u_m)]dt \geq \int_0^T [g(t,y^*) + h(u^*)]dt,$$

which, combined with (4.1), indicates that

$$L(y^*, u^*) = d. \tag{4.6}$$

By (4.5) and (4.6), we obtain that (y^*, u^*) is optimal for problem (P_1). This completes the proof.

Proof of Theorem 1.4. We may set $d = \mathrm{Inf}_{(y,u)\in\tilde{D}_{ad}} L(y,u)$ because $\tilde{D}_{ad} \neq \emptyset$. It is clear that $d > -\infty$. Let $(y_m, u_m) \in \tilde{D}_{ad}$ be such that

$$d \leq L(y_m, u_m) < d + \frac{1}{m}. \tag{4.7}$$

By (H_1), (H_7), and (H_{10}), we have that

$$\begin{array}{l} \{u_m\} \text{ is bounded in } L^2(Q), \\ \{y_m\} \text{ is bounded in } L^{2\tilde{r}}(Q), \\ \{y_m(x,0)\} \text{ is bounded in } L^2(\Omega). \end{array} \tag{4.8}$$

By (4.8) and by the same argument as in the proof of Theorem 1.3, we infer that $\{\frac{\partial y_m}{\partial t} + Ay_m\}$ is bounded in $L^2(Q)$, which together with (4.8) implies that $\{y_m\}$ is bounded in Y (cf. [15]).

Thus there exist a pair $(y^*, u^*) \in Y \times L^2(Q)$ and subsequences of $\{y_m\}$ and $\{u_m\}$, still denoted by themselves, such that

$$y_m \to y^* \text{ weakly in } Y, \text{ strongly in } L^2(Q) \text{ as } m \to \infty, \tag{4.9}$$

$$u_m \to u^* \text{ weakly in } L^2(Q) \text{ as } m \to \infty, \tag{4.10}$$

$$(y_m(x,0), y_m(x,T)) \to (y^*(x,0), y^*(x,T)) \text{ weakly in } L^2(\Omega) \times L^2(\Omega) \text{ as } m \to \infty. \tag{4.11}$$

By (4.9), (4.10), and (4.11), one can easily check that (y^*, u^*) satisfies (1.3).

On the other hand, since S is convex and closed, d_S is lower semicontinuous in the weak topology of $L^2(\Omega) \times L^2(\Omega)$. Thus it follows from (4.11) that

$$0 = \varliminf_{m\to\infty} d_S(y_m(x,0), y_m(x,T)) \geq d_S(y^*(x,0), y^*(x,T)),$$

which yields that $(y^*(x,0), y^*(x,T)) \in S$. Thus we get that

$$(y^*, u^*) \in \tilde{D}_{ad}. \tag{4.12}$$

Now by (4.12) and by the same argument as in the proof of Theorem 1.3, we infer that (y^*, u^*) is optimal for problem (P_2). This completes the proof. □

REFERENCES

[1] V. Barbu, *Analysis and Control of Nonlinear Infinite Dimensional System*, Academic Press, New York, 1993.

[2] V. Barbu, *Optimal Control of Variational Inequalities*, Res. Notes Math. 100, Pitman, Boston, London, Melbourne, 1984.

[3] V. Barbu, *Exact controllability of the superlinear heat equation*. Appl. Math. Optim., 42 (2000), pp. 73–89.

[4] V. Barbu and N. Pavel, *Optimal control problems with two point boundary conditions*, J. Optim. Theory Appl., 77 (1993), pp. 51–78.

[5] V. Barbu and N. Pavel, *Periodic optimal control in Hilbert space*, Appl. Math. Optim., 33 (1996), pp. 169–188.

[6] M. Bergounioux and F. Troltzsch, *Optimal conditions and generalized bang-bang principle for a state constrained semilinear parabolic problem*, Numer. Funct. Anal. Optim., 15 (1996), pp. 517–537.

[7] E. Casas, *Pontryagin's principle for state-constrained boundary control problems of semilinear parabolic equations*, SIAM, J. Control Optim., 35 (1997), pp. 1297–1327.
[8] E. Casas and J. Yong, *Maximum principle for state-constrained optimal control problems governed by quasilinear equations*, Differential Integral Equations, 130 (1996), pp. 179–200.
[9] F. H. Clarke, *Optimization and Non smooth Analysis*, John Wiley, New York, 1983.
[10] H. Fattorini and H. Frankowska, *Necessary conditions for infinite dimensional control problems*, Math. Control Signals Systems, 4 (1991), pp. 225–257.
[11] H. O. Fattorini and T. Murphy, *Optimal problems for nonlinear parabolic boundary control systems*, SIAM J. Control Optim., 32 (1994), pp. 1577–1596.
[12] X. Li and J. Yong, *Necessary conditions for optimal control of distributed parameter systems*, SIAM J. Control Optim., 29 (1991), pp. 895–908.
[13] X. Li and J. Yong, *Optimal Control Theory for Infinite Dimensional Systems*, Birkhäuser Boston, Cambridge, MA, 1995.
[14] J. L. Lions, *Some Methods in the Mathematical Analysis of Systems and Their Control*, Science Press, Beijing, China, Gordon and Breach, Science Publishers, New York, 1981.
[15] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin, Heidelberg, New York, 1971.
[16] G. Wang and S. Chen, *Maximum principle for optimal control of some parabolic systems with two point boundary conditions*, Numer. Funct. Anal. Optim., 20 (1999), pp. 163–174.
[17] G. Wang, *Optimal control of parabolic differential equations with two point boundary state constraints*, SIAM J. Control Optim., 38 (2000), pp. 1639–1654.
[18] G. Wang, *Optimal control of parabolic variational inequality involving state constraint*, Nonlinear Anal., 42 (2000), pp. 789–801.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 5, pp. 1540–1555

DYNAMIC MEAN-VARIANCE PORTFOLIO SELECTION WITH NO-SHORTING CONSTRAINTS*

XUN LI†, XUN YU ZHOU†, AND ANDREW E. B. LIM‡

Abstract. This paper is concerned with mean-variance portfolio selection problems in continuous-time under the constraint that short-selling of stocks is prohibited. The problem is formulated as a stochastic optimal linear-quadratic (LQ) control problem. However, this LQ problem is *not* a conventional one in that the control (portfolio) is constrained to take nonnegative values due to the no-shorting restriction, and thereby the usual Riccati equation approach (involving a "completion of squares") does not apply directly. In addition, the corresponding Hamilton–Jacobi–Bellman (HJB) equation inherently has no smooth solution. To tackle these difficulties, a continuous function is constructed via two Riccati equations, and then it is shown that this function is a viscosity solution to the HJB equation. Solving these Riccati equations enables one to explicitly obtain the efficient frontier and efficient investment strategies for the original mean-variance problem. An example illustrating these results is also presented.

Key words. continuous-time, mean-variance portfolio selection, short-selling prohibition, efficient frontier, stochastic LQ control, HJB equation, viscosity solution

AMS subject classifications. 91B28, 93E20

PII. S0363012900378504

1. Introduction.

Research on portfolio selection dates back to the 1950s with Markowitz's pioneering work [24] on mean-variance efficient portfolios for a single-period investment. The most important contribution of Markowitz's work is the introduction of quantitative and scientific approaches to risk management and analysis. When short-selling of stocks is not allowed, efficient portfolios are obtained computationally via solving a quadratic programming problem. Later, Merton [26] derived an analytical solution to the single-period mean-variance problem under the assumption that the covariance matrix is positive definite and short-selling is allowed.

While it is natural to extend Markowitz's work to multiperiod and continuous-time portfolio selections, these extensions have, by and large, taken a somewhat different tack to Markowitz's original formulation; see, e.g., [1, 10, 14, 27, 28] for the multiperiod case and [4, 7, 8, 13, 17, 25] for the continuous-time case. Specifically, rather than treating the Var $X(T)$ and $EX(T)$ of a portfolio as separate quantities and finding the relationship between them, a single quantity, the expected utility of terminal wealth $EU(X(T))$, is considered instead. The utility function U commonly has a power, log, exponential, or quadratic form. One disadvantage of this approach is that the relationship between risk and return is contained only implicitly in the utility function. Hence, it is less clear in general what relationship exists between the risk and the return of the derived policy. It should be noted that mean-variance analysis and expected utility formulation are two *different* tools for dealing with portfolio selections. As a consequence, optimal portfolios determined by utility functions

*Received by the editors September 20, 2000; accepted for publication (in revised form) June 12, 2001; published electronically January 18, 2002. This research was partially supported by the RGC Earmarked Grants CUHK 4054/98E and CUHK 4435/99E.

http://www.siam.org/journals/sicon/40-5/37850.html

†Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, People's Republic of China (xli@se.cuhk.edu.hk, xyzhou@se.cuhk.edu.hk).

‡Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027 (lim@ieor.columbia.edu).

are usually not mean-variance efficient. One exception is the case of the quadratic utility function; see Duffie and Richardson [7], where this relationship is shown in the setting of the related mean-variance hedging problem. For comparisons of the performance of the mean-variance versus utility approaches, the reader is referred to [11, 12, 14, 31, 36].

One difficulty in extending Markowitz's idea to the multiperiod or continuous-time settings is that the variance Var $X(T)$ involves a term $[EX(T)]^2$ that is hard to analyze due to its nonseparability in the sense of dynamic programming; see [37, p. 20] for a more detailed discussion on this point. Only recently have Li and Ng [20] faithfully extended Markowitz's mean-variance model to the multiperiod setting by using the idea of embedding the problem in a tractable auxiliary problem.

In the paper by Zhou and Li [37], the continuous-time mean-variance problem in which short-selling of stocks is allowed is studied by incorporating the embedding technique used in Li and Ng [20]. However, the main contribution of [37] is not the explicit mean-variance efficient frontier it obtained per se; rather it is the unifying framework, namely, that of the stochastic linear-quadratic (LQ) optimal control, that it introduced in order to solve certain finance problems including mean-variance portfolio selection. The so-called indefinite stochastic LQ control theory has been developed extensively in recent years (see, e.g., [2, 3, 21, 34]), and this in turn provides a powerful tool for solving some finance problems that are linear-quadratic in nature [19, 22, 37].

The objective of this paper is to investigate continuous-time mean-variance portfolio selection in the case where short-selling of stocks is not allowed. (However, shorting the riskless asset—the bond—is still allowed.) This belongs to the realm of so-called constrained portfolio selection, which essentially renders the market incomplete. In the past decade, the constrained portfolio selection problem has been extensively studied (see, e.g., [6, 16, 30, 32, 33]). However, again the expected utility model has been mainly adopted. In particular, Xu and Shreve in their two-part paper [32, 33] investigated a utility maximization problem with a no-shorting constraint using a duality analysis. In [6, 18], the duality results of [32, 33] are extended to a general class of portfolio selection problems in incomplete markets, including those with constraints. The main results in [6, 18] establish the existence of a solution to the dual problem and show how it can be used to construct a solution of the original portfolio optimization problem. One important difference between the approach we adopt and the duality methods in [6, 18] is that existence and optimality of the candidate portfolio in this paper are established using the theory of *viscosity solutions* and the viscosity verification theorem in [38]. This enables us to sidestep the considerable technicalities encountered in [6, 18] when studying existence through convex duality.

In this paper we continue to use stochastic LQ control as the framework for studying the constrained mean-variance portfolio problem. Compared with [22, 37], the distinctive feature of this paper is that shorting is prohibited. As a consequence, a major difficulty in the present case is that the control (portfolio) is constrained, while the LQ theory typically requires the control to be unconstrained (the reason is that the optimal control constructed through the Riccati equation may not satisfy the control constraint). This means that the elegant Riccati approach does not apply directly. We sidestep this problem by studying the Hamilton–Jacobi–Bellman (HJB) equation. (Recall that the Riccati equation *is* essentially the HJB equation after separating the time and spatial variables.) However, the HJB equation has no classical (i.e., smooth) solutions in our case due to the presence of the control constraint. To cope

with this difficulty, we first conjecture a continuous solution to the HJB equation via *two* Riccati equations, and then show that it is indeed the *viscosity solution* to the equation. Further, using the viscosity verification theorem established in [38], we explicitly obtain the optimal strategy along with the efficient frontier.

The outline of this paper is as follows. In section 2, we formulate the mean-variance portfolio problem under a short-selling prohibition. In section 3, we study a stochastic LQ control problem of which portfolio selection is a special case, and we obtain the viscosity solution to the corresponding HJB equation along with the optimal feedback control. Section 4 is devoted to the derivation of the efficient investment strategies and efficient frontier for the portfolio selection problem. In section 5, we present a numerical example to illustrate the results obtained. Finally, section 6 concludes the paper.

2. Problem formulation and preliminaries.

2.1. Notation.

We make use of the following notation:

M'	:	the transpose of any matrix or vector M;
$\|M\|$	:	$\sqrt{\sum_{i,j} m_{ij}^2}$ for any matrix or vector $M = (m_{ij})$;
$\mathbb{R}^n$	:	n-dimensional real Euclidean space;
$\mathbb{R}^n_+$	:	the subset of $\mathbb{R}^n$ consisting of elements with nonnegative components.

The underlying uncertainty is generated by a fixed filtered complete probability space $(\Omega, \mathcal{F}, \mathbf{P}, \{\mathcal{F}_t\}_{t\geq 0})$ on which is defined a standard $\{\mathcal{F}_t\}_{t\geq 0}$-adapted m-dimensional Brownian motion $W(t) \equiv (W^1(t), \ldots, W^m(t))'$. Given a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ with a filtration $\{\mathcal{F}_t | a \leq t \leq b\}(-\infty \leq a < b \leq +\infty)$, and a Hilbert space $\mathcal{H}$ with the norm $\|\cdot\|_{\mathcal{H}}$, define the Banach space

$$L^2_{\mathcal{F}}(0,T;\mathcal{H}) = \left\{ \varphi(\cdot) \,\middle|\, \begin{array}{l} \varphi(\cdot) \text{ is an } \mathcal{F}_t\text{-adapted, } \mathcal{H}\text{-valued measurable process on } [a,b] \\ \text{and } E\int_a^b \|\varphi(t,\omega)\|^2_{\mathcal{H}} dt < +\infty \end{array} \right\}$$

with the norm

$$\|\varphi(\cdot)\|_{\mathcal{F},2} = \left(E\int_a^b \|\varphi(t,\omega)\|^2_{\mathcal{H}} dt \right)^{\frac{1}{2}} < +\infty.$$

2.2. Problem formulation.

We consider a financial market where $m+1$ assets are traded continuously on a finite horizon $[0,T]$. One asset is a bond, whose price $\mathbb{P}_0(t)$, $t \geq 0$, evolves according to the differential equation

$$\left\{ \begin{array}{rcl} d\mathbb{P}_0(t) &=& r(t)\mathbb{P}_0(t)dt, \quad t \in [0,T], \\ \mathbb{P}_0(0) &=& p_0 > 0, \end{array} \right. \tag{2.1}$$

where $r(t)$ (> 0) is the interest rate of the bond. The remaining m assets are stocks, and their prices are modeled by the stochastic differential equations

$$\left\{ \begin{array}{rcl} d\mathbb{P}_i(t) &=& \mathbb{P}_i(t)\{b_i(t)dt + \sum_{j=1}^m \sigma_{ij}(t)W^j(t)\}, \quad t \in [0,T], \\ \mathbb{P}_i(0) &=& p_i > 0, \end{array} \right. \tag{2.2}$$

where $b_i(t)(> r(t))$ is the appreciation rate and $\sigma_{ij}(t)$ is the volatility coefficient. Denote $b(t) := (b_1(t), \ldots, b_m(t))'$ and $\sigma(t) := (\sigma_{ij}(t))$. We assume throughout that $r(t)$, $b(t)$, and $\sigma(t)$ are deterministic, Borel-measurable, and bounded on $[0,T]$. In addition, we assume that the nondegeneracy condition

$$\sigma(t)\sigma(t)' \geq \delta I \quad \forall t \in [0,T], \tag{2.3}$$

where $\delta > 0$ is a given constant, is satisfied. Also, we define the relative risk coefficient

$$\theta(t) \triangleq \sigma^{-1}(t)(b(t) - r(t)\mathbf{1}), \tag{2.4}$$

where $\mathbf{1}$ is the m-dimensional column vector with each component equal to 1.

Suppose an agent has an initial wealth $X_0 > 0$, and the total wealth of his position at time $t \geq 0$ is $X(t)$. Then it is well-known that $X(t),\ t \geq 0$, follows (see, e.g., [37])

$$\begin{cases} dX(t) = \{r(t)X(t) + \sum_{i=1}^{m}(b_i(t) - r(t))u_i(t)\}dt + \sum_{j=1}^{m}\sum_{i=1}^{m}\sigma_{ij}(t)u_i(t)dW^j(t), \\ X(0) = X_0, \end{cases} \tag{2.5}$$

where $u_i(t),\ i = 0, 1, \ldots, m$, denotes the total market value of the agent's wealth in the ith bond/stock. We call $u(t) := (u_1(t), \ldots, u_m(t))$ the portfolio (which changes over time t). An important restriction considered in this paper is the prohibition of short-selling the stocks, i.e., it must be satisfied that $u_i(t) \geq 0,\ i = 1, \ldots, m$. On the other hand, borrowing from the money market (at the interest rate $r(t)$) is still allowed; that is, $u_0(t)$ is not explicitly constrained.

Mean-variance portfolio selection refers to the problem of finding an allowable investment policy (i.e., a dynamic portfolio satisfying all the constraints) such that the expected terminal wealth satisfies $EX(T) = d$ while the risk measured by the variance of the terminal wealth

$$\text{Var } X(T) = E[X(T) - EX(T)]^2 = E[X(T) - d]^2$$

is minimized.

We impose throughout this paper the following assumption.

Assumption 2.1. The value of the expected terminal wealth d satisfies $d \geq X_0 e^{\int_0^T r(s)ds}$.

Remark 2.1. Assumption 2.1 states that the investor's expected terminal wealth d cannot be less than $X_0 e^{\int_0^T r(s)ds}$, which coincides with the amount that he/she would earn if all of the initial wealth were invested in the bond for the entire investment period. Clearly, this is a reasonable assumption, for the solution of the problem under $d < X_0 e^{\int_0^T r(s)ds}$ is foolish for rational investors.

DEFINITION 2.1. *A portfolio $u(\cdot)$ is said to be admissible if $u(\cdot) \in L^2_{\mathcal{F}}(0, T; \mathbb{R}^m_+)$.*

DEFINITION 2.2. *The mean-variance portfolio selection problem is formulated as the following optimization problem parameterized by $d \geq X_0 e^{\int_0^T r(s)ds}$:*

$$\begin{array}{ll} \min & \text{Var } X(T) \equiv E[X(T) - d]^2, \\ \text{subject to} & \begin{cases} EX(T) = d, \\ u(\cdot) \in L^2_{\mathcal{F}}(0, T; \mathbb{R}^m_+), \\ (X(\cdot), u(\cdot)) \text{ satisfy (2.5)}. \end{cases} \end{array} \tag{2.6}$$

Moreover, the optimal control of (2.6) *is called an efficient strategy, and* $(\text{Var } X(T), d)$, *where* Var $X(T)$ *is the optimal value of* (2.6) *corresponding to d, is called an efficient point. The set of all efficient points, when the parameter d runs over $[X_0 e^{\int_0^T r(s)ds}, +\infty)$, is called the efficient frontier.*

Since (2.6) is a convex optimization problem, the equality constraint $EX(T) = d$ can be dealt with by introducing a Lagrange multiplier $\mu \in \mathbb{R}$. In this way the portfolio

problem (2.6) can be solved via the following optimal stochastic control problem (for every fixed μ):

$$\begin{array}{ll} \min & E\Big\{[X(T)-d]^2+2\mu[EX(T)-d]\Big\}, \\ \text{subject to} & \left\{\begin{array}{l} u(\cdot)\in L^2_{\mathcal{F}}(0,T;\mathbb{R}^m_+), \\ (X(\cdot),u(\cdot)) \text{ satisfy (2.5)}, \end{array}\right. \end{array} \tag{2.7}$$

where the factor 2 in front of the multiplier μ is introduced in the objective function just for convenience. Clearly, this problem is equivalent to

$$(A(\mu)): \quad \begin{array}{ll} \min & E\Big\{\frac{1}{2}[X(T)-(d-\mu)]^2\Big\}, \\ \text{subject to} & \left\{\begin{array}{l} u(\cdot)\in L^2_{\mathcal{F}}(0,T;\mathbb{R}^m_+), \\ (X(\cdot),u(\cdot)) \text{ satisfy (2.5)}, \end{array}\right. \end{array}$$

in the sense that the two problems have exactly the same optimal control.

3. A general constrained stochastic LQ problem. The problem $A(\mu)$ formulated in the previous section is a stochastic optimal LQ control problem. This problem has two features which distinguish it from conventional LQ problems. One is that the running cost of this problem is identically zero; that is, it is an *indefinite* stochastic LQ control problem, the theory of which has been developed extensively in recent years (see, for example, [2, 3, 21, 34, 35]). The other feature, which also gives rise to the main difficulty of the problem, is that the control is constrained. Therefore, the conventional "completion of squares" approach to the unconstrained LQ problem, which involves the Riccati equation, will no longer apply. In this section, we solve a class of constrained, indefinite stochastic LQ problems of which $A(\mu)$ is a special case.

Consider the controlled linear stochastic differential equation

$$\left\{\begin{array}{rcl} dx(t) &=& [A(t)x(t)+B(t)u(t)+f(t)]ds+\sum_{j=1}^m D_j(t)u(t)dW^j(t), \quad t\in[s,T], \\ x(s) &=& y\in\mathbb{R}, \end{array}\right. \tag{3.1}$$

where $A(t)$ and $f(t)\in\mathbb{R}$ are scalars, $B(t)'\in\mathbb{R}^m_+$ and $D_j(t)'\in\mathbb{R}^m$ $(j=1,\dots,m)$ are column vectors. In addition, we assume that the matrix $\sum_{j=1}^m D_j(t)'D_j(t)$ is nonsingular.

The class of admissible controls associated with (3.1) is the set $\mathcal{U}[s,T]=L^2_{\mathcal{F}}(s,T;\mathbb{R}^m_+)$. Given $u(\cdot)\in\mathcal{U}[s,T]$, the pair $(x(\cdot),u(\cdot))$ is referred to as an admissible pair if $x(\cdot)\in L^2_{\mathcal{F}}(s,T;\mathbb{R})$ is a solution of the stochastic differential equation (3.1) associated with $u(\cdot)\in\mathcal{U}[s,T]$. Our objective is to find an optimal $u(\cdot)$ that minimizes the quadratic (terminal) cost function

$$J(s,y;u(\cdot))=E\left\{\tfrac{1}{2}x(T)^2\right\}. \tag{3.2}$$

The value function associated with the LQ problem (3.1)–(3.2) is defined by

$$V(s,y)=\inf_{u(\cdot)\in\mathcal{U}[s,T]} J(s,y;u(\cdot)). \tag{3.3}$$

3.1. HJB equation. Since the Riccati equation approach is not applicable in this case, we study the corresponding HJB equation instead, which is the following partial differential equation:

$$\left\{\begin{array}{l} v_t(t,x)+\inf_{u\geq 0}\left\{v_x(t,x)[A(t)x+B(t)u+f(t)]+\frac{1}{2}v_{xx}(t,x)u'D(t)'D(t)u\right\}=0, \\ v(T,x)=\frac{1}{2}x^2, \end{array}\right. \tag{3.4}$$

where $D(t)' = (D_1(t)', \ldots, D_m(t)')$. Unfortunately, owing essentially to the nonnegativity constraint of the control, the HJB equation does not have a smooth solution, as opposed to the unconstrained case in which the solution to the HJB equation is a quadratic function which can be constructed via the Riccati equation. The idea here is to construct a function, show that it is a *viscosity* solution (see the appendix for the definition) to the HJB equation, and then employ the verification theorem to construct the optimal control.

Before we start, we recall some results from convex analysis.

LEMMA 3.1. *Let s be a continuous, strictly convex quadratic function*

$$s(z) \triangleq \tfrac{1}{2}\|(\mathcal{D}')^{-1}z + (\mathcal{D}')^{-1}\mathcal{B}'\|^2 \tag{3.5}$$

over $z \in [0,\infty)^m$, where $\mathcal{B}' \in \mathbb{R}^m_+$, $\mathcal{D} \in \mathbb{R}^{m\times m}$, and $\mathcal{D}'\mathcal{D} > 0$. Then s has a unique minimizer $\bar{z} \in [0,\infty)^m$, i.e.,

$$\|(\mathcal{D}')^{-1}\bar{z} + (\mathcal{D}')^{-1}\mathcal{B}'\|^2 \le \|(\mathcal{D}')^{-1}z + (\mathcal{D}')^{-1}\mathcal{B}'\|^2 \quad \forall z \in [0,\infty)^m. \tag{3.6}$$

The Kuhn–Tucker conditions for the minimization of s in (3.5) over $[0,\infty)^m$ lead to the Lagrange multiplier vector $\bar{\nu} \in [0,\infty)^m$ such that $\bar{\nu} = \nabla s(\bar{z}) = (\mathcal{D}'\mathcal{D})^{-1}\bar{z} + (\mathcal{D}'\mathcal{D})^{-1}\mathcal{B}'$ and $\bar{\nu}'\bar{z} = 0$.

LEMMA 3.2. *Let h be a continuous, strictly convex quadratic function*

$$h(z) \triangleq \tfrac{1}{2}z'\mathcal{D}'\mathcal{D}z - \alpha\mathcal{B}z \tag{3.7}$$

over $z \in [0,\infty)^m$, where $\mathcal{B}' \in \mathbb{R}^m_+$, $\mathcal{D} \in \mathbb{R}^{m\times m}$, and $\mathcal{D}'\mathcal{D} > 0$.

(i) *For every $\alpha \ge 0$, h has the unique minimizer $\alpha D^{-1}\bar{\xi} \in [0,\infty)^m$, where $\bar{\xi} = (\mathcal{D}')^{-1}\bar{z} + (\mathcal{D}')^{-1}\mathcal{B}'$. Here $\bar{z}$ is the minimizer of $s(z)$ specified in Lemma* 3.1. *Furthermore, $\bar{z}'\mathcal{D}^{-1}\bar{\xi} = 0$ and*

$$h(\alpha\bar{\nu}) = h(\alpha\mathcal{D}^{-1}\bar{\xi}) = -\tfrac{1}{2}\alpha^2\|\bar{\xi}\|^2. \tag{3.8}$$

(ii) *For every $\alpha < 0$, h has the unique minimizer* 0.

Lemma 3.1 and Lemma 3.2(i) are proved in section 5.2 and Lemma 3.2 of [33], while Lemma 3.2(ii) is obvious.

Remark 3.1. It should be noted that the vector $\bar{\xi}$ is independent of the parameter α.

Now let us come back to the LQ problem (3.1)–(3.2). Let

$$\bar{z}(t) := \arg\min_{z(t)\in[0,\infty)^m} \tfrac{1}{2}\|(D(t)')^{-1}z(t) + (D(t)')^{-1}B(t)'\|^2 \tag{3.9}$$

and

$$\bar{\xi}(t) := (D(t)')^{-1}\bar{z}(t) + (D(t)')^{-1}B(t)'. \tag{3.10}$$

Note that $\bar{\xi}(t)$ is a column vector *independent* of x. Let $\overline{P}(t)$, $\overline{g}(t)$, and $\overline{c}(t)$, respectively, denote the solutions of the following differential equations (the first being a special Riccati equation)

$$\begin{cases} \dot{\overline{P}}(t) = [-2A(t) + \|\bar{\xi}(t)\|^2]\overline{P}(t), \\ \overline{P}(T) = 1, \\ \overline{P}(t) > 0 \qquad \forall t \in [0,T], \end{cases} \tag{3.11}$$

$$\begin{cases} \dot{\overline{g}}(t) = [-A(t) + \|\bar{\xi}(t)\|^2]\overline{g}(t) - f(t)\overline{P}(t), \\ \overline{g}(T) = 0, \end{cases} \tag{3.12}$$

$$\begin{cases} \dot{\overline{c}}(t) = -f(t)\overline{g}(t) + \tfrac{1}{2}\|\bar{\xi}(t)\|^2\overline{P}(t)^{-1}\overline{g}(t)^2, \\ \overline{c}(T) = 0, \end{cases} \tag{3.13}$$

and $\widetilde{P}(t)$, $\widetilde{g}(t)$, and $\widetilde{c}(t)$, respectively, denote the solutions of the following differential equations (the first being *another* special Riccati equation)

$$\begin{cases} \dot{\widetilde{P}}(t) = -2A(t)\widetilde{P}(t), \\ \widetilde{P}(T) = 1, \\ \widetilde{P}(t) > 0 \qquad \forall t \in [0,T], \end{cases} \tag{3.14}$$

$$\begin{cases} \dot{\widetilde{g}}(t) = -A(t)\widetilde{g}(t) - f(t)\widetilde{P}(t), \\ \widetilde{g}(T) = 0, \end{cases} \tag{3.15}$$

$$\begin{cases} \dot{\widetilde{c}}(t) = -f(t)\widetilde{g}(t), \\ \widetilde{c}(T) = 0. \end{cases} \tag{3.16}$$

In the next subsection, we shall show that

$$V(t,x) = \begin{cases} \frac{1}{2}\overline{P}(t)x^2 + \overline{g}(t)x + \overline{c}(t) & \text{if } x + e^{-\int_t^T A(s)ds}\displaystyle\int_t^T f(z)e^{\int_z^T A(s)ds}dz \le 0, \\ \frac{1}{2}\widetilde{P}(t)x^2 + \widetilde{g}(t)x + \widetilde{c}(t) & \text{if } x + e^{-\int_t^T A(s)ds}\displaystyle\int_t^T f(z)e^{\int_z^T A(s)ds}dz > 0 \end{cases} \tag{3.17}$$

is a viscosity solution of the HJB equation (3.4), and

$$u^*(t,x) = \begin{cases} -D(t)^{-1}\bar{\xi}(t)\left(x + e^{-\int_t^T A(s)ds}\displaystyle\int_t^T f(z)e^{\int_z^T A(s)ds}dz\right) \\ \qquad\qquad \text{if } x + e^{-\int_t^T A(s)ds}\displaystyle\int_t^T f(z)e^{\int_z^T A(s)ds}dz \le 0, \\ 0 \qquad\qquad \text{if } x + e^{-\int_t^T A(s)ds}\displaystyle\int_t^T f(z)e^{\int_z^T A(s)ds}dz > 0 \end{cases} \tag{3.18}$$

is the associated optimal feedback control.

Remark 3.2. Equations (3.11)–(3.13) appear naturally in stochastic LQ problems with nonhomogeneous terms in the dynamics. They can be derived by conjecturing the value function to be a quadratic function (as in (3.17)), plugging in the HJB equation (3.4), and then comparing the terms of x^2, x, and the constant, respectively. See [35, pp. 317–318] for a detailed derivation.

3.2. Value function and optimal control. This subsection is devoted to verifying the aforementioned results. First we show that V constructed in (3.17) is a viscosity solution to the HJB equation (3.4).

We start with (3.11). Clearly

$$\overline{P}(t) = e^{\int_t^T (2A(s) - \|\bar{\xi}(s)\|^2)ds} \tag{3.19}$$

is the solution of (3.11). Note, in particular, that the constraint $\overline{P}(t) > 0$ is automatically satisfied. Defining $\overline{\eta}(t) := \frac{\overline{g}(t)}{\overline{P}(t)}$, it follows from (3.11) and (3.12) that

$$\dot{\overline{\eta}}(t) = \frac{\overline{P}(t)\dot{\overline{g}}(t) - \dot{\overline{P}}(t)\overline{g}(t)}{\overline{P}(t)^2} = \frac{A(t)\overline{P}(t)\overline{g}(t) - f(t)\overline{P}(t)^2}{\overline{P}(t)^2} = A(t)\overline{\eta}(t) - f(t).$$

Solving this equation with $\overline{\eta}(T) = 0$ yields

$$\overline{\eta}(t) = e^{-\int_t^T A(s)ds}\int_t^T f(z)e^{\int_z^T A(s)ds}dz. \tag{3.20}$$

Hence,

$$\overline{g}(t) = \overline{P}(t)\overline{\eta}(t) = e^{\int_t^T (A(s) - \|\bar{\xi}(s)\|^2) ds} \int_t^T f(z) e^{\int_z^T A(s) ds} dz.$$

Substituting these expressions into (3.13), we obtain

$$\begin{aligned} \dot{\bar{c}}(t) &= -f(t)\overline{g}(t) + \tfrac{1}{2}\|\bar{\xi}(t)\|^2 \overline{P}(t)^{-1}\overline{g}(t)^2 \\ &= \left[-f(t) + \tfrac{1}{2}\|\bar{\xi}(t)\|^2 e^{-\int_t^T A(s) ds} \int_t^T f(z) e^{\int_z^T A(s) ds} dz \right] \\ &\quad \cdot e^{\int_t^T (A(s) - \|\bar{\xi}(s)\|^2) ds} \int_t^T f(z) e^{\int_z^T A(s) ds} dz. \end{aligned}$$

Therefore,

$$\begin{aligned} \bar{c}(t) &= \int_t^T \left[f(v) - \tfrac{1}{2}\|\bar{\xi}(v)\|^2 e^{-\int_v^T A(s) ds} \int_v^T f(z) e^{\int_z^T A(s) ds} dz \right] \\ &\quad \cdot e^{\int_v^T (A(s) - \|\bar{\xi}(s)\|^2) ds} \int_v^T f(z) e^{\int_z^T A(s) ds} dz dv. \end{aligned}$$

Now we define the region Γ_1 in the (t, x)-plane as

$$\Gamma_1 := \left\{ (t, x) \in [0, T] \times \mathbb{R} \;\middle|\; x + e^{-\int_t^T A(s) ds} \int_t^T f(z) e^{\int_z^T A(s) ds} dz < 0 \right\}.$$

In Γ_1, V as given by (3.17) is sufficiently smooth for the terms in (3.4) to be well-defined, with

$$V_t(t, x) = \tfrac{1}{2}\dot{\overline{P}}(t)x^2 + \dot{\overline{g}}(t)x + \dot{\bar{c}}(t), \quad V_x(t, x) = \overline{P}(t)x + \overline{g}(t), \quad V_{xx}(t, x) = \overline{P}(t).$$

Substituting these into the left-hand side (LHS) of (3.4), we obtain

$$\begin{aligned} \text{LHS} &= V_t(t, x) + V_x(t, x)[A(t)x + f(t)] + \inf_{u \geq 0} \left[\tfrac{1}{2} V_{xx}(t, x) u' D(t)' D(t) u + V_x(t, x) B(t) u \right] \\ &= \left[\tfrac{1}{2}\dot{\overline{P}}(t)x^2 + \dot{\overline{g}}(t)x + \dot{\bar{c}}(t) \right] + [\overline{P}(t)x + \overline{g}(t)][A(t)x + f(t)] \\ &\quad + \inf_{u \geq 0} \left\{ \tfrac{1}{2}\overline{P}(t) u' D(t)' D(t) u + [\overline{P}(t)x + \overline{g}(t)] B(t) u \right\} \\ &= \left[\tfrac{1}{2}\dot{\overline{P}}(t) + A(t)\overline{P}(t) \right] x^2 + \left[\dot{\overline{g}}(t) + A(t)\overline{g}(t) + f(t)\overline{P}(t) \right] x + \left[\dot{\bar{c}}(t) + f(t)\overline{g}(t) \right] \\ &\quad + \overline{P}(t) \inf_{u \geq 0} \left\{ \tfrac{1}{2} u' D(t)' D(t) u + [x + \overline{\eta}(t)] B(t) u \right\}. \end{aligned} \tag{3.21}$$

By using Lemma 3.2 with $\alpha = -[x + \overline{\eta}(t)] > 0$, it follows that the minimizer of (3.21) is achieved by

$$\begin{aligned} u^*(t, x) &= -D(t)^{-1}\bar{\xi}(t)[x + \overline{\eta}(t)] \\ &= -D(t)^{-1}\bar{\xi}(t) \left[x + e^{-\int_t^T A(s) ds} \int_t^T f(z) e^{\int_z^T A(s) ds} dz \right]. \end{aligned} \tag{3.22}$$

Substituting $u^*(t,x)$ back into (3.21) and noting (3.11), (3.12), and (3.13), it immediately follows that the LHS $= 0$. This implies that V satisfies the HJB equation (3.4) in Γ_1.

Remark 3.3. Although the minimizer (3.22) of (3.21) involves the parameter $\bar{\xi}(t)$, as defined by (3.9) and (3.10), it is important to recognize that $\bar{\xi}(t)$ does not depend on x. In particular, this means that $\overline{P}(t)$, $\overline{g}(t)$, and $\overline{c}(t)$, which also depend on $\bar{\xi}(t)$, do not depend on x. Hence, the expressions for $V_t(t,x)$, $V_x(t,x)$, and $V_{xx}(t,x)$ do not involve terms of the form $\overline{P}_x(t)$, $\overline{g}_x(t)$, and $\overline{c}_x(t)$, etc. It is precisely for this reason that closed form expressions for the value function can still be obtained.

Remark 3.4. Although the expression of $\bar{\xi}(\cdot)$ is not explicitly analytical as it involves $\bar{z}(\cdot)$, it can easily be obtained numerically via solving the quadratic program in (3.9) off line.

Next we proceed to the region Γ_2 defined by

$$\Gamma_2 := \left\{ (t,x) \in [0,T] \times \mathbb{R} \;\middle|\; x + e^{-\int_t^T A(s)ds} \int_t^T f(z) e^{\int_z^T A(s)ds} dz > 0 \right\}.$$

Similar to the derivations for the previous case, we obtain

$$\begin{cases} \widetilde{P}(t) &= e^{2\int_t^T A(s)ds}, \\ \widetilde{g}(t) &= e^{\int_t^T A(s)ds} \displaystyle\int_t^T f(z) e^{\int_z^T A(s)ds} dz, \\ \widetilde{c}(t) &= \displaystyle\int_t^T f(v) e^{\int_v^T A(s)ds} \int_v^T f(z) e^{\int_z^T A(s)ds} dz dv, \\ \widetilde{\eta}(t) &= \frac{\widetilde{g}(t)}{\widetilde{P}(t)} = e^{-\int_t^T A(s)ds} \displaystyle\int_t^T f(z) e^{\int_z^T A(s)ds} dz. \end{cases}$$

In Γ_2, V is once again sufficiently smooth for the derivatives in (3.4) to be well-defined, and

$$V_t(t,x) = \tfrac{1}{2}\dot{\widetilde{P}}(t)x^2 + \dot{\widetilde{g}}(t)x + \dot{\widetilde{c}}(t), \quad V_x(t,x) = \widetilde{P}(t)x + \widetilde{g}(t), \quad V_{xx}(t,x) = \widetilde{P}(t).$$

Substituting into the LHS of (3.4), we obtain

$$\begin{aligned} \text{LHS} &= V_t(t,x) + V_x(t,x)[A(t)x + f(t)] + \inf_{u \geq 0}\left[\tfrac{1}{2} V_{xx}(t,x) u' D(t)' D(t) u + V_x(t,x) B(t) u\right] \\ &= \left[\tfrac{1}{2}\dot{\widetilde{P}}(t)x^2 + \dot{\widetilde{g}}(t)x + \dot{\widetilde{c}}(t)\right] + [\widetilde{P}(t)x + \widetilde{g}(t)][A(t)x + f(t)] \\ &\quad + \inf_{u \geq 0}\left\{\tfrac{1}{2}\widetilde{P}(t) u' D(t)' D(t) u + [\widetilde{P}(t)x + \widetilde{g}(t)] B(t) u\right\} \\ &= \left[\tfrac{1}{2}\dot{\widetilde{P}}(t) + A(t)\widetilde{P}(t)\right]x^2 + \left[\dot{\widetilde{g}}(t) + A(t)\widetilde{g}(t) + f(t)\widetilde{P}(t)\right]x + \left[\dot{\widetilde{c}}(t) + f(t)\widetilde{g}(t)\right] \\ &\quad + \widetilde{P}(t) \inf_{u \geq 0}\left\{\tfrac{1}{2} u' D(t)' D(t) u + [x + \widetilde{\eta}(t)] B(t) u\right\}. \end{aligned} \tag{3.23}$$

Since $x + \widetilde{\eta}(t) > 0$, the minimizer of (3.23) is

$$u^*(t,x) = 0. \tag{3.24}$$

Substituting $u^*(t,x)$ into (3.23), it is easy to show that V satisfies the HJB equation (3.4) in Γ_2.

Finally, the switching curve Γ_3 defined by

$$\Gamma_3 := \left\{ (t,x) \in [0,T] \times \mathbb{R} \;\middle|\; x + e^{-\int_t^T A(s)ds} \int_t^T f(z) e^{\int_z^T A(s)ds} dz = 0 \right\}$$

is where the nonsmoothness of V occurs. First, a direct calculation shows that $V(t,x) = \frac{1}{2}\overline{P}(t)x^2 + \overline{g}(t)x + \overline{c}(t) = \frac{1}{2}\widetilde{P}(t)x^2 + \widetilde{g}(t)x + \widetilde{c}(t) = 0$ on Γ_3. Therefore, $V(t,x)$ is continuous at $(t,x) \in \Gamma_3$. In addition, we also easily obtain

$$\begin{cases} V_t(t,x) = \frac{1}{2}\dot{\overline{P}}(t)x^2 + \dot{\overline{g}}(t)x + \dot{\overline{c}}(t) = \frac{1}{2}\dot{\widetilde{P}}(t)x^2 + \dot{\widetilde{g}}(t)x + \dot{\widetilde{c}}(t) = 0, \\ V_x(t,x) = \overline{P}(t)x + \overline{g}(t) = \widetilde{P}(t)x + \widetilde{g}(t) = 0. \end{cases}$$

That is, $V(t,x)$ is also continuously differentiable at points on Γ_3. However, V_{xx} does not exist on Γ_3, since $\overline{P}(t) \not\equiv \widetilde{P}(t)$. This means that V does not possess the necessary smoothness properties to qualify as a classical solution of the HJB equation (3.4). For this reason, we are required to work within the framework of viscosity solutions. From Definition 6.1 in the appendix, it can be shown that for any $(t,x) \in \Gamma_3$,

$$\begin{cases} D^{1,2,+}_{t,x} V(t,x) = \{0\} \times \{0\} \times [\widetilde{P}(t), +\infty), \\ D^{1,2,-}_{t,x} V(t,x) = \{0\} \times \{0\} \times (-\infty, \overline{P}(t)]. \end{cases} \tag{3.25}$$

For the HJB equation (3.4), we define $G(t,x,u,p,P) = p[A(t)x + B(t)u + f(t)] + \frac{1}{2}Pu'D(t)'D(t)u$. For any $(q,p,P) \in D^{1,2,+}_{t,x} V(t,x)$, where $(t,x) \in \Gamma_3$, we have

$$q + \inf_{u \geq 0} G(t,x,u,p,P) = \inf_{u \geq 0} \left\{ \tfrac{1}{2} Pu'D(t)'D(t)u \right\} \geq \inf_{u \geq 0} \left\{ \tfrac{1}{2} \widetilde{P}(t)u'D(t)'D(t)u \right\} = 0.$$

Therefore, V is a viscosity subsolution of the HJB equation (3.4). On the other hand, for $(q,p,P) \in D^{1,2,-}_{t,x} V(t,x)$, where $(t,x) \in \Gamma_3$, we have

$$q + \inf_{u \geq 0} G(t,x,u,p,P) = \inf_{u \geq 0} \left\{ \tfrac{1}{2} Pu'D(t)'D(t)u \right\} \leq \inf_{u \geq 0} \left\{ \tfrac{1}{2} \overline{P}(t)u'D(t)'D(t)u \right\} = 0.$$

Therefore, V is also a viscosity supersolution of the HJB equation (3.4). Finally, it is easy to see that the terminal condition $V(T,x) = \frac{1}{2}x^2$ is satisfied. Hence, it follows from Definition 6.1 that $V(t,x)$ is a viscosity solution of the HJB equation (3.4). Moreover, for any $(t,x) \in \Gamma_3$, take $(q^*(t,x), p^*(t,x), P^*(t,x), u^*(t,x)) := (0, 0, \widetilde{P}(t), 0) \in D^{1,2,+}_{t,x} V(t,x) \times \mathcal{U}[s,T]$; then

$$q^*(t,x) + G(t,x,u^*(t,x), p^*(t,x), P^*(t,x)) = 0.$$

It then follows from the *verification theorem* in [38, Theorem 3.1] that $u^*(t,x)$ defined by (3.18) is the optimal feedback control.

Remark 3.5. We mention, once again, that our proof that the control (3.18) is optimal for the problem (3.1)–(3.2) is based on the viscosity verification theorem from [38, Theorem 3.1]. This enables us to solve the constrained LQ problem (3.1)–(3.2) without the technicalities of the duality analysis in [6, 18].

4. Efficient strategies and efficient frontier. In this section we apply the general results established in the previous section to the problem $A(\mu)$ formulated in section 2. Set

$$x(t) = X(t) - (d - \mu).$$

Problem $A(\mu)$ is equivalent to the following problem:

$$\min E\Big\{\tfrac{1}{2}x(T)^2\Big\}, \tag{4.1}$$

subject to

$$\begin{cases} dx(t) = [A(t)x(t) + B(t)u(t) + f(t)]dt + \sum_{j=1}^{m} D_j(t)u(t)dW^j(t), \\ x(0) = X_0 - (d-\mu), \end{cases} \tag{4.2}$$

where $u(\cdot) \in L^2_{\mathcal{F}}(0,T;\mathbb{R}^m_+)$ and

$$\begin{cases} A(t) = r(t), & B(t) = (b_1(t) - r(t), \dots, b_m(t) - r(t)), \\ f(t) = (d-\mu)r(t), & D_j(t) = (\sigma_{1j}(t), \dots, \sigma_{mj}(t)). \end{cases} \tag{4.3}$$

Now, corresponding to (3.9) and (3.10), set

$$\bar{\pi}(t) := \arg \min_{\pi(t)\in[0,\infty)^m} \tfrac{1}{2}\|\sigma(t)^{-1}\pi(t) + \sigma(t)^{-1}(b(t) - r(t)\mathbf{1})\|^2 \tag{4.4}$$

and

$$\bar{\theta}(t) := \sigma(t)^{-1}\bar{\pi}(t) + \sigma(t)^{-1}(b(t) - r(t)\mathbf{1}). \tag{4.5}$$

4.1. An optimal strategy. Before analyzing the efficient frontier of the original portfolio selection problem (2.6), we first present the optimal investment strategy for the problem $A(\mu)$. The optimal control obtained in (3.18) translates into the following strategy:

$$\begin{aligned} & u^*(t,X) \equiv (u_1^*(t,X), \dots, u_m^*(t,X))' \\ = & \begin{cases} -(\sigma(t)')^{-1}\bar{\theta}(t)\Big[x + (d-\mu)(1 - e^{-\int_t^T r(s)ds})\Big] \\ \qquad\qquad \text{if } x + (d-\mu)(1 - e^{-\int_t^T r(s)ds}) \le 0 \\ 0 \qquad\qquad \text{if } x + (d-\mu)(1 - e^{-\int_t^T r(s)ds}) > 0, \end{cases} \\ = & \begin{cases} -(\sigma(t)\sigma(t)')^{-1}[\bar{\pi}(t) + (b(t) - r(t)\mathbf{1})]\Big[X - (d-\mu)e^{-\int_t^T r(s)ds}\Big] \\ \qquad\qquad \text{if } X - (d-\mu)e^{-\int_t^T r(s)ds} \le 0, \\ 0 \qquad\qquad \text{if } X - (d-\mu)e^{-\int_t^T r(s)ds} > 0. \end{cases} \end{aligned} \tag{4.6}$$

THEOREM 4.1. *An optimal investment strategy to the problem $A(\mu)$ is given by* (4.6).

4.2. Efficient frontier. In this subsection, we derive the efficient frontier for the portfolio selection problem (2.6), i.e., we specify the relation between the variance and the expected value of the terminal wealth for every efficient strategy. First of all, note that

$$\begin{aligned} E\Big\{\tfrac{1}{2}x(T)^2\Big\} &= E\Big\{\tfrac{1}{2}[X(T) - (d-\mu)]^2\Big\} \\ &= E\Big\{\tfrac{1}{2}[X(T) - d]^2\Big\} + \mu[EX(T) - d] + \tfrac{1}{2}\mu^2. \end{aligned}$$

Hence, for every fixed μ, we have

$$\begin{aligned}
&\min_{u(\cdot)\in\mathcal{U}[0,T]} E\Big\{\tfrac{1}{2}[X(T)-d]^2+\mu[EX(T)-d]\Big\}\\
&=\min_{u(\cdot)\in\mathcal{U}[0,T]} E\Big\{\tfrac{1}{2}x(T)^2\Big\}-\tfrac{1}{2}\mu^2\\
&=V(0,x(0))-\tfrac{1}{2}\mu^2\\
&=\tfrac{1}{2}P(0)x(0)^2+g(0)x(0)+c(0)-\tfrac{1}{2}\mu^2\\
&=\tfrac{1}{2}P(0)[X_0-(d-\mu)]^2+g(0)[X_0-(d-\mu)]+c(0)-\tfrac{1}{2}\mu^2,
\end{aligned}$$

where $P(\cdot), g(\cdot)$, and $c(\cdot)$ are either $\bar{P}(\cdot), \bar{g}(\cdot)$, and $\bar{c}(\cdot)$ or $\tilde{P}(\cdot), \tilde{g}(\cdot)$, and $\tilde{c}(\cdot)$, respectively, depending on whether or not $X_0-(d-\mu)e^{-\int_0^T r(s)ds}\le 0$ (see (3.17)). Now, if $X_0-(d-\mu)e^{-\int_0^T r(s)ds}\le 0$, we have a concave quadratic function in μ

$$\begin{aligned}
&\min_{u(\cdot)\in\mathcal{U}[0,T]} E\Big\{\tfrac{1}{2}[X(T)-d]^2+\mu[EX(T)-d]\Big\}\\
&=\tfrac{1}{2}\overline{P}(0)[X_0-(d-\mu)]^2+\overline{g}(0)[X_0-(d-\mu)]+\overline{c}(0)-\tfrac{1}{2}\mu^2\\
&=\tfrac{1}{2}e^{-\int_0^T\|\bar{\theta}(s)\|^2ds}\Big[X_0e^{\int_0^T r(s)ds}-(d-\mu)\Big]^2-\tfrac{1}{2}\mu^2.
\end{aligned}$$

If $X_0-(d-\mu)e^{-\int_0^T r(s)ds}>0$, we have a linear function in μ

$$\begin{aligned}
&\min_{u(\cdot)\in\mathcal{U}[0,T]} E\Big\{\tfrac{1}{2}[X(T)-d]^2+\mu[EX(T)-d]\Big\}\\
&=\tfrac{1}{2}\widetilde{P}(0)[X_0-(d-\mu)]^2+\widetilde{g}(0)[X_0-(d-\mu)]+\widetilde{c}(0)-\tfrac{1}{2}\mu^2\\
&=\tfrac{1}{2}\Big[X_0e^{\int_0^T r(s)ds}-(d-\mu)\Big]^2-\tfrac{1}{2}\mu^2\\
&=\tfrac{1}{2}\Big(X_0e^{\int_0^T r(s)ds}-d\Big)^2+\Big(X_0e^{\int_0^T r(s)ds}-d\Big)\mu.
\end{aligned}$$

Therefore we conclude that under the optimal investment strategy (4.6) the optimal cost for problem (2.7) is

(4.7)

$$\begin{aligned}
&\min_{u(\cdot)\in\mathcal{U}[0,T]} E\Big\{[X(T)-d]^2+2\mu[EX(T)-d]\Big\}\\
&=\begin{cases}
e^{-\int_0^T\|\bar{\theta}(s)\|^2ds}\Big[X_0e^{\int_0^T r(s)ds}-(d-\mu)\Big]^2-\mu^2 & \text{if } X_0-(d-\mu)e^{-\int_0^T r(s)ds}\le 0,\\
\Big[X_0e^{\int_0^T r(s)ds}-(d-\mu)\Big]^2-\mu^2 & \text{if } X_0-(d-\mu)e^{-\int_0^T r(s)ds}>0.
\end{cases}
\end{aligned}$$

Note that the above value still depends on the Lagrange multiplier μ. To obtain the optimal value (i.e., the minimum variance Var $X(T)$) and optimal strategy for the original portfolio selection problem (2.6) one needs to maximize the value in (4.7) over $\mu\in\mathbb{R}$ according to the Lagrange duality theorem [23]. A simple calculation shows that (4.7) attains its maximum value

$$\frac{\Big(d-X_0e^{\int_0^T r(s)ds}\Big)^2}{e^{\int_0^T\|\bar{\theta}(s)\|^2ds}-1}\quad\text{at}\quad\mu^*=\frac{d-X_0e^{\int_0^T r(s)ds}}{1-e^{\int_0^T\|\bar{\theta}(s)\|^2ds}}.$$

(Note that in the calculation we made use of the fact that

$$X_0-(d-\mu^*)e^{-\int_0^T r(s)ds}=\frac{de^{-\int_0^T r(s)ds}-X_0}{e^{-\int_0^T\|\bar{\theta}(s)\|^2ds}-1}\le 0,$$

due to Assumption 2.1.)

The above discussion leads to the following theorem.

THEOREM 4.2. *The efficient strategy of portfolio selection problem* (2.6) *corresponding to the expected terminal wealth* $EX(T) = d$*, as a function of time* t *and wealth* X*, is*

$$u^*(t,X) \equiv (u_1^*(t,X), \ldots, u_m^*(t,X))' = \begin{cases} -(\sigma(t)\sigma(t)')^{-1}[\bar{\pi}(t) + (b(t) - r(t)\mathbf{1})]\Big[X - (d-\mu^*)e^{-\int_t^T r(s)ds}\Big] \\ \qquad\qquad \text{if } X - (d-\mu^*)e^{-\int_t^T r(s)ds} \le 0, \\ 0 \qquad\qquad \text{if } X - (d-\mu^*)e^{-\int_t^T r(s)ds} > 0, \end{cases} \tag{4.8}$$

where $\mu^* = \frac{d - X_0 e^{\int_0^T r(s)ds}}{1 - e^{\int_0^T \|\bar{\theta}(s)\|^2 ds}}$*, and* $\bar{\pi}(\cdot)$ *and* $\bar{\theta}(\cdot)$ *are defined in* (4.4) *and* (4.5)*, respectively. Moreover, the efficient frontier is*

$$\text{Var } X(T) = \frac{(d - X_0 e^{\int_0^T r(s)ds})^2}{e^{\int_0^T \|\bar{\theta}(s)\|^2 ds} - 1} \equiv \frac{(EX(T) - X_0 e^{\int_0^T r(s)ds})^2}{e^{\int_0^T \|\bar{\theta}(s)\|^2 ds} - 1}. \tag{4.9}$$

Remark 4.1. The form of the efficient strategy (4.8) suggests that it should put all the money in the bond if the current wealth is large enough.

Remark 4.2. The so-called *mutual fund theorem*, due originally to Tobin [31] for single-period investment, is a natural consequence of the mean-variance theory and is the foundation of the CAPM (*Capital Asset Pricing Model*; Sharpe [29]). It basically asserts that any mean-variance efficient portfolio is a convex combination of the riskless asset and a prescribed portfolio containing only the risky assets. (The latter is called the *tangent fund*.) As an immediate consequence, in all the efficient portfolios the allocations among the risky assets have constant proportions—the same as those in the tangent fund. In particular, it implies that those proportions should not depend on the total wealth of the investor. It then follows from (4.8) that the mutual fund theorem does not hold under the short-selling prohibition, because the fraction of wealth in stocks in an efficient portfolio does depend on the wealth of the agent. However, we see that a modified form of the mutual fund theorem still holds true in the present case. Specifically, we now have two modes depending on whether $X - (d-\mu^*)e^{-\int_t^T r(s)ds} \le 0$. In each mode the allocations among the stocks keep constant proportions.

Remark 4.3. The efficient frontier in the mean-standard deviation diagram is still a straight line, as with the single-period mean-variance setting (see, e.g., [24]). To be specific, let $\sigma_{X(T)}$ be the standard deviation of the terminal wealth; then (4.9) gives

$$EX(T) = X_0 e^{\int_0^T r(s)ds} + \sigma_{X(T)} \sqrt{e^{\int_0^T \|\bar{\theta}(s)\|^2 ds} - 1}, \tag{4.10}$$

which is also called the capital market line.

5. An example. In this section, a numerical example is presented to demonstrate the results in the previous section. Let $m = 3$. The interest rate of the bond and the appreciation rate of the m stocks are $r = \frac{2}{100}$ and $(b_1, b_2, b_3)' = (\frac{4}{100}, \frac{5}{100}, \frac{6}{100})'$, respectively, and the volatility matrix is

$$\sigma = \begin{bmatrix} 1 & 0 & \frac{2}{3} \\ 0 & 1 & 0 \\ 0 & 0 & \frac{2}{3} \end{bmatrix}.$$

Then we have

$$\sigma^{-1} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{3}{2} \end{bmatrix}$$

and $(b_1 - r, b_2 - r, b_3 - r)' = (\frac{2}{100}, \frac{3}{100}, \frac{4}{100})'$. Hence, $\theta := \sigma^{-1}(b - r\mathbf{1}) = (\frac{-2}{100}, \frac{3}{100}, \frac{6}{100})'$. Obviously, $s(\pi) \triangleq \frac{1}{2}\|\sigma^{-1}\pi + \theta\|^2$ over $[0,\infty)^m$ has a unique minimizer $\bar{\pi} = \left(\frac{2}{100}, 0, 0\right)'$ with the minimum value $s(\bar{\pi}) = \frac{1}{2}\|\sigma^{-1}\bar{\pi} + \theta\|^2 = \frac{9}{4000}$. Then we have

$$\|\bar{\theta}\|^2 = \|\sigma^{-1}\bar{\pi} + \theta\|^2 = \frac{9}{2000}$$

and

$$(\sigma\sigma')^{-1}[\bar{\pi} + (b - r\mathbf{1})] = \left(\ 0, \quad \frac{3}{100}, \quad \frac{9}{100} \ \right)'.$$

Therefore, Theorem 4.2 implies that an efficient strategy is

$$\begin{aligned} u^*(t,X) &\equiv (u_1^*(t,X), u_2^*(t,X), u_3^*(t,X))' \\ &= \begin{cases} \begin{bmatrix} 0 \\ 3/100 \\ 9/100 \end{bmatrix} [(d - \mu^*)e^{\frac{2}{100}(t-T)} - X] & \text{if } X - (d-\mu^*)e^{\frac{2}{100}(t-T)} \le 0, \\ 0 & \text{if } X - (d-\mu^*)e^{\frac{2}{100}(t-T)} > 0, \end{cases} \end{aligned}$$

where $\mu^* = \frac{d - X_0 e^{rT}}{1 - e^{\|\bar{\theta}\|^2 T}} = \frac{d - X_0 e^{\frac{2}{100}T}}{1 - e^{\frac{9}{2000}T}}$. The efficient frontier is

$$\text{Var } X(T) = \frac{(d - X_0 e^{rT})^2}{e^{\|\bar{\theta}\|^2 T} - 1} = \frac{\left[EX(T) - X_0 e^{\frac{2}{100}T}\right]^2}{e^{\frac{9}{2000}} - 1}.$$

6. Conclusion. This paper investigates a continuous-time mean-variance portfolio selection problem where short-selling is not allowed. The efficient strategies and efficient frontier are derived explicitly based on stochastic LQ control technique and viscosity solution theory. This also demonstrates that stochastic LQ control is a powerful framework to treat some finance problems.

An immediate open problem is to extend the results in this paper to the case in which all the market coefficients are random processes. This is a challenging problem because the HJB equation becomes a backward stochastic partial differential equation due to the randomness of coefficients for which viscosity solution theory is still largely unexplored.

Appendix: Viscosity solutions. We list here some basic terminologies from the theory of viscosity solutions which are referred to in the paper.

Let

$$G(t,x,u,p,P) = \tfrac{1}{2}\sigma(t,x,u)'P\sigma(t,x,u) + p'h(t,x,u) - L(t,x,u),$$

where $\sigma : [0,T) \times \mathbb{R}^n \times \mathbb{R}^{n_u} \to \mathbb{R}^n$, $h : [0,T) \times \mathbb{R}^n \times \mathbb{R}^{n_u} \to \mathbb{R}^n$, and $L : [0,T) \times \mathbb{R}^n \times \mathbb{R}^{n_u} \to \mathbb{R}$. Consider the second-order PDE

$$\begin{cases} v_t + \inf\limits_{u \ge 0} G(t,x,u,v_x,v_{xx}) = 0, \quad (t,x) \in [0,T) \times \mathbb{R}^n, \\ v(T,x) = g(x), \end{cases} \tag{6.1}$$

where $g : \mathbb{R}^n \to \mathbb{R}$.

Clearly the HJB equation (3.4) is a special case of (6.1). It is well-known that (6.1) does not in general have classical (smooth) solutions. A generalized concept of solution, called a viscosity solution, is introduced in [5]. The main result in [35] is that under certain mild conditions there exists a unique viscosity solution in the first-order case. In the second-order case, uniqueness is proven in [15]. See also [9, 35] for more details about the viscosity solution and its application in stochastic control.

DEFINITION 6.1. *Let $v \in C([0,T] \times \mathbb{R}^n)$ and $(t_0, x_0) \in (0,T) \times \mathbb{R}^n$. Then the second-order superdifferential of v at (t_0, x_0) is defined by*

$$D^{1,2,+}_{t,x} v(t_0,x_0) = \Big\{ (\varphi_t(t_0,x_0),\, \varphi_x(t_0,x_0), \varphi_{xx}(t_0,x_0)) \;\Big|\; \varphi \in C^\infty((0,T)\times\mathbb{R}^n) \text{ and } v-\varphi \text{ has a local maximum at } (t_0,x_0) \Big\}, \tag{6.2}$$

and the second order subdifferential of v is defined by

$$D^{1,2,-}_{t,x} v(t_0,x_0) = \Big\{ (\varphi_t(t_0,x_0),\, \varphi_x(t_0,x_0), \varphi_{xx}(t_0,x_0)) \;\Big|\; \varphi \in C^\infty((0,T)\times\mathbb{R}^n) \text{ and } v-\varphi \text{ has a local minimum at } (t_0,x_0) \Big\}. \tag{6.3}$$

Moreover, v is a viscosity solution of (6.1) *if*

$$v(T,x) = g(x) \quad \forall\, x \in \mathbb{R}^n, \tag{6.4}$$

and

$$q + \inf_{u\in U} G(t,x,u,p,P) \geq 0 \quad \forall (q,p,P) \in D^{1,2,+}_{t,x} v(t,x), \tag{6.5}$$

$$q + \inf_{u\in U} G(t,x,u,p,P) \leq 0 \quad \forall (q,p,P) \in D^{1,2,-}_{t,x} v(t,x), \tag{6.6}$$

for all $(t,x) \in [0,T) \times \mathbb{R}^n$.

In particular, v is called a *viscosity subsolution* if it satisfies (6.4)–(6.5), and a *viscosity supersolution* if it satisfies (6.4) and (6.6).

REFERENCES

[1] J.Y. CAMPBELL, A.W. LO, AND A.C. MACKINLAY, *The Econometrics of Financial Markets*, Princeton University Press, Princeton, NJ, 1997.

[2] S. CHEN, X. LI, AND X.Y. ZHOU, *Stochastic linear quadratic regulators with indefinite control weight costs,* SIAM J. Control Optim., 36 (1998), pp. 1685–1702.

[3] S. CHEN AND X. ZHOU, *Stochastic linear quadratic regulators with indefinite control weight costs.* II, SIAM J. Control Optim., 39 (2000), pp. 1065–1081.

[4] J. COX AND C.F. HUANG, *Optimal consumption and portfolio policies when asset prices follow a diffusion process,* J. Econom. Theory, 49 (1989), pp. 33–83.

[5] M.G. CRANDALL AND P.L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations,* Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[6] J. CVITANIC AND I. KARATZAS, *Convex duality in constrained portfolio optimization,* Ann. Appl. Probab., 2 (1992), pp. 767–818.

[7] D. DUFFIE AND H. RICHARDSON, *Mean-variance hedging in continuous time,* Ann. Appl. Probab., 14 (1991), pp. 1–15.

[8] B. DUMAS AND E. LUCIANO, *An exact solution to a dynamic portfolio choice problem under transactions costs,* J. Finance, 46 (1991), pp. 577–595.

[9] W.H. FLEMING AND H.M. SONER, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[10] J.C. Francis, *Investments: Analysis and Management*, McGraw-Hill, New York, 1976.
[11] R.R. Grauer, *A comparison of growth optimal and mean-variance investment policies,* J. Financial and Quantitative Analysis, 16 (1981), pp. 1–21.
[12] R.R. Grauer and N.H. Hakansson, *On the use of mean-variance and quadratic approximations in implementing dynamic investment strategies: A comparison of the returns and investment policies,* Management Sci., 39 (1983), pp. 856–871.
[13] S.J. Grossman and Z. Zhou, *Equilibrium analysis of portfolio insurance,* J. Finance, 51 (1996), pp. 1379–1403.
[14] N.H. Hakansson, *Capital growth and the mean-variance approach to portfolio selection,* J. Financial and Quantitative Analysis, 6 (1971), pp. 517–557.
[15] R. Jensen, *The maximum principle for viscosity solutions of second order fully nonlinear partial differential equations,* Arch. Ration. Mech. Anal., 101 (1988), pp. 1–27.
[16] I. Karatzas and S.G. Kou, *Hedging American contingent claims with constrained portfolios,* Finance and Stochastics, 2 (1998), pp. 215–258.
[17] I. Karatzas, J.P. Lehoczky, and S.E. Shreve, *Optimal portfolio and consumption decisions for a "small investor" on a finite horizon,* SIAM J. Control Optim., 25 (1987), pp. 1557–1586.
[18] I. Karatzas and S.E. Shreve, *Methods of Mathematical Finance*, Springer-Verlag, New York, 1999.
[19] M. Kohlmann and X. Y. Zhou, *Relationship between backward stochastic differential equations and stochastic controls: A linear-quadratic approach,* SIAM J. Control Optim., 38 (2000), pp. 1392–1407.
[20] D. Li and W.L. Ng, *Optimal dynamic portfolio selection: Multi-period mean-variance formulation,* Math. Finance, 10 (2000), pp. 387–406.
[21] A.E.B. Lim and X.Y. Zhou, *Optimal stochastic LQR control with integral quadratic constraints and indefinite control weights,* IEEE Trans. Automat. Control, 44 (1999), pp. 1359–1369.
[22] A.E.B. Lim and X.Y. Zhou, *Mean-variance portfolio selection with random parameters,* Math. Oper. Res., to appear.
[23] D.G. Luenberger, *Optimization by Vector Space Methods*, Wiley, New York, 1968.
[24] H. Markowitz, *Portfolio selection,* J. Finance, 7 (1959), pp. 77–91.
[25] R.C. Merton, *Lifetime portfolio selection under uncertainty: the continuous-time case,* Rev. Econom. Statist., 51 (1969), pp. 247–257.
[26] R.C. Merton, *An analytical derivation of the efficient portfolio frontier,* J. Financial and Economics Anal., 7 (1972), pp. 1851–1872.
[27] J. Mossin, *Optimal multi-period portfolio policies,* J. Business, 41 (1968), pp. 215–229.
[28] P.A. Samuelson, *Lifetime portfolio selection by dynamic stochastic programming,* Rev. Econom. Statist., 51 (1969), pp. 236–246.
[29] W.F. Sharpe, *Capital asset prices: A theory of market equilibrium under conditions of risk,* J. Finance, 19 (1964), pp. 425–442.
[30] H. Shirakawa, *Optimal consumption and portfolio selection with incomplete markets and upper and lower bound constraints,* Math. Finance, 4 (1994), pp. 1–24.
[31] J. Tobin, *Liquidity preference as behavior towards risk,* Rev. Econom. Statist., 25 (1958), pp. 65–86.
[32] G.L. Xu and S.E. Shreve, *A duality method for optimal consumption and investment under short-selling prohibition:* I.*General Market Coefficients,* Ann. Appl. Probab., 2 (1992), pp. 87–112.
[33] G.L. Xu and S.E. Shreve, *A duality method for optimal consumption and investment under short-selling prohibition:* II. *Constant market coefficients,* Ann. Appl. Probab., 2 (1992), pp. 314–328.
[34] D. Yao, S. Zhang, and X.Y. Zhou, *Stochastic linear quadratic control via semidefinite programming,* SIAM J. Control Optim., 40 (2001), pp. 801–823.
[35] J. Yong and X.Y. Zhou, *Stochastic Controls: Hamiltonian Systems and HJB Equations*, Springer-Verlag, New York, 1999.
[36] Y.G. Zhao and W.T. Ziemba, Mean-variance versus expected utility in dynamic investment analysis, working paper, Faculty of Commerce and Business Administration, University of British Columbia, Vancouver, BC, Canada, 2000.
[37] X.Y. Zhou and D. Li, *Continuous time mean-variance portfolio selection: A stochastic LQ framework,* Appl. Math. Optim., 42 (2000), pp. 19–33.
[38] X.Y. Zhou, J. Yong, and X. Li, *Stochastic verification theorems within the framework of viscosity solutions,* SIAM J. Control Optim., 35 (1997), pp. 243–253.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 5, pp. 1556–1575

SAFE COOPERATIVE ROBOT DYNAMICS ON GRAPHS*

ROBERT W. GHRIST† AND DANIEL E. KODITSCHEK‡

Abstract. This paper introduces the use of vector fields to design, optimize, and implement reactive schedules for safe cooperative robot patterns on planar graphs. We consider automated guided vehicles (AGVs) operating upon a predefined network of pathways. In contrast to the case of locally Euclidean configuration spaces, regularization of collisions is no longer a local procedure, and issues concerning the global topology of configuration spaces must be addressed. The focus of the present inquiry is the definition, design, and algorithmic construction of controllers for achievement of safe, efficient, cooperative patterns in the simplest nontrivial example (a pair of robots on a Y-network) by means of a hierarchical event-driven state feedback law.

Key words. configuration spaces, AGV, graph network, hierarchical control

AMS subject classifications. Primary, 93C85, 68T40; Secondary, 93C25, 37N35

PII. S0363012900368442

1. Introduction.

Recent literature suggests the growing awareness of a need for "reactive" scheduling wherein one desires not merely a single deployment of resources but a plan for successive redeployments against a changing environment [19]. However, scheduling problems have been traditionally solved by appeal to a discrete representation of the domain at hand. Thus the need for "tracking" changing goals introduces a conceptual dilemma: there is no obvious topology by which proximity to the target of a given deployment can be measured. In contrast to problems entailing the management of information alone, problems in many robotics and automation settings involve the management of *work*—the exchange of energy in the presence of geometric constraints. In these settings, it may be desirable to postpone the imposition of a discrete representation long enough to gain the benefit of the natural topology that accompanies the original domain.

This paper explores the use of vector fields for reactive scheduling of safe cooperative robot patterns on graphs. The word "safe" means that obstacles—designated illegal portions of the configuration space—are avoided. The word "cooperative" connotes situations wherein physically distributed agents are collectively responsible for executing the schedule. The word "pattern" refers to tasks that cannot be encoded simply in terms of a point goal in the configuration space. The word "reactive" will be interpreted as requiring feedback so that the desired pattern rejects perturbations: conditions close but slightly removed from those desired remain close and, indeed, converge toward the exactly desired pattern.

1.1. Setting: AGVs on a guidepath network of wires.

An automated guided vehicle (AGV) is an unmanned powered cart "capable of following an external guidance signal to deliver a unit load from destination to destination," where, in most

*Received by the editors February 28, 2000; accepted for publication (in revised form) September 24, 2001; published electronically February 6, 2002. This research was supported in part by National Science Foundation grant IRI-9510673 [DK] and by National Science Foundation grant DMS-9971629 [RG]. A sketch of these ideas appeared in [10].

http://www.siam.org/journals/sicon/40-5/36844.html

†School of Mathematics and CDSNS, Georgia Institute of Technology, Atlanta, GA 30017 (ghrist@math.gatech.edu).

‡Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, MI 48109-2110 (kod@eecs.umich.edu).

common applications, the guidepath signal is buried in the floor [6]. Thus the AGV's work space is a network of wires—a graph. The motivation to choose AGV-based materials handling systems over more conventional fixed conveyors rests not simply in their ease of reconfigurability but in the potential they offer for graceful response to perturbations in normal plant operation. In real production facilities, the flow of work in process fluctuates constantly in the face of unanticipated work station downtime, variations in process rate, and, indeed, variations in materials transport and delivery rates [8]. Of course, realizing their potential robustness against these fluctuations in work flow remains an only partially fulfilled goal of contemporary AGV systems.

Choreographing the interacting routes of multiple AGVs in a nonconflicting manner presents a novel, complicated, and necessarily online planning problem. Nominal routes might be designed offline, but they can never truly be traversed with the nominal timing for all the reasons described above. Even under normal operating conditions, no single nominal schedule can suffice to coordinate the work flow as the production volume or product mix changes over time: new vehicles need to be added or deleted, and the routing scheme needs to be adapted. In any case, abnormal conditions—unscheduled process down times, blocked work stations, failed vehicles—continually arise, demanding altered routes.

The traffic control schemes deployed in contemporary AGV systems are designed to simplify the real-time route planning and adaptation process by "blocking zone control" strategies. The work space is partitioned into a small number of cells, and, regardless of the details of their source and destination tasks, no two AGVs are ever allowed into the same cell at the same time [6]. Clearly, this simplification results in significant loss of a network's traffic capacity.

In this paper, we will consider a centralized approach that employs dynamical systems theory to focus on real-time responsiveness and efficiency as opposed to computational complexity or average throughput. Without a doubt, beyond a certain maximum number of vehicles, the necessity to compute in the high dimensional configuration space will limit the applicability of any algorithms that arise. However, this point of view seems not to have been carefully explored in the literature. Indeed, we will sketch some ideas about how an approach that starts from the coupled version of the problem may lend sufficient insight to move back and forth between the individuals' and the group's configuration spaces even in real time. For the sake of concreteness we will work in the so-called pickup and delivery (i.e., where loads are picked up at certain points and dropped off at others, as opposed to "stop and go," where an AGV network stands in for an assembly line [3]) paradigm of assembly or fabrication (where a desired steady state "pattern"—a scheduled series of visits to specific work stations by specific AGVs—is dispatched ahead of time), and we will not be concerned with warehousing-style AGV applications.

1.2. Organization of the paper.

Section 2. We introduce the salient properties of a feedback controlled dynamical system on a graph by addressing the closed loop motion planning problem of a single AGV on its wire network. In this setting, configuration spaces are not required, although the nonmanifold structure of the work space necessitates a mild adaptation of the dynamical systems machinery, specified in Appendix A. We describe a simple hybrid controller built from *edge point fields*—locally defined dynamics that realize single letter patterns—which generalize the scheme Burridge, Rizzi, and Koditschek have proposed in [5].

Section 3. Turning to the central topic, we address the case of multiple AGVs in

the simplest possible setting—two AGVs on a Y-graph—as a local exemplar of the general problem. The contributions of this section include

1. an intrinsic coordinate system for the configuration space;
2. a detailed analysis of the topology of the configuration space, affording immediate recourse to previously developed methods of safe controller design [12];
3. the construction of a "circulating flow" on this space that executes a stable safe periodic pattern as a canonical example of dynamically controlled collision-free behavior suitable to more general settings of the problem.

Section 4. Because limit cycles are likely too rigid a means of arbitrary pattern specification in the more generalized settings of the problem, we return to the notion of building a palette of control laws that realize safe "letters" along with a hybrid (logical level) scheme for concatenating them to produce arbitrary patterns in the form of periodic attracting orbits whose limit set is any desired "word," within a "monotone cycle" grammar as formalized in Theorem 3. Section 4 ends with a constructive procedure for incorporating performance guarantees in the construction of these grammars, concluding with a more speculative view of potential extensions of this work.

Appendix A is included to place on a rigorous foundation the use of vector fields on graphs and configuration spaces thereof.

2. Notation and background.

2.1. Graph topology. A *graph*, Γ, consists of a finite collection of 0-dimensional vertices $\mathcal{V}:=\{v_i\}_1^N$, and 1-dimensional edges $\mathcal{E}:=\{e_j\}_1^M$ assembled as follows. Each edge is homeomorphic to the closed interval $[0,1]$ attached to $\mathcal{V}$ along its boundary points $\{0\}$ and $\{1\}$.[1] We place upon Γ the quotient topology given by the endpoint identifications: neighborhoods of a point in the interior of e_j are homeomorphic images of interval neighborhoods of the corresponding point in $[0,1]$, and neighborhoods of a vertex v_i consist of the union of homeomorphic images of half-open neighborhoods of the endpoints for all incident edges.

The configuration spaces we consider in section 3 and throughout are subsets of self-products of graphs. The topology of $\Gamma \times \Gamma$ is easily understood in terms of the topology of Γ as follows [17]. Let $(x,y) \in \Gamma \times \Gamma$ denote an ordered pair in the product. Then any small neighborhood of (x,y) within $\Gamma \times \Gamma$ is the union of neighborhoods of the form $\mathcal{N}(u) \times \mathcal{N}(v)$, where $\mathcal{N}(\cdot)$ denotes the neighborhood within Γ. In other words, the products of neighborhoods form a *basis* of neighborhoods in the product space.

Given a graph, Γ, outfitted with a finite number N of noncolliding AGVs constrained to move on Γ, the (labeled) configuration space of safe motions is defined as

$$\mathcal{C}:= (\Gamma \times \cdots \times \Gamma) - \mathcal{N}(\Delta), \tag{2.1}$$

where $\Delta:=\{(x_i) \in \Gamma \times \cdots \times \Gamma : x_j = x_k \text{ for some } j \neq k\}$ denotes the pairwise diagonal and $\mathcal{N}(\cdot)$ denotes the (small) neighborhood.

[1] In our model, we will disallow "homoclinic" edges whose boundary points are attached to the same vertex. With respect to the application setting, this is very natural since vertices correspond to work stations along a path. It is hard to imagine networks designed with loops that do not service any work stations. In the worst case, there is established precedent in the AGV technology literature for introducing additional "transfer point" technology to a factory setting solely for purposes of traffic control [3].

For general graphs, the topological features of $\mathcal{C}$ can be extremely complicated, as measured by, say, the rank of the fundamental group (see [17] for definitions). Even in the case where the work space, Γ, is contractible (and thus, the product of its n copies is contractible), removal of this collision diagonal often creates spaces with a large fundamental group. For example, given a graph Γ_K with K edges all connected at a single point (forming a K-pronged "star"), it follows from the more general results in [9] that the fundamental group of the configuration space $\Gamma_K \times \Gamma_K - \mathcal{N}(\Delta)$ is a free group on $K^2 - 3K + 1$ generators; i.e., the number of "independent" closed paths in this space (with respect to continuous deformation) grows quadratically with K.

Mathematically, it is usually most interesting to pass to the quotient of $\mathcal{C}$ by the action of the permutation group on N elements, thus forgetting the identities of the AGV elements; however, as such spaces are almost completely divorced from any applications involving coordinated transport, we work on the "full" configuration space $\mathcal{C}$. We do not treat the general aspects of this problem comprehensively in this paper; rather, we restrict our attention to the simplest nontrivial example, which illustrates nicely the relevant features present in the more general situation.

In order to proceed, it is necessary to clarify what we mean by a vector field on a simplicial complex that fails to be a manifold. This is a nontrivial issue: for example, in the case of a graph, the tangent space to a vertex with incidence number greater than two is not well defined. We defer a more detailed discussion of these statements to Appendix A. The essential difference is that we construct *semiflows*—flows which possess unique forward orbits.

2.2. Edge point fields for single AGV control. In the context of describing and executing *patterns* or periodic motions on a graph, one desires a set of building blocks for moving from one goal to the next. We introduce the terminology and philosophy for constructing patterns by way of the simplest possible examples: a single AGV on a graph. This avoids the additional topological complications present in the context of cooperative motion.

To this end, we introduce the class of *edge point fields* as a dynamical toolbox for a hybrid controller. Given a specified goal point $g \in e_j$ within an edge of Γ, an *edge point field* is a locally defined vector field X_g on Γ with the following properties:

Locally defined. X_g is defined on a neighborhood $\mathcal{N}(e_j)$ of the goal edge e_j within the graph topology, and forward orbits under X_g are uniquely defined.

Point attractor. Every forward orbit of X_g asymptotically approaches the unique fixed point $g \in e_j$.[2]

Navigation-like. X_g admits a C^0 Lyapunov function, $\Phi_g : \Gamma \to R$.

The following existence lemma (whose trivial proof we omit) holds.

LEMMA 1. *Given any edge $e_j \subset \Gamma$ which is contractible within Γ, there exists an edge point field X_g for any desired goal $g \in e_j$.*

As a remark, we note that, as is usual in the traditional dynamical systems settings, the orbits of an edge point field may take an infinite amount of time to reach their destination. We can always rectify this situation by modifying the flow in a neighborhood of the goal via a sublinear term, e.g., $\dot{x} = -x^{1/3}$. This comment applies to vector fields used throughout the remainder of this work.

2.3. Discrete regulation of patterns. By an excursion on a graph, we mean a (possibly infinite) sequence of edges from the graph, $E = e_{i_1} \dots e_{i_N} \dots \in \mathcal{E}^Z$, having

[2] When it is not clear from the context, we shall denote the goal point achieved by an edge point flow as $\mathbf{g}(X_g) = \{g\}$.

the property that each pair of contiguous edges e_{i_j} and $e_{i_{j+1}}$ share a vertex in common. The set of excursions forms a language, $\mathcal{L}$, the so-called *subshift* on the alphabet defined by the named edges (we assume each name is unique) [13]. Given a legal block, $B = e_{i_1} \dots e_{i_M} \in \mathcal{L}$, we say that an excursion realizes that pattern if its periodic extension eventually reaches the "goal" $BBBBB\dots$ under the iterates of the block shift. In other words, after some transient behavior, the excursion consists of repetitions of the block B (terminating possibly with the empty edge).

In a previous paper [5], Burridge, Rizzi, and Koditschek introduced a very simple but effective discrete event controller for regulating patterns on abstract graphs representing a "prepares" relation imposed on families of controllers over general smooth manifolds. We introduce this prepares relation below and prune it as in [5]. The resulting ordering on the controllers yields a controller transition logic that enlarges the basin of any one member of the family to include the union of all "higher" controllers. This simple idea has a much longer history. It was in introduced in robotics as "preimage backchaining" [14], pursued in [15] as a method for building verifiable hardened automation via the metaphor of a family of funnels, and pursued in [7] as a means of prescribing sensor specifications from goals and action sets. In the discrete event systems literature, an optimal version of this procedure has been introduced in [4], and a generalization has recently been proposed in [18].

Let $\mathcal{E}^0{:=}B \subset \mathcal{E}$ denote the edges of Γ that appear in the block of letters specifying the desired pattern. Denote by

$$\mathcal{E}^{n+1} \subset \mathcal{E} - \bigcup_{k \le n} \mathcal{E}^k$$

those edges that share a vertex with an edge in $\mathcal{E}^n$ but are not in any of the previously defined subsets. This yields a finite partition of $\mathcal{E}$ into "levels," $\{\mathcal{E}^p\}_{p=0}^{P}$, such that for each edge, $e_i^p \in \mathcal{E}^p$, there can be found a legal successor edge, $e_j^{p-1} \in \mathcal{E}^{p-1}$, such that $e_i^p e_j^{p-1} \in \mathcal{L}$ is a legal block in the language. Note that we have implicitly assumed that $\mathcal{E}^0$ is reachable from the entire graph—otherwise, there will be some "leftover" component of $\mathcal{E}$ forming the last cell in the partition starting within which it is not possible to achieve the pattern. Note as well that we impose some ordering of each cell $\mathcal{E}^p = \{e_i^p\}_{i=1}^{M_p}$: the edges of $\mathcal{E}^0 = B$ are ordered by their appearance in the block; the ordering of edges in higher level cells is arbitrary.

We may now define a "graph control" law $G{:}\mathcal{E} \to \mathcal{E}$ as follows. From the nature of the partition $\{\mathcal{E}^p\}$ above, it is clear that the least legal successor function,

$$L(p,i){:=}\begin{cases} i+1 \bmod M & : p = 0, \\ \min\{j \le M_p : e_i^p e_j^{p-1} \in \mathcal{L}\} & : p > 0 \end{cases} \tag{2.2}$$

is well defined. From this, we construct the graph controller:

$$G(e_i^p){:=}e_{L(p,i)}^{p-1}. \tag{2.3}$$

It follows almost directly from the definition of this function that its successive application to any edge leads eventually to a repetition of the desired pattern.

PROPOSITION 2. *The iterates of G on $\mathcal{E}$ achieve the pattern B.*

2.4. Hybrid edge point fields. A semiflow, $(X)^t$, on the graph induces excursions in $\mathcal{L}$ parametrized by an initial condition as follows. The first letter corresponds to the edge in which the initial condition is located. (Initial conditions at vertices

are assigned to the incident edge along which the semiflow points.) The next letter is added to the sequence by motion through a vertex from one edge to the next.

We will say of two edge point fields X_1, X_2 on a graph, Γ, that X_1 *prepares* X_2, denoted $X_1 \succ X_2$, if the goal of the first is in the domain of attraction of the second, $\mathsf{g}(X_1) \subset \mathcal{N}(X_2)$. Given any finite collection of edge point fields on Γ, we will choose some $0 < \alpha < 1$ and assume that their associated Lyapunov functions have been scaled in such a fashion that $X_1 \succ X_2$, implies $(\Phi_1)^{-1}[0, \alpha] \subset \mathcal{N}(X_2)$. In other words, an α crossing of the trajectory $\Phi_1 \circ (X_1)^t$ signals arrival in $\mathcal{N}(X_2)$.

Suppose now that for every edge in some pattern block, $e_i^0 \in \mathcal{E}^0$, there has been designated a goal point g_i^0 along with an edge point field X_i^0 taking that goal: $\mathsf{g}(X_i^0) = g_i^0$. Assume as well that the edge point field associated with each previous edge in the pattern prepares the flow associated with the next edge; in other words, using the successor function (2.2), we have

$$\mathsf{g}\left(X_j^0\right) \subset \mathcal{N}\left(X_{L(p,j)}^0\right).$$

Now construct edge point fields on all the edges of Γ such that the tree representation of their $\succ$ relations is exactly the tree pruned from the original graph above:

$$\mathsf{g}\left(X_j^p\right) \subset \mathcal{N}\left(X_{L(p,j)}^{p-1}\right).$$

We are finally in a position to construct a hybrid semiflow on Γ. This feedback controller will run the piecewise smooth vector field, $\dot{x} = X$, as follows:

$$X := \begin{cases} X_j^p & : x \in e_j^p \text{ and } \Phi_j^p > \alpha, \\ X_{L(p,j)}^{p-1} & : x \in e_{L(p,j)}^{p-1} \text{ or } \Phi_j^p \leq \alpha. \end{cases} \tag{2.4}$$

It is clear from the construction that progress from edge to edge of the state of this flow echoes the graph transition rule G constructed above.

PROPOSITION 3. *The edge transitions induced by the hybrid controller* (2.4) *are precisely the iterates of the graph map* G (2.3) *in the language* $\mathcal{L}$.

3. The Y-graph. We now turn our attention to the safe control of multiple AGVs on a graph work space via vector fields. Whereas the case of a single AGV on a graph could be controlled by vector fields on the graph itself, the safe coordination of multiple agents necessitates vector field controls on the appropriate configuration space—a space whose topological features are by no means obvious.

For the remainder of this work, we consider the simplest example of a nontrivial configuration space: that associated with the *Y-graph*, Υ, having four vertices $\{v_i\}_0^3$ and three edges $\{e_i\}_1^3$. Each edge e_i attaches a vertex v_i to the central vertex v_0. Although this is a simple scenario compared to what one finds in a typical setting, there are several reasons why this example is in many respects canonical.

1. *Simplicity.* Any graph may be constructed by gluing index-K radial graphs together for various K. The $K = 3$ model we consider is the simplest nontrivial case and is instructive for understanding the richness and challenges of local cooperative dynamics on graphs.
2. *Genericity with respect to graphs.* Graphs which consist of copies of Υ glued together, the *trivalent graphs*, are generic: any nontrivial graph may be perturbed in a neighborhood of the vertex set so as to be trivalent. For example, the 4-valent graph resembling the letter "X" may be perturbed slightly to

resemble the letter "H"—a trivalent graph. An induction argument shows that this is true for all graphs. Hence the dynamics on an arbitrary graph are approximated by patching together dynamics on copies of Υ.

3. *Genericity with respect to local dynamics.* Finally, pairwise local AGV interactions on an arbitrary graph restrict themselves precisely to the dynamics of two agents on Υ as follows. Given a vertex v of a graph Γ, assume that two AGVs x and y are on different edges e_1 and e_2 incident to v and moving toward v with the goal of switching positions. A collision is imminent unless one AGV "moves out of the way" onto some other edge e_3 incident to v. The local interactions thus restrict themselves to dynamics of a pair of AGVs on the subgraph defined by $\{v; e_1, e_2, e_3\}$. Hence the case we treat in this paper is the generic scenario for the local resolution of collision singularities in cooperative dynamics on graphs and forms a basis for decentralized control of large numbers of independent agents.

3.1. Intrinsic coordinates. The configuration space $\mathcal{C}$ of two points on Υ is a subset of the cartesian product $\Upsilon \times \Upsilon$. Since Υ (and indeed any graph which is physically relevant to the setting of this paper) is embedded in a factory floor or ceiling and thus planar, the configuration space $\mathcal{C}$ embeds naturally in R^4. We wish to modify this embedding to facilitate both analysis on and visualization of the configuration space. We will present alternate embeddings in both higher and lower dimensional Euclidean spaces for these purposes.

We begin by representing the configuration space within a higher dimensional Euclidean space via *intrinsic* coordinates—coordinates independent of the graphs embedded in space. We illustrate this coordinate system with the Y-graph Υ, noting that a few simple modifications yield coordinate schemes for general graphs.

Let $\{e_i\}_1^3$ denote the three edges in Υ, parametrized so that the closure of each edge e_i is identified with $[0,1]$ oriented so that $[0,1]$ is mapped to $[v_0, v_i]$. Any point in Υ is thus given by a vector x in the $\{e_i\}$ basis whose magnitude $|x| \in [0,1]$ determines the position of the point in the e_i direction. For $|x| > 0$, denote by $\iota(x)$ the value of i so that $x = |x| e_{\iota(x)}$. This parameterization embeds Υ as the positive unit axis frame in R^3. Likewise, a point in $\mathcal{C}$ is given as a pair of distinct vectors (x, y), i.e., as the positive unit axis frame in R^3 cross itself sitting inside of $R^3 \times R^3 \cong R^6$. We have thus embedded the configuration space of two distinct points on Υ in the positive orthant of R^6. It is clear that one can embed the more general configuration space of N points on Υ in R^{3N} in this manner.

This coordinate system is particularly well suited to describing vector fields on $\mathcal{C}$ and implementing numerical simulations of dynamics, as the coordinates explicitly track the physical position of each point on the graph.

3.2. A topological analysis. Visualizing $\mathcal{C}$ as a subset of R^4 or R^6 is unenlightening. More useful for visualization purposes is the following construction which embeds $\mathcal{C}$ within R^3.

THEOREM 1. *The configuration space $\mathcal{C}$ associated with a pair of AGVs restricted to the Y-graph Υ is homeomorphic to a punctured disc with six 2-simplices attached as per Figure* 3.1.

Proof. Recall that $\mathcal{C}$ consists of pairs of distinct vectors (x, y) in intrinsic coordinates. Restrict attention to the subspace $\mathcal{D} \subset \mathcal{C}$ defined by

$$\mathcal{D} := \{(x, y) \in \mathcal{C} : \iota(x) \neq \iota(y)\}, \tag{3.1}$$

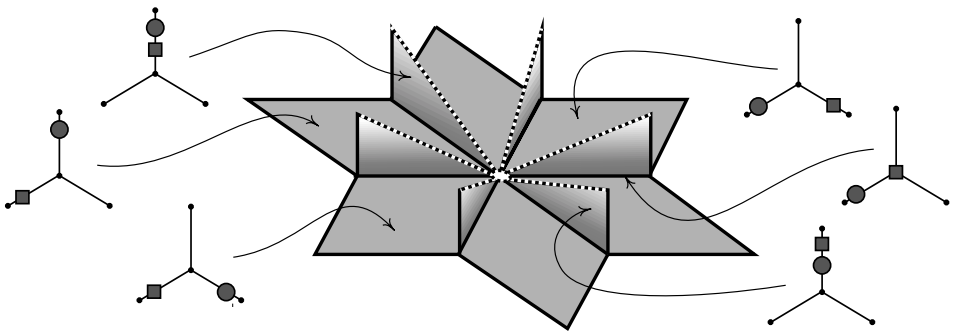


FIG. 3.1. *The configuration space $\mathcal{C}$ embedded in R^3. Dashed lines refer to open boundaries; sample configurations for representative 2-cells are illustrated to the sides.*

where an undefined index is considered to be not equal to one which is defined. Thus $\mathcal{D}$ consists of configurations for which both AGVs do not occupy the same edge interior.

The set $\mathcal{D}$ has a cellular decomposition as follows. There are 2 AGVs and 3 edges in Υ; hence there are $3 \cdot 2 = 6$ cells $\mathcal{D}_{i,j} \subset \mathcal{D}$, where $i := \iota(x) \neq \iota(y) =: j$. Since (the closure of) each edge in Υ is homeomorphic to $[0,1]$ (determined by $|\cdot|$), the cell $\mathcal{D}_{i,j}$ is homeomorphic to $([0,1] \times [0,1]) - \{(0,0)\}$, where, of course, the origin $(0,0)$ is removed as it belongs to the diagonal Δ. A path in $\mathcal{D}$ can move from cell to cell only along the subsets where the index of one AGV changes, e.g., $|x| = 0$ or $|y| = 0$. Thus the edges $\{0\} \times (0,1]$ and $(0,1] \times \{0\}$ of the punctured square $\mathcal{D}_{i,j}$ are attached, respectively, to $\mathcal{D}_{k,j}$ and $\mathcal{D}_{i,k}$, where k is the unique index not equal to i or j.

Furthermore, each 2-cell $\mathcal{D}_{i,j}$ has a product structure as follows: decompose $\mathcal{D}_{i,j}$ along the lines of constant $\theta := \tan^{-1}(\frac{|y|}{|x|})$. It is clear that θ is the angle in the unit's first quadrant in which $\mathcal{D}_{i,j}$ sits. Hence each $\mathcal{D}_{i,j}$ is decomposed into a product of a closed interval $S_{i,j} := \theta \in [0, \pi/2]$ (an "angular" coordinate) with the half-open interval $(0,1]$ (a "radial" coordinate). As this product decomposition is respected along the gluing edges, we have a decomposition of all of $\mathcal{D}$ into the product of $(0,1] \times S$, where S is a cellular complex given by gluing the six segments $S_{i,j}$ end-to-end cyclically along their endpoints. The set S is a 1-manifold without boundary since each $S_{i,j}$ is a closed interval, each of whose endpoints is glued to precisely one other $S_{i,j}$. Hence, by the classification of 1-manifolds, S is homeomorphic to a circle. We have thus decomposed $\mathcal{D}$ as the cross product of a circle with $(0,1]$—a punctured unit disc.

The complement of $\mathcal{D}$ in $\mathcal{C}$ consists of those regions where $\iota(x) = \iota(y)$. For each $i = 1 \ldots 3$, the subset of $\mathcal{C}$ where $\iota(x) = \iota(y) = i$ is homeomorphic to $((0,1] \times (0,1]) - \{|x| = |y|\}$: this consists of two disjoint triangular "fins." A total of six such fins are thus attached to $\mathcal{D}$ along the six edges where $|x|$ or $|y| = 0$. In the coordinates of the product decomposition for $\mathcal{D}$, these fins emanate along the radial lines where θ equals zero or $\pi/2$, yielding the topological space illustrated in Figure 3.1. □

COROLLARY 4. *Given any point goal $g \in \mathcal{D} \subset \mathcal{C}$, there exists an explicit navigation function (of class piecewise real-analytic) generating a semiflow which sends all but a measure-zero set of initial conditions to g under the gradient semiflow.*

Proof. The subset $\mathcal{D} \subset \mathcal{C}$ is homeomorphic to a punctured disc $S \times (0,1]$ and may easily be compactified to an annulus with boundary $S \times [\epsilon, 1]$ by removing an open neighborhood of the diagonal. Then the conditions for the theorems of Koditschek and Rimon [12] are met since an annulus is a *sphereworld*. Hence not only does a

navigation function Φ on this subspace exist, but an explicit procedure for determining Φ is given [12]. One may then extend Φ to the remainder of $\mathcal{C}$ as follows: choose a point (x, y) on the fin, and define

$$\Phi(x,y) := \begin{cases} \frac{1}{1-|x|}\Phi(0,y), & |x| < |y|, \\ \frac{1}{1-|y|}\Phi(x,0), & |y| < |x|, \end{cases} \tag{3.2}$$

so that Φ increases sharply along the fins.[3] This directs the gradient flow to monotonically "descend" away from the diagonal and onto $\mathcal{D}$. Note that $\mathcal{D}$ is forward-invariant under the dynamics and that, upon prescribing the vector field on the fins to point into $\mathcal{D}$, we have defined a semiflow and hence a well-defined navigational procedure. □

This result is very satisfying in the sense that it guarantees a navigation function by applying existing theory to a situation which, from the definition alone, would not appear to be remotely related to a sphereworld. However, a deeper analysis of configuration spaces of graphs [9] reveals that, for more than two AGVs, the configuration space of a graph is never a sphereworld.[4]

We thus consider alternate methods for realizing compatible goals by means of a vector field on the configuration space, focusing, in particular, on the use of attracting periodic orbits as a controller component in the "toolbox" for building up the sort of hybrid feedback laws to be considered carefully in section 4.

3.3. Example: A circulating flow. We begin with a simple example of a vector field on $\mathcal{C}$ which possesses an attracting limit cycle as a goal. This "circulating field," which cycles a pair of AGVs through states on the boundary of $\mathcal{D} \subset \mathcal{C}$, is a canonical example of (1) a meaningful semiflow with limit cycle and (2) a practical field for implementing collision avoidance in a hybrid controller (cf. item 3 in the preface to section 3). Figure 3.2 (right) illustrates the flow restricted to $\mathcal{D}$.

THEOREM 2. *There exists a piecewise-smooth vector field X on $\mathcal{C}$ which has the following properties:*

1. *X defines a nonsingular semiflow on $\mathcal{C}$.*
2. *The diagonal Δ is repelling with respect to X.*
3. *Every orbit of X approaches a unique attracting limit cycle on $\mathcal{C}$ which cycles through all possible ordered pairs of distinct edge states.*

Proof. Recall that $\mathcal{D}$ denotes that portion of the configuration space corresponding to a placement of the AGVs on distinct edges of the graph; from the proof of Theorem 1, $\mathcal{D}$ is homeomorphic to a punctured disc. The intrinsic coordinates on the configuration space $\mathcal{C}$ are illustrated in Figure 3.2 (left), where only $\mathcal{D}$ is shown for simplicity. The reader should think of this as a collection of six square coordinate planes, attached together pairwise along axes with the origin removed.[5] The six triangular fins are then attached as per Figure 3.1.

Recall that any point in the graph is represented as a vector $x = |x|e_i$ for some i. Denote by $\hat{e}_i$ the unit tangent vector in each tangent space $T_x e_i$ pointing in the

[3] This construction does not satisfy the formal requirements for a "navigation function" since it is not bounded on the closure of the configuration space. There is a straightforward procedure detailed in [12] that can be used to complete the construction.

[4] Any configuration space of any graph is *aspherical*; there are no essential closed spheres of dimension larger than one, in contrast to a sphereworld. Thus, although a navigation is guaranteed to result in any case, the explicit constructions [12] are inapplicable.

[5] In the natural product metric on $\mathcal{C}$, these six 2-cells are flat Euclidean squares.

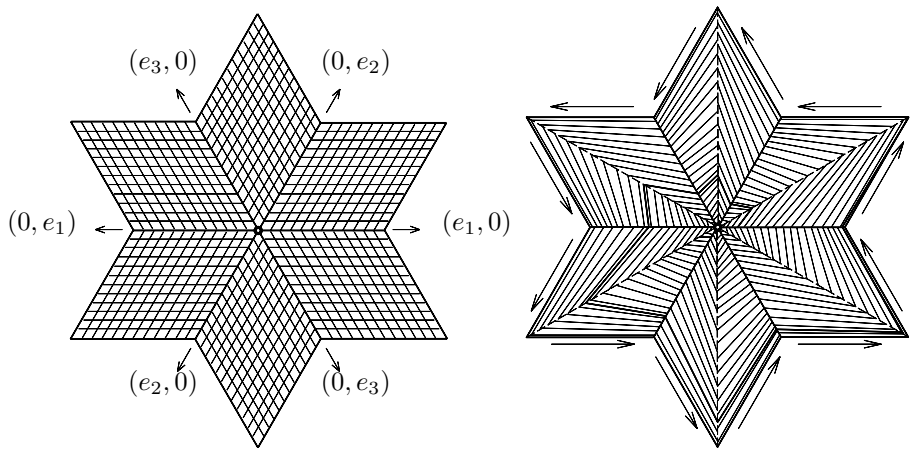


FIG. 3.2. *Left: The coordinate system on the unfinned region $\mathcal{D}$ of $\mathcal{C}$. Right: The circulating flow with a typical orbit.*

positive (outward) direction toward the endpoint v_i. The vector field we propose is the following. Given $(x, y) \in \mathcal{C}$,

1. if $\iota(x) = \iota(y)$, then

$$\left.\begin{array}{l} \dot{x} = -|y|\hat{e}_{\iota(x)}, \\ \dot{y} = |y|(1-|y|)\hat{e}_{\iota(y)}, \end{array}\right\} \quad 0 < |x| < |y|,$$
$$\left.\begin{array}{l} \dot{x} = |x|(1-|x|)\hat{e}_{\iota(x)}, \\ \dot{y} = -|x|\hat{e}_{\iota(y)}, \end{array}\right\} \quad 0 < |y| < |x|; \tag{3.3}$$

2. if $\iota(x) = \iota(y) + 1$ or $|x| = 0$, then

$$\left.\begin{array}{l} \dot{x} = |y|\hat{e}_{(\iota(y)+1)}, \\ \dot{y} = |y|(1-|y|)\hat{e}_{\iota(y)}, \end{array}\right\} \quad 0 \le |x| < |y|,$$
$$\left.\begin{array}{l} \dot{x} = |x|(1-|x|)\hat{e}_{\iota(x)}, \\ \dot{y} = -|x|\hat{e}_{\iota(y)}, \end{array}\right\} \quad 0 < |y| \le |x|; \tag{3.4}$$

3. if $\iota(y) = \iota(x) + 1$ or $|y| = 0$, then

$$\left.\begin{array}{l} \dot{x} = -|y|\hat{e}_{\iota(x)}, \\ \dot{y} = |y|(1-|y|)\hat{e}_{\iota(y)}, \end{array}\right\} \quad 0 < |x| \le |y|,$$
$$\left.\begin{array}{l} \dot{x} = |x|(1-|x|)\hat{e}_{\iota(x)}, \\ \dot{y} = |x|\hat{e}_{(\iota(x)+1)}, \end{array}\right\} \quad 0 \le |y| < |x|. \tag{3.5}$$

Note that all addition operations on $\iota(x)$ and $\iota(y)$ are performed mod three.

The vector field is nonsingular as follows: if $|x||y| \neq 0$, then the vector field is by inspection nonsingular. If $|x| = 0$, then $|y| > 0$ since the points are distinct. It then follows from (3.4) that the vector field on this region has $d|x|/dt = |y| \neq 0$. A similar argument holds for the case where $|y| = 0$.

The vector field defines a semiflow as follows: on those regions where $0 \neq |x| \neq |y| \neq 0$, the vector field is smooth and hence defines a true flow. Along the lines where $|x| = |y|$, the vector field is not smooth but nevertheless is constructed so as to define unique solution curves; hence the region $\mathcal{D}$, where $\iota(x) \neq \iota(y)$, is invariant under the flow. Finally, along the branch line curves where $|x| = 0$ or $|y| = 0$, the vector field

points into the branch lines from the fins, implying that the dynamics is a semiflow (see the remarks in Appendix A).

This vector field admits a C^0 Lyapunov function $\Phi : \mathcal{C} \to [0, 1)$ of the form

$$\Phi(x,y) := \begin{cases} 1 - |(|x| - |y|)| & : \iota(x) = \iota(y), \\ 1 - \max\{|x|, |y|\} & : \iota(x) \neq \iota(y). \end{cases} \tag{3.6}$$

From (3.3), one computes that on the fins (where $\iota(x) = \iota(y)$),

$$\frac{d\Phi}{dt} = -\left|\left(\frac{d|x|}{dt} - \frac{d|y|}{dt}\right)\right| < 0 \tag{3.7}$$

since here $|x| \neq |y|$. Furthermore, on the disc $\mathcal{D}$ ($\iota(x) \neq \iota(y)$), Φ changes as $\frac{d\Phi}{dt} = \Phi(\Phi - 1)$. Hence Φ strictly decreases off of the boundary of the disc

$$\partial\mathcal{D} := \{(x,y) : |x| = 1 \text{ or } |y| = 1\} = \Phi^{-1}(0). \tag{3.8}$$

It follows from the computation of $d\Phi/dt$ that the diagonal set Δ of $\Upsilon \times \Upsilon$ is repelling and that the boundary cycle $\partial\mathcal{D}$ is an attracting limit cycle. □

This example illustrates how one can use a relatively simple vector field on the configuration space to construct a pattern which is free from collisions. Indeed, as part of a hybrid control scheme, one could use this circulating flow to resolve potential collisions between AGVs in a general setting by localizing the dynamics near a pairwise collision to those on a trivalent subgraph. In practice, the fact that the outer vertices of the Y-graph are never quite reached by an interior orbit is irrelevant: a near-approach suffices for any practical application.

4. Patterns and vector fields for monotone cycles. In this section, we consider the problem of constructing vector fields which are tuned to trace out specific collision-free patterns—scheduled series of visits to specific work stations by the pair of AGVs whose regularity we wish to achieve at steady state, and return back to from any temporary perturbation or disruption. We begin with a specification of a suitable language for describing patterns.

4.1. A grammar for patterns. The setting we envisage is as follows: the three ends of the graph Υ are stations at which an AGV can perform some function. The AGV pair is required to execute an ordered sequence of functions, requiring an interleaved sequence of visitations. In order to proceed with vector field controls for cooperative patterns, it is helpful to construct the appropriate symbolic language, as introduced in section 2 for single AGV systems. Denote the pair of AGV states as x and y, respectively. Also, denote the three docking stations as vertices v_1 through v_3 as in Figure 3.1. The grammar $\mathcal{G}$ we use is defined as follows:

- (**xi**): These represent configurations for which the AGV x is docked at the vertex v_i, $i = 1 \ldots 3$. The AGV y is at an unspecified undocked position.
- (**yi**): These represent configurations for which the AGV y is docked at the vertex v_i, $i = 1 \ldots 3$. The AGV x is at an unspecified undocked position.
- (**xiyj**): These represent configurations for which the AGV x is docked at vertex v_i, while the AGV y is simultaneously docked at the vertex v_j, $j \neq i$.

For example, the word (**x1**)(**y2**)(**x3y2**) executes a sequence in which the first AGV docks at Station v_1 and then undocks while the second AGV docks at Station v_2. Finally, the AGVs simultaneously dock at Stations v_3 and v_2, respectively.

As we have assumed from the beginning, the one-dimensional nature of the graph-constraints precludes the presence of multiple agents at a single docking station; hence there are exactly twelve symbols in the grammar $\mathcal{G}$. From this assumption, it follows that particular attention is to be paid to those trajectories which do not make excursions onto the "fins" of the configuration space. It is obvious from the physical nature of the problem that planning paths which involve traveling on the fins is not a locally optimal trajectory with respect to minimizing distance or elapsed time. It suffices to say that we restrict our attention for the moment to trajectories and limit cycles for patterns, in particular, which are constrained to the region $\mathcal{D} \subset \mathcal{C}$.

We identify each symbol with a region of the boundary of the unbranched portion of $\mathcal{C}$; namely, $\partial\mathcal{D}$ is partitioned into twelve *docking zones* as in Figure 4.1. Note further that there is a cyclic ordering, $\prec$, on $\mathcal{G}$ induced by the orientation on the boundary of the disc along which the zones lie. By a cyclic ordering, we mean a way of determining whether a point q lies between any ordered pair of points (p_1, p_2).

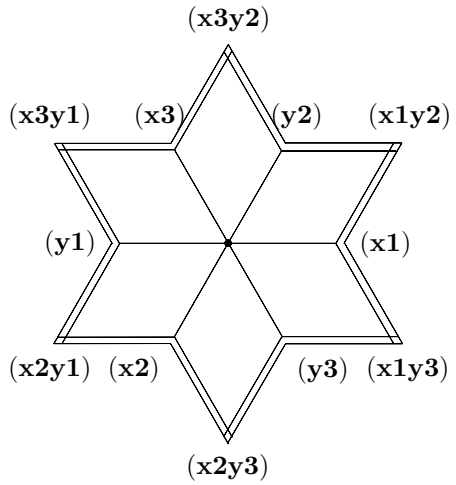


FIG. 4.1. *Labels for the cyclically ordered grammar* $\mathcal{G}$.

We proceed with the analysis of limit cycles on $\mathcal{C}$. Consider the class of *pattern vector fields*, $\mathcal{X}_P$, on $\mathcal{C}$ defined as follows. For every $X \in \mathcal{X}_P$,

1. X defines a semiflow on $\mathcal{C}$ and a true flow off the nonmanifold set of $\mathcal{C}$;
2. there is a unique limit cycle γ which is attracting and which traces out a nonempty word in the grammar $\mathcal{G}$;
3. the diagonal set Δ is a repeller with respect to X;
4. there are no fixed invariant sets of X which attract a subset of positive measure save γ.

Denote by $\mathcal{X}_M$ the subset of $\mathcal{X}_P$ for which the limit cycle, γ, lies in $\mathcal{D}$. The question of which words in the grammar $\mathcal{G}$ are admissible for the class $\mathcal{X}_P$ has a simple answer in terms of the cyclic ordering $\prec$. A word $\mathbf{w}$ composed of elements $\mathbf{w} = w_1 w_2 \dots w_n$ in the grammar $\mathcal{G}$ said to be *monotone* with respect to the cyclic ordering $\prec$ if $w_{i-1} \prec w_i \prec w_{i+1}$ for every i (index operations all mod n).

THEOREM 3. *Within the class of vector fields* $\mathcal{X}_M$*, the limit cycles trace out monotone words in the cyclically ordered grammar* $(\mathcal{G}, \prec)$.

Proof. The idea of the proof is simple and follows from the observation that any limit cycle of the flow must be embedded (the curve does not intersect itself). After a small perturbation, one may assume that the boundary zone $\partial\mathcal{D}$ is visited by γ in

a finite number of points, $Q := \gamma \cap \partial\mathcal{D}$. Consider two points $p, q \in Q$, which are consecutive in the limit cycle: that is, there is an embedded subarc $\alpha \subset \gamma$ which connects p to q within the interior of $\mathcal{D}$. The arc α separates $\mathcal{D}$ into two topological discs (this is the Jordan curve theorem [17]); hence γ must lie entirely within the closure of one of these discs. This implies that the limit cycle cannot visit any point $x \in \partial\mathcal{D}$ satisfying $p \prec x \prec q$. Repeating this argument for all pairs of consecutive points yields the monotonicity property. □

Although the only admissible words in the grammar $\mathcal{G}$ are those which are monotone, it is possible to realize many if not all of the nonmonotone cycles as limit cycles for a semiflow on the *full* configuration space $\mathcal{C}$; one must design the semiflow so as to utilize the fins for "jumping" over regions of $\mathcal{D}$ cut off by the limit cycle. Such vector fields quickly become very convoluted, even for relatively simple nonmonotone limit cycles, and a more explicit constructive procedure would need stronger motivation from the application domain than we are presently aware of.

4.2. Isotopy classes of limit cycles. Given a limit cycle γ which traces out a pattern by visiting the boundary zone $\partial\mathcal{D}$ in the ordered set $Q \subset \partial\mathcal{D}$, one wants to know which other limit cycles minimize a given performance functional while still visiting Q in the proper sequence. The mathematical framework for dealing with this problem is the notion of *isotopy classes* of curves.

Two subsets A_0 and A_1 of a set B are said to be (ambiently) *isotopic rel* C (where $C \subset B$) if there exists a continuous 1-parameter family of homeomorphisms $f_t : B \to B$ such that

1. f_0 is the identity map on B,
2. $f_1(A_0) = A_1$, and
3. $f_t|_C$ is the identity map on C for all t.

As t increases, f_t deforms B, pushing A_0 to A_1 without cutting or tearing the spaces and without disturbing C.

There are two ways in which optimization questions relate to isotopy classes of limit cycles: (1) Given an element of the grammar $\mathcal{G}$, in which isotopy class (rel the docking zones) of curves does an optimal limit cycle reside? (2) Within a given isotopy class of cycles rel Q, which particular cycle is optimal?

For a monotone limit cycle on $\mathcal{D}$, question (1) focuses on the location of the cycle with respect to the central point $(0, 0)$, which is deleted from the disc $\mathcal{D}$. It is a standard fact from planar topology that every curve in the punctured disc has a well-defined *winding number*, which measures how many times the cycle goes about the origin, and, furthermore, that this number is -1, 0, or 1 if the cycle is an embedded curve. This winding number determines the isotopy class of the curve in $\mathcal{D}$. Hence the problem presents itself as follows: given an element of the grammar $\mathcal{G}$, which isotopy class rel the docking zones is optimal (with respect to any or all of the functionals defined)? Is the winding number zero or nonzero?[6]

To address this question, we define the *gap angles* associated to a limit cycle. For the remainder of this section, we will place standard polar coordinates on the region $\mathcal{D}$ (given as a subset of the plane as per Figures 3.2 and 4.1) with the central puncture corresponding to the origin. Given a set of "docked states"—or points $Q = \{q_1, q_2, \ldots, q_J\}$ ordered with respect to time—we define the gap angles to be the successive differences in the angular coordinates of the q_j: thus $\angle_j := P(q_{j+1}) - P(q_j)$, where P denotes projection of points in $\mathcal{D}$ onto their angular coordinates, and

[6]The difference between $+1$ and -1 is the orientation of time.

subtraction is performed with respect to the orientation on $\partial\mathcal{D}$.

For simplicity, we consider the optimization-isotopy problem in the case of a discrete cost functional $\mathbf{W}_d$, defined to be the intersection number of the path with the branch locus of $\mathcal{C}$, i.e., the number of times an AGV occupies the central vertex (the shared resource in the problem). Similar arguments are possible for other natural performance metrics.

PROPOSITION 5. *Given a cyclically ordered set of points $Q = \{q_j\}_1^J$ on the boundary of $\mathcal{D}$, consider the class of embedded monotone cycles on $\mathcal{D}$ which trace out the points of Q.*

1. *There is a $\mathbf{W}_d$-minimizing embedded monotone cycle on $\mathcal{D}$ having winding number zero with respect to the origin if there is a gap angle greater than π.*
2. *Conversely, if there are no gap angles greater than π, then there is a $\mathbf{W}_d$-minimizing embedded cycle of index ± 1.*

Proof. Define the gap angles $\{\angle_j\}_1^J$ to be the differences of the angles between the points q_j and q_{j+1} (in standard planar polar coordinates with all indices mod J). Since $\sum_j \angle_j = 2\pi$, there can be at most one gap angle greater than π. To simplify the problem, use a 1-parameter family P_t of maps from the identity P_0 to the projection $P = P_1$, which deforms $\mathcal{D}$ to the boundary circle $S := \partial\mathcal{D}$ by projecting along radial lines. The index of a curve on $\mathcal{D}$ is invariant under this deformation as is the functional $\mathbf{W}_d$.

Denote by γ_j the subarc of γ between points q_j and q_{j+1} (all indices mod J). Denote by α_j the subarc of the boundary S between points q_j and q_{j+1}, where the arc is chosen to subtend the gap angle $\angle_j$. Since the boundary curve $S = \cup_j \alpha_j$ is a curve of index ± 1, the arcs γ_j and α_j are isotopic in $\mathcal{D}$ rel their endpoints for all j if and only if γ is a curve of index ± 1.

Assume first that there is a gap angle $\angle_j > \pi$ with γ an index ± 1 curve on S which intersects the branch angles $\Theta = \{n\pi/3 : n \in Z\}$ in a minimal number of points among all other closed curves on S which visit the points Q in the specified order. It follows that the arc $P(\gamma_j)$ subtends an angle greater than π and thus increments $\mathbf{W}_d$ by at least three. One may replace γ_j by a curve γ_j', which substitutes for the arc γ_j, one which wraps around "the other way" monotonically. This changes the index of γ from nonzero to zero since the arc γ_j' is no longer isotopic to α_j. Also, it is clear that this either decreases the number of intersections with Θ or leaves this number unchanged.

We must show that the replacement arc γ_j' can be chosen in such a way that it does not intersect the remainder of γ. However, since γ is a curve of index ± 1, we may isotope each arc γ_i to the boundary curve α_i without changing the value of $\mathbf{W}_d$. Thus we may remove γ_j and replace it with the curve which is, say, a geodesic (in the natural metric geometry) from q_j to q_{j+1}. As this curve does not approach the boundary S apart from its ends, the new curve γ' is an embedded curve of index zero without an increase in $\mathbf{W}_d$.

Now assume, on the contrary, that γ is a $\mathbf{W}_d$ minimizer of index zero which has all gap angles strictly less than π. Then each arc from γ_i must intersect the branch set Θ in at most three components since, otherwise, the subtended arc would be in excess of $4\pi/3$. In the case where there exists an arc with exactly three intersections with the branch set, this arc may be replaced by an arc which goes around the singularity in the other direction without changing the number of intersections with the branch set (since there are a total of six branch lines); however, the index of the curve is toggled between zero and nonzero.

The final case is that in which each arc intersects the branch set in at most two places. However, since γ is a curve of index zero, some arc γ_j must not be isotopic to α_j. Hence the projection deformations P_t must push γ_j to a curve in the boundary S whose subtended gap angle is $2\pi - \angle_j > \pi$. Thus γ_j intersects the branch set in at least three places, yielding a contradiction. Replacing γ_j by the appropriate arc which is isotopic to α_j yields a $\mathbf{W}_d$-minimal cycle of nonzero index. □

4.3. Tuning cycles. Designing a customized "pattern" of two AGVs on the Y-graph is as simple as drawing a vector field on $\mathcal{C}$ with a stable limit cycle tracing out the desired motion. The problem then is how to specify such a vector field in coordinates. Since we focus on those limit cycles which are contained within $\mathcal{D}$, we can exploit the fact that $\mathcal{D}$ is topologically a punctured disc. We thus give an explicit coordinate-change between the natural polar coordinates on a disc and the intrinsic coordinates of section 3. Once we possess an explicit coordinate change (and its inverse), we can design a vector field in polar coordinates (an easy task to do in these coordinates) and then take the push-forward of the vector field under the coordinate change.

It will be convenient to keep track of which "wedge" of the annular region a point (r, θ) is. To do so, we introduce a parity function

$$P(\theta) := (-1)^{\{\lfloor 3\theta/\pi \rfloor + \lfloor 6\theta/\pi \rfloor\}}, \tag{4.1}$$

where $\lfloor t \rfloor$ is the integer-valued floor function. Recall the notation for the intrinsic coordinates for a point x on the graph Υ: $x = |x|\hat{e}_{\iota(x)}$, where $|x| \in [0, 1]$ is the distance from x to the central vertex, and $\hat{e}_{\iota(x)}$ is the unit tangent vector pointing along the direction of the $\iota(x)$-edge. Here the index $\iota(x)$ is an integer (defined modulo 3) and will be undefined in the case when $|x| = 0$, i.e., x is at the central vertex.

LEMMA 6. *The following is a piecewise-linear homeomorphism from the punctured unit disc in R^2 to the subset $\mathcal{D}$. Define $F(r, \theta) = (x, y)$, where*

$$\begin{aligned} \iota(x) &= \left\lfloor -\tfrac{3}{2\pi}(\theta - \pi) \right\rfloor \quad , \quad |x| = \begin{cases} r & \mathcal{P}(\theta) = +1, \\ r\left|\cot \tfrac{3}{2}\theta\right| & \mathcal{P}(\theta) = -1, \end{cases} \\ \iota(y) &= \left\lfloor -\tfrac{3}{2\pi}\theta \right\rfloor \quad , \quad |y| = \begin{cases} r\left|\tan \tfrac{3}{2}\theta\right| & \mathcal{P}(\theta) = +1, \\ r & \mathcal{P}(\theta) = -1. \end{cases} \end{aligned} \tag{4.2}$$

The inverse of this homeomorphism is given by $F^{-1}(x, y) = (r, \theta)$, where

$$\begin{aligned} \theta &= \begin{cases} \tfrac{2}{3}\tan^{-1}\tfrac{|y|}{|x|} - \tfrac{2\pi}{3}(\iota(y) + 1), & \iota(y) = \iota(x) + 1, \\ & \quad or\ |x| = 0, \\ -\tfrac{2}{3}\tan^{-1}\tfrac{|y|}{|x|} - \tfrac{2\pi}{3}(\iota(x) - 1), & \iota(x) = \iota(y) + 1, \\ & \quad or\ |y| = 0, \end{cases} \\ r &= \begin{cases} |x| & \mathcal{P}(\theta) = +1, \\ |y| & \mathcal{P}(\theta) = -1. \end{cases} \end{aligned} \tag{4.3}$$

Note that all θ values are defined modulo 2π, and all index values are integers defined modulo 3.

Proof. Begin by working on the region $\mathcal{D}_{1,2} \subset \mathcal{D}$, where $\iota(x) = 1$ and $\iota(y) = 2$. As noted earlier, this subspace is isometric to the positive unit square in R^2 with the origin removed. We need to map this to the subset $\{(r, \theta) : r \in (0, 1], \theta \in [0, \pi/3]\}$. The simplest such homeomorphism is to first shrink along radial lines, leaving the

angle invariant; hence

$$r = \begin{cases} |x| & : |x| \le |y|, \\ |y| & : |y| \le |x|. \end{cases} \tag{4.4}$$

Next, we squeeze the quarter-circle into a sixth of a circle by multiplying the angle by 2/3, leaving the radial coordinate invariant:

$$\theta = \frac{2}{3} \tan^{-1} \frac{|y|}{|x|}. \tag{4.5}$$

This gives the basic form of F^{-1} as per (4.3). To extend this to the remainder of $\mathcal{D}$, it is necessary to carefully keep track of $\iota(x)$ and $\iota(y)$ and subtract the appropriate angle from the computation of θ. Also, the condition of $|x| \le |y|$, etc., in (4.4) is incorrect on other domains of $\mathcal{D}$ since the inequalities flip as one traverses from square to square: the parity function $\mathcal{P}(\theta)$ keeps track of which "wedge" one is working on.

To determine F from F^{-1} is a tedious but unenlightening calculation, made more unpleasant by the various indices to be kept track of. Briefly, given r and θ on the first sixth of the unit disc, one knows from (4.4) that either $|x| = r$ or $|y| = r$, depending on whether θ is above or below $\pi/4$. To solve for the other magnitude, one inverts (4.5) to obtain $|y| = r\left|\tan \frac{3}{2}\theta\right|$ or $|x| = r\left|\cot \frac{3}{2}\theta\right|$, respectively. To generalize this to the other $\mathcal{D}_{i,j}$ domains of $\mathcal{D}$, it is necessary to take absolute values and to use the parity function $\mathcal{P}(\theta)$ as before. Finally, the computation of the index is obtainable from the combinatorics of the coordinate system as illustrated in Figure 3.2. □

For the design of limit cycles, it is easier to work on the polar disc and write out an explicit vector field $X = (\dot{r}, \dot{\theta})$ with a limit cycle. To transform this into intrinsic coordinates, one takes the push-forward of X with respect to F, obtaining the piecewise-smooth vector field

$$\begin{cases} \begin{pmatrix} |\dot{x}| = \dot{r}, \\ |\dot{y}| = \dot{r}\left|\tan(\frac{3}{2}\theta)\right| + \frac{3}{2} r\dot{\theta} \sec^2(\frac{3}{2}\theta) \end{pmatrix} & \mathcal{P}(\theta) = +1, \\ \begin{pmatrix} |\dot{x}| = \dot{r}\left|\cot(\frac{3}{2}\theta)\right| + \frac{3}{2} r\dot{\theta} \csc^2(\frac{3}{2}\theta), \\ |\dot{y}| = \dot{r} \end{pmatrix} & \mathcal{P}(\theta) = -1, \end{cases} \tag{4.6}$$

which simplifies to

$$\begin{cases} \begin{pmatrix} |\dot{x}| = \dot{r}, \\ |\dot{y}| = \dot{r}\dfrac{|y|}{|x|} + \dfrac{3}{2}\dot{\theta}\dfrac{|x|}{1 + \left(\frac{|y|}{|x|}\right)^2} \end{pmatrix} & \mathcal{P}(\theta) = +1, \\ \begin{pmatrix} |\dot{x}| = \dot{r}\dfrac{|x|}{|y|} + \dfrac{3}{2}\dot{\theta}\dfrac{|y|}{1 + \left(\frac{|x|}{|y|}\right)^2}, \\ |\dot{y}| = \dot{r} \end{pmatrix} & \mathcal{P}(\theta) = -1. \end{cases} \tag{4.7}$$

We present a more explicit example. Given a simple closed curve γ in R^2 which has nonzero winding number with respect to the origin, γ may be parametrized as $\{(r, \theta) : r = f(\theta)\}$ for some periodic positive function f. To construct a vector field on R^2 whose limit sets consist of the origin as a source and γ as an attracting limit cycle, it suffices to take the push-forward of the vector field $\dot{r} = r(1 - r),\ \dot{\theta} = \omega$ under the

planar homeomorphism $\phi : (r, \theta) \mapsto (f(\theta)r, \theta)$, which rescales linearly in the angular component. The calculations follow:

$$\begin{aligned} \phi_* \begin{pmatrix} \dot{r} \\ \dot{\theta} \end{pmatrix} &= D\phi \begin{pmatrix} \dot{r} \\ \dot{\theta} \end{pmatrix} \Big|_{r \mapsto \frac{r}{f}} = \begin{bmatrix} f & rf' \\ 0 & 1 \end{bmatrix} \begin{pmatrix} r(1-r) \\ \omega \end{pmatrix} \Big|_{r \mapsto \frac{r}{f}} \\ &= \begin{pmatrix} r\left(1 - \frac{r - f'\omega}{f}\right) \\ \omega \end{pmatrix}. \end{aligned} \tag{4.8}$$

Hence, given $f(\theta)$, we may tune a vector field to trace out the desired limit cycle and then use (4.2) and (4.3) to map it into intrinsic coordinates.

4.4. Optimal chords within a hybrid controller. To design optimal cycles with winding number zero, then we turn to constructing customized portions of limit cycles, or *chords* which can be pieced together via a state-actuated hybrid controller, much as in section 2. In other words, instead of building a simple fixed vector field with a limit cycle, we will use a set of vector fields which vary discretely in time and which may be pieced together so as to tune a limit cycle to the desired specifications. There is nothing in this construction which relies on the index-zero property, and thus these chords can be used to generate all monotone limit cycles on $\mathcal{C}$.

Let G denote a word representing a desired monotone limit cycle on the configurations space $\mathcal{C}$. Choose points $\{q_i\}$ on the boundary of $\mathcal{D}$ which correspond to the docking zones for the cycle given by G. Choose arcs α_i on $\mathcal{D}$ which connect q_i to q_{i+1} (using cyclic index notation). The arcs α_i are assumed given in the intrinsic coordinates on $\mathcal{D}$, as would be the case if one were determining a length-minimizing curve.

In the case where the limit cycle $\alpha := \cup_i \alpha_i$ is an embedded curve of nonzero index, the procedure of the previous subsection determines a vector field X_α on $\mathcal{C}$ which realizes α as an attracting limit cycle with the appropriate dynamics on the complementary region. Recall that one translates α to a curve on the disc model via the homeomorphism of (4.3). Then, representing the limit cycle α as a function $f_\alpha(\theta)$, one takes the vector field of (4.8) and, if desired, takes the image of this vector field under (4.7).

If, however, this is not the case, consider the arc α_j for a fixed j, and construct an index ± 1 cycle $\beta^j = \cup_i \beta^j_i$ which has docking zones $\{q_i\}$ such that $\beta^j_j = \alpha_j$. Then the vector field X^j as constructed above has β as an attracting limit cycle. Denote by Φ^j the Lyapunov function which measures proximity to β: $\Phi^j(p) := \left\| p - \beta^j \right\|$ (with distance measured in say the product metric on $\mathcal{C}$). Then consider the modified Lyapunov function $\Psi^j(p) := \Phi^j(p) + \|p - q_{j+1}\|$, which measures the distance to the endpoint of the arc β^j_j in addition to the proximity to β^j.

Repeat this procedure for each j, yielding the vector fields $\{X^j\}$ which attract, respectively, to limit cycles β^j. It follows that X^j prepares X^{j+1} since the goal point of X^j, q_{j+1} lies on the attracting set of X^{j+1}. The Lyapunov functions $\{\Psi^j\}$ serve as a set of funnels which channel the orbit into the sequence of arcs α_j, forming α. One scales the Ψ^j so that a $\Psi^j < \epsilon$ event triggers the switching in the hybrid controller from X^j to X^{j+1}:

$$X := \begin{cases} X^1 & : \quad \Phi^j > \epsilon \quad \forall\, j, \\ X^j & : \quad \Phi^j < \epsilon \text{ and } \Psi^j > \epsilon. \end{cases} \tag{4.9}$$

By construction, the hybrid controller (4.9) realizes a limit cycle within ϵ of α as the attracting set.

5. Future directions. A point of primary concern is the adaptability of the global topological approach to systems which increase in complexity, either through more intricate graphs or through increased numbers of AGVs. The latter is of greater difficulty than the former since the dimension of the resulting configuration space is equal to the number of AGVs. Hence, no matter how simple the underlying graph is, a system with ten independent AGVs will require a dynamical controller on a (topologically complicated) ten-dimensional space—a formidable problem both from the topological, dynamical, and computational viewpoints.

However, there are some approaches which may facilitate working with such spaces. Consider the model space $\mathcal{C}$ with which this paper is concerned: although a two-dimensional space $\mathcal{C}$ is homeomorphic to the product of a graph (a circle with six radial edges attached) with the interval $(0, 1]$. In fact, if we consider the circulating flow of (3.3)–(3.5), one can view this as a product field of a semiflow on the graph (which "circulates") with a vector field on the factor $(0, 1]$ (which "pushes out" to the boundary).

A similar approach is feasible for arbitrary graphs [9].

THEOREM 4. *Given any graph* Γ *(except the graph homeomorphic to a circle), the configuration space of* N *distinct points on* Γ *can be deformation retracted to a subcomplex whose dimension is bounded above by the number of vertices of* Γ *of valency greater than two.*[7]

This theorem implies the existence of low dimensional *spines* which carry all of the topology of the configuration space. For example, the configuration space of N points on the Y-graph can be continuously deformed to a one-dimensional graph, regardless of the size of N. Since the full space can be deformation retracted onto the spine, a vector field defined on the spine can be pulled back continuously to the full configuration space, thus opening up the possibility of reducing the control problem to that on a much "smaller" space. Additional results about the topology of configuration spaces on graphs may yield computationally tractable means of dealing with complex path planning: for example, having a presentation for the fundamental group of a configuration space of a graph in terms of a suitably simple set of cycles would be extremely well-suited to a hybrid control algorithm based on "localized" vector fields supported on small portions of the full configuration space.

Results connected with computational issues for configuration spaces of graphs are also being developed. Abrams has developed a "discretization" algorithm for converting the configuration space of a graph into a cubical complex [1]. This is then perfectly suited to the recent algorithms in computational homology [11] which prefer cube complex structures and can quickly determine geodesic paths.

The optimization problem is another avenue for inquiry. The fact that a dynamical approach allows for increased density of AGVs on a graph (as compared with blocking-zone strategies) would indicate an increased efficiency with respect to, say, elapsed time of flight. However, a more careful investigation of the tuning of optimal cycles is warranted. A careful treatment of the geometry of configuration spaces of graphs is essential to the optimization problem: it follows from the recent thesis of Abrams [1] that these spaces always possess a remarkable geometric property (*NPC* or *nonpositively curved*) which implies, among other things, that geodesics are unique

[7] Added in proof: A similar result has been shown in [16].

within their homotopy class. Such properties, though rare in the world of topological spaces, appear to be not at all uncommon among real-world robotic systems [2].

We believe that the benefits associated with using the full configuration space to tune optimal dynamical cycles justifies a careful exploration of these challenging spaces.

Appendix A. The topology and dynamics of graphs.

In this appendix, we provide a careful basis for the use of vector fields on configuration spaces of graphs. In the setting of manifolds, all of the constructions used in this paper are entirely natural and well defined. However, on spaces like $\mathcal{C}$, the most fundamental of notions (like the existence and uniqueness theorems for ODEs) are not in general valid.

We begin by defining vector fields on graphs. For present purposes, it is convenient to work with an intrinsic formulation (i.e., directly in the graph rather than via an embedding) of these objects. To this end, denote by v a vertex with K incident edges $\{e_i\}_1^K$ and by $\{X_i\}_1^K$ a collection of nonsingular vector fields locally defined on a neighborhood of the endpoint of each e_i (homeomorphic to $[0,1)$).

LEMMA 7. *A set of nonsingular vector fields $\{X_i\}$ on the local edge set of a graph Γ generates a well-defined semiflow on Γ if the following hold:*

1. *Each edge field X_i generates a well-defined local semiflow on $(0,1)$.*
2. *The magnitude of the endpoint vectors $\|X_i(0)\|$ (taken with respect to the attaching homeomorphisms) are all identical.*
3. *Among the* signs *of the endpoint vectors $X_i(0)$ (either positive if pointing into $[0,\epsilon)$ or negative if pointing out) there is a single positive sign.*

Proof. Since the vector field is well defined away from the vertex, it is only necessary to have the magnitudes $\|X_i(0)\|$ agree in order to have a well-defined function $\|X\|$ on Γ. In order to make this a well-defined field of directions, we must also consider in which direction the vector is pointing. Again, this is determined off of the vertex by (1). Condition (3) means that at the vertex there is a unique direction along which the vector field is pointing out: all other edges point in. Hence the direction field, as well as the magnitude field, is well defined.

The semiflow property follows naturally from this. Assume that the Nth edge of Γ has the positive sign. Then, given an initial point $x \in \Gamma$, if $x \in e_N$, then the orbit of x under the local field X_N remains in e_N and is well defined. If $x \in e_j$ for some $j \neq N$, then the union of the edges $e_j \cup e_N$ is a manifold homeomorphic to R on which the vector fields X_j and X_N combine to yield a well-defined vector field, since the directions are "opposite." As we are now on a manifold, the standard existence theorem implies that x has a forward orbit (which passes through the vertex and continues into e_N). Thus every point on Γ has a well-defined forward orbit. □

In the case where the vector fields have singularities, it is a simpler matter. If the singularities are not at the vertex, then there is no difference. If there is a singularity at the vertex, then condition (3) in Lemma 7 is void—all such vector fields are well defined.

In order to extend these results to the configuration space of this paper, consider the space $\mathcal{C} = \Upsilon \times \Upsilon - \Delta$, and let $(x,y) \in \mathcal{C}$ denote a point on the branch set of $\mathcal{C}$. Because of the structure of Υ and the fact that the diagonal points are deleted, it follows that at most one AGV may occupy a nonmanifold point of Υ. Hence a neighborhood of (x,y) in $\mathcal{C}$ has a natural product structure $N \cong \Upsilon \times R$. Let $P : N \to \Upsilon$ denote projection onto the first factor.

LEMMA 8. *A nonsingular vector field X on the individual cells of $\mathcal{C}$ generates*

a well-defined semiflow if (1) *the projection of the local vector fields onto the graph factor,* $P_*(X|_{\{x\}\times\Upsilon})$*, satisfies Lemma* 7 *for each point* x *in the branch set of* $\mathcal{C}$ *and* (2) *the projections of the vector fields on the branch set to the* R*-factor are equal up to the attaching maps.*

Proof. Off of the branch set, the space is a manifold, and hence the vector field gives a well defined flow. If p is a point on the branch line, condition (2) implies that the vector field is well defined with respect to the attaching maps and the net effect in the R-factor is a drift in this direction. In the graph factor, condition (1) and the proof of Lemma 7 imply that there is a unique forward orbit through p. □

Intuitively, this condition means that, as in the case of a graph, the vector field must point "in" on all but one sheet of the configuration space in order to have well-defined orbits. We may thus lift the criteria of Lemma 7 to the product configuration space. All of the vector fields in this paper are so constructed.

REFERENCES

[1] A. Abrams, *Configuration Spaces of Graphs and Brownian Motion*, Ph.D. thesis, University of California at Berkeley, Berkeley, CA, 2000.

[2] A. Abrams and R. Ghrist, *Shape Complexes for Reconfigurable Robotic Systems*, in preparation, 2001.

[3] Y. A. Bozer and M. M. Srinivasan, *Tandem configurations for automated guided vehicle systems and the analysis of single vehicle loops*, IIE Transactions, 23 (1991), pp. 72–82.

[4] Y. Brave and M. Heymann, *On optimal attraction of discrete-event processes*, Inform. Sci., 67 (1993), pp. 245–276.

[5] R. R. Burridge, A. A. Rizzi, and D. E. Koditschek, *Sequential composition of dynamically dexterous robot behaviors*, Int. J. Rob. Res., 18 (1999), pp. 534–555.

[6] G. A. Castleberry, *The AGV Handbook*, Braun-Brumfield, Ann Arbor, MI, 1991.

[7] M. Erdmann, *Understanding action and sensing by designing actions-based sensors*, Int. J. Rob. Res., 14 (1995), pp. 483–509.

[8] S. B. Gershwin, *Manufacturing Systems Engineering*, Prentice-Hall, Englewood Cliffs, NJ, 1994.

[9] R. Ghrist, *Configuration spaces and braid groups on graphs in robotics*, in Braids, Links, and Mapping Class Groups: The Proceedings of Joan Birman's 70th Birthday, AMS/IP Stud. Adv. Math. 24, AMS, Providence, RI, 2001, pp. 29–39. ArXiv preprint math.GT/9905023.

[10] R. Ghrist and D. E. Koditschek, *Safe cooperative robot dynamics on graphs*, in Proceedings of the 8th International Symposium on Robotic Research, Y. Nakayama, ed., Springer-Verlag, New York, 1998, pp. 81–92.

[11] W. Kalies, K. Mischaikow, and G. Watson, *Cubical approximation and computation of homology*, in Conley Index Theory, Banach Center Publ. 47, Polish Acad. Sci., Warsaw, Poland, 1999, pp. 115–131.

[12] D. E. Koditschek and E. Rimon, *Robot navigation functions on manifolds with boundary*, Adv. Appl. Math., 11 (1990), pp. 412–442.

[13] D. Lind and B. Marcus, *An Introduction to Symbolic Dynamics and Coding*, Cambridge University Press, Cambridge, UK, 1995.

[14] T. Lozano-Perez, M. T. Mason, and R. H. Taylor, *Automatic synthesis of fine-motion strategies for robots*, Int. J. Rob. Res., 3 (1984), pp. 3–23.

[15] M. T. Mason, *The mechanics of manipulation*, in Proceedings of the IEEE International Conference on Robotics and Automation, IEEE Robotics and Automation Society, Piscataway, NJ, 1985, pp. 544–548.

[16] R. J. Milgram and S. Kaufman, *Topological Characterization of Safe Coordinated Vehicle Motion*, in preparation.

[17] J. R. Munkres, *Topology: A First Course*, Prentice-Hall, Englewood Cliffs, NJ, 1975.

[18] R. Sengupta and S. Lafortune, *An optimal control theory for discrete event systems*, SIAM J. Control Optim., 36 (1998), pp. 488–541.

[19] S. Smith, *Reactive scheduling systems*, in Intelligent Scheduling Systems, D. Brown and W. Schering, eds., Kluwer Academic Publishers, Boston, 1995, pp. 155–192.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 5, pp. 1576–1584

FILTERING OF NONLINEAR STOCHASTIC FEEDBACK SYSTEMS*

F. CARRAVETTA†, A. GERMANI‡, R. LIPTSER§, AND C. MANES‡

Abstract. This paper concerns the filtering problem for a class of stochastic nonlinear systems where the drift term may depend either on some external function (*open-loop system*) or on the system output (*closed-loop system*), through a *controller*. Such systems are denoted *feedback systems*. The following result is proven: for feedback systems, the optimal filter in the open-loop case remains optimal when the feedback is closed. The proof is obtained by showing equivalence of suitable expressions for the estimators of the open-loop and closed-loop systems, obtained using the Kallianpur–Striebel formula [G. Kallianpur and C. Striebel, *Ann. Math. Statist.*, 39 (1968), pp. 785–801].

Key words. nonlinear filtering, closed-loop systems, Kallianpur–Striebel formula, Girsanov theorem

AMS subject classifications. 93E03, 93E11

PII. S0363012998347146

1. Introduction.

Consider the class of nonlinear stochastic systems described by the equations

$$\begin{aligned} dX_t^\phi &= f(t, X_t^\phi, u(t, \phi_{[0,t]}))dt + b(t, X_t^\phi)dW_t', \\ dY_t^\phi &= h(t, X_t^\phi)dt + B(t)dW_t'', \end{aligned} \tag{1.1}$$

where $X_t^\phi \in \mathbb{R}^n$ is the system state, $Y_t^\phi \in \mathbb{R}^m$ is the observation process, and $u(t, \phi_{[0,t]}) \in \mathbb{R}^p$ is the input function, generated by some driving function ϕ. f, h are vector functions of suitable dimensions. $W_t' \in \mathbb{R}^n$ and $W_t'' \in \mathbb{R}^m$ are independent Wiener processes. (Without loss of generality, we consider square diffusion matrices b and B.)

If in system (1.1) the driving function ϕ is *replaced* by the system output Y, we obtain the following system:

$$\begin{aligned} dX_t &= f(t, X_t, u(t, Y_{[0,t]}))dt + b(t, X_t)dW_t', \\ dY_t &= h(t, X_t)dt + B(t)dW_t''. \end{aligned} \tag{1.2}$$

So the term $u(t, Y_{[0,t]})$ represents a causal map of the observation process into the input, describing a behavior of some feedback control device (*the controller*). We will refer to system (1.1) as the *open-loop* system, and to system (1.2) as the *closed-loop* system.

*Received by the editors November 12, 1998; accepted for publication (in revised form) September 17, 2001; published electronically February 6, 2002. This work was partially supported by ASI (Italian Aerospace Agency).

http://www.siam.org/journals/sicon/40-5/34714.html

†Istituto di Analisi dei Sistemi e Informatica del CNR, Viale Manzoni 30, 00185 Roma, Italy (carravetta@iasi.rm.cnr.it).

‡Istituto di Analisi dei Sistemi e Informatica del CNR, Viale Manzoni 30, 00185 Roma, Italy and Dipartimento di Ingegneria Elettrica, Università degli Studi dell'Aquila, 67040 Monteluco di Roio, L'Aquila, Italy (germani@ing.univaq.it, manes@ing.univaq.it).

§Department of Electrical Engineering–Systems, Tel Aviv University, 69978 Ramat Aviv, Israel (liptser@eng.tau.ac.il).

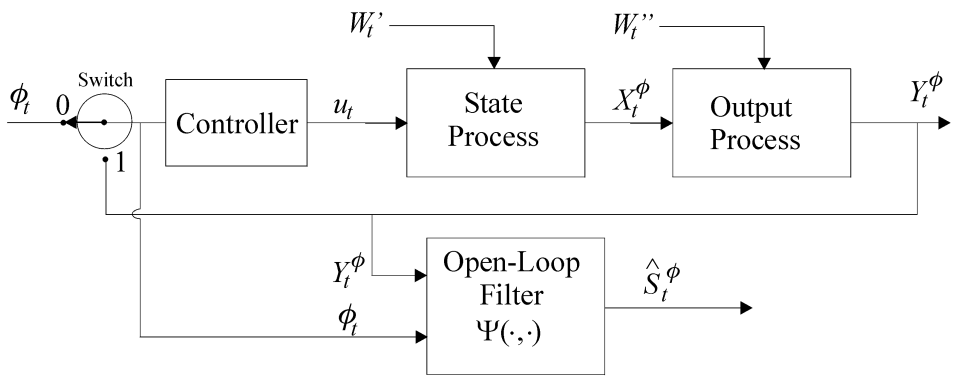


FIG. 1.1. *The optimal filter for the open-loop system.*

Let $F : \mathbb{R}^n \mapsto \mathbb{R}^{n'}$ be a function of the system state that defines a signal to be estimated for the open- and closed-loop systems:

$$S_t^\phi = F\big(X_t^\phi\big), \tag{1.3}$$

$$S_t = F\big(X_t\big). \tag{1.4}$$

Assume for every fixed t there is a function $\Psi_t(y_{[0,t]}; \phi_{[0,t]})$, $(y_t, \phi(t),\ t \geq 0$, are continuous vector functions valued in $\mathbb{R}^p)$ such that

$$\Psi_t(Y_{[0,t]}^\phi; \phi_{[0,t]}) = E\big(S_t^\phi / Y_{[0,t]}^\phi\big). \tag{1.5}$$

This is the *open-loop filter*, i.e., the optimal filter for the open-loop system (1.1), forced by the system output and by the forcing term ϕ. For every t, also assume that there exists a function $\Phi_t(y_{[0,t]})$ such that

$$\Phi_t(Y_{[0,t]}) = E\big(S_t / Y_{[0,t]}\big),\ P\text{-a.s.} \tag{1.6}$$

This is the *closed-loop filter*, i.e., the optimal filter for the closed-loop system (1.2) that is forced by the system output only.

The following question arises:

$$\Psi_t(Y_{[0,t]}; Y_{[0,t]}) \stackrel{?}{=} \Phi_t(Y_{[0,t]}),\ P\text{-a.s.} \tag{1.7}$$

Stated in other words, if we apply the *open-loop filter to the closed-loop system*, then does the estimate agree with the optimal state-estimate for the closed-loop system?

This problem is also depicted in Figures 1.1 and 1.2, where a switch can commute from position 0 (open-loop system) to position 1 (closed-loop system). For each position of the switch there is a different optimal filter. An affirmative answer to question (1.7) means that the filter of Figure 1.1 remains optimal when the switch commutes from position 0 to position 1.

The question if (1.7) holds or not is not only interesting by itself but is important in many applications. For instance, in all cases in which a finite-dimensional filter exists for the open-loop system (see [4]), identity (1.7) proves that the filter remains optimal and finite-dimensional also when the feedback is closed. Another

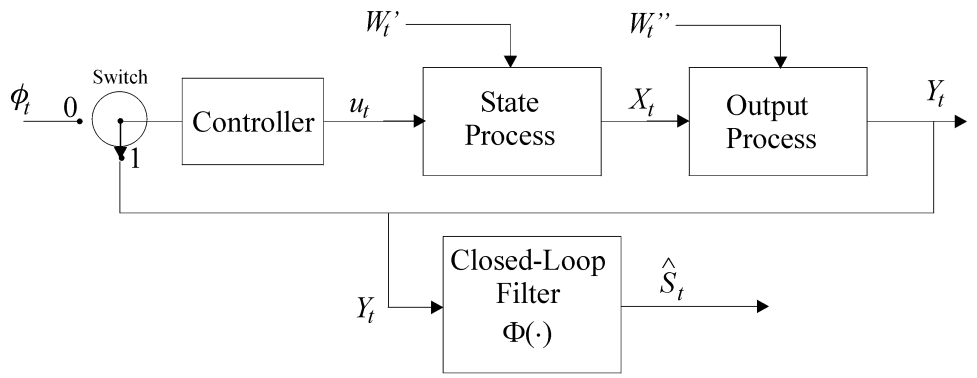


FIG. 1.2. *The optimal filter for the closed-loop system.*

interesting application is when $\Phi_t(Y_{[0,t]})$ is computed by the Monte Carlo method via $\Psi_t(Y^{\phi}_{[0,t]}; \phi_{[0,t]})$.

Up to now, the correctness of (1.7) has been proved only for particular cases of the problem, such as in the case of the linear-Gaussian system under nonlinear feedback [6], [10], [11] of the type

$$\begin{aligned} dX_t &= (A(t, Y_{[0,t]})X_t + u(t, Y_{[0,t]}))dt + F(t, Y_{[0,t]})dW_t', \\ dY_t &= C(t, Y_{[0,t]})X_t dt + G(t, Y_{[0,t]})dW_t'', \end{aligned}$$

which is important from an application point of view.

In this paper, we give an affirmative answer to question (1.7) for the nonlinear models (1.1), (1.2), under some not very restrictive assumptions.

The paper is organized as follows: section 2 reports the rigorous statement of the problem, and section 3 presents the main theorem. Conclusions follow.

2. Problem statement. On a probability space $\{\Omega, \mathcal{F}, P\}$, consider two independent Wiener processes W_t' and W_t'', $t \in [0, \infty)$, of dimension n and m, respectively, and a random vector $\mathcal{X} \in \mathbb{R}^n$. Let $\mathcal{F}^t$ be the nondecreasing family of σ-algebras generated by $\{(\mathcal{X}, W_s', W_s''),\ 0 \le s \le t\}$. Throughout the paper, $\mathcal{C}_{[0,\infty)}(\mathbb{R}^q)$ shall denote the space of $\mathbb{R}^q$-valued continuous functions over the interval $[0, \infty)$. On this space, let $\mathcal{B}_t^q$, $t \ge 0$, be the σ-algebra generated by cylinder sets of the form

$$\big\{\varphi \in \mathcal{C}_{[0,\infty)}(\mathbb{R}^q) :\ \phi(t_k) \in B_k;\ t_k \le t;\ k{=}1,\ldots,\bar{k};\ \bar{k} \in \mathbb{N};\ B_k \in \mathcal{B}(\mathbb{R}^q)\big\}, \tag{2.1}$$

where $\mathcal{B}(\mathbb{R}^q)$ is the Borel σ-algebra of $\mathbb{R}^q$. Moreover, let $\mathcal{B}_\infty^q = \vee_{t\ge 0}\mathcal{B}_t^q$. Let $\mathcal{R}_+$ be the Borel σ-algebra on $\mathbb{R}_+$.

Given a process ξ_t, let $\sigma_t(\xi)$ be the σ-algebra generated by $\{\xi_s,\ 0 \le s \le t\}$.

For a given $\phi \in \mathcal{C}_{[0,\infty)}(\mathbb{R}^m)$, consider the *open-loop* model:

$$\begin{aligned} dX_t^\phi &= f(t, X_t^\phi, u(t,\phi))dt + b(t, X_t^\phi)dW_t', \quad & X_0^\phi &= \mathcal{X}, \\ dY_t^\phi &= h(t, X_t^\phi)dt + B(t)dW_t'', & Y_0^\phi &= 0, \\ S_t^\phi &= F\big(X_t^\phi\big). \end{aligned} \tag{2.2}$$

Consider also the *closed-loop* model:

$$
\begin{aligned}
dX_t &= f(t, X_t, u(t,Y))dt + b(t, X_t)dW'_t, \quad X_0 = \mathcal{X},\\
dY_t &= h(t, X_t)dt + B(t)dW''_t, \qquad\qquad\quad Y_0 = 0,\\
S_t &= F\big(X_t\big).
\end{aligned} \tag{2.3}
$$

In both models the state space is $\mathbb{R}^n$, the observation space is $\mathbb{R}^m$, and the signal space is $\mathbb{R}^{n'}$.

For models (2.2) and (2.3), we make the following assumptions:

(i) The function $u:\ \mathbb{R}_+ \times \mathcal{C}_{[0,\infty]}(\mathbb{R}^m) \mapsto \mathbb{R}^p$ is $\mathbb{R}_+ \otimes \mathcal{B}^m_\infty$-measurable and $\{\mathcal{B}^m_t\}_{t\geq 0}$-adapted.

(ii) For any $t \in \mathbb{R}_+$, the functions $f(t,\cdot,\cdot)$, $h(t,\cdot)$, $F(\cdot)$ have bounded components.

(iii) There exist an increasing function $L(t)$ and a measure $\mu(dt)$ on $\mathbb{R}_+$, with $\int_0^t \mu(ds) < \infty$, $t > 0$, so that (here $\|\cdot\|$ is the Euclidean norm)

$$
\begin{aligned}
&\big\|f(t,x',u(t,y')) - f(t,x'',u(t,y''))\big\|\\
&\qquad\qquad \leq L(t)\left(\|x'-x''\| + \int_0^t \|y'_s - y''_s\|\mu(ds)\right),\\
&\big\|h(t,x') - h(t,x''))\big\| \leq L(t)(\|x'-x''\|),\\
&\big\|b(t,x') - b(t,x''))\big\| \leq L(t)(\|x'-x''\|).
\end{aligned} \tag{2.4}
$$

(iv) Matrices $\mathfrak{D}_t := BB^*(t)$ and $\mathfrak{d}_t := bb^*(t,x)$ (* is the transposition symbol) are uniformly nonsingular, respectively, in $\mathbb{R}_+$ and in $\mathbb{R}_+ \times \mathbb{R}^n$, with bounded inverse.

(v) (*Open-loop filter.*) There exists a function $\Psi : \mathbb{R}_+ \times \mathcal{C}_{[0,\infty]}(\mathbb{R}^m) \times \mathcal{C}_{[0,\infty]}(\mathbb{R}^m) \mapsto \mathbb{R}^{n'}$, $\mathcal{R}_+ \otimes \mathcal{B}^m_\infty \otimes \mathcal{B}^m_\infty$-measurable and $\{\mathcal{B}^m_t \otimes \mathcal{B}^m_t\}_{t\geq 0}$-adapted, such that

$$
\Psi_t(Y^\phi;\phi) = E\big(S^\phi_t/\sigma_t(Y^\phi)\big),\ P\text{-a.s.},\ \forall t \in \mathbb{R}_+. \tag{2.5}
$$

(vi) (*Closed-loop filter.*) There exists a function $\Phi : \mathbb{R}_+ \times \mathcal{C}_{[0,\infty]}(\mathbb{R}^m) \mapsto \mathbb{R}^{n'}$, $\mathcal{R}_+ \otimes \mathcal{B}^m_\infty$-measurable and $\{\mathcal{B}^m_t\}_{t\geq 0}$-adapted, such that

$$
\Phi_t(Y) = E\big(S_t/\sigma_t(Y)\big),\ P\text{-a.s.},\ \forall t \in \mathbb{R}_+. \tag{2.6}
$$

Note that thanks to the assumption of $\{\mathcal{B}_t\}_{t\geq 0}$-measurability of the function u, the term $u(t,Y)$ performs a causal mapping of the observation process into the input. Moreover, note that condition (iii) guarantees existence and uniqueness of strong solutions of (2.2) and of (2.3), adapted to $\mathcal{F}^t$.

3. Main result. The main result of this paper is given by the following theorem, which answers question (1.7).

THEOREM 3.1. *Consider the open-loop and the closed-loop nonlinear stochastic models* (2.2) *and* (2.3). *Let the assumptions* (i)–(vi) *be satisfied. Then the functions* Ψ_t *and* Φ_t *defined in* (2.5) *and* (2.6) *are such that*

$$
\Psi_t(Y;Y) = \Phi_t(Y),\ P\text{-a.s.},\ \forall\ t \in \mathbb{R}_+. \tag{3.1}
$$

Before proving this theorem, we have to state some preliminary results. Throughout the paper we will use the following notation:

$$
\big\|h(t,x)\big\|^2_{\mathfrak{D}_t^{-1}} = h^*(t,x)\left(BB^*\right)^{-1}(t)h(t,x). \tag{3.2}
$$

Moreover, for a given process ξ taking values on $\mathcal{C}_{[0,\infty)}(\mathbb{R}^q)$, we shall denote by μ_ξ^t the measure induced by the process on $\{\mathcal{C}_{[0,\infty)}(\mathbb{R}^q), \mathcal{B}_t^q\}$.

Let $F_t : \mathcal{C}_{[0,\infty]}(\mathbb{R}^n) \mapsto \mathbb{R}^{n'}$ be the bounded function defined by the equality $F_t(z) = F(z(t))$, where F is the function defining the signals for systems (2.2), (2.3).

LEMMA 3.2 (Kallianpur–Striebel formula for $\Psi_t(Y^\phi;\phi)$). *For any $t \geq 0$, the open-loop filter can be written as*

$$\Psi_t(Y^\phi;\phi) = \frac{\int_{\mathcal{C}_{[0,\infty)}(\mathbb{R}^n)} F_t(z)\Lambda_t(z,Y^\phi)\mu^t_{X^\phi}(dz)}{\int_{\mathcal{C}_{[0,\infty)}(\mathbb{R}^n)} \Lambda_t(z,Y^\phi)\mu^t_{X^\phi}(dz)}, \tag{3.3}$$

where

$$\Lambda_t(X^\phi,Y^\phi) = \exp\left(\int_0^t h^*(s,X_s^\phi)\mathfrak{D}_s^{-1}dY_s^\phi - \frac{1}{2}\int_0^t \left\|h(s,X_s^\phi)\right\|^2_{\mathfrak{D}_s^{-1}}ds\right). \tag{3.4}$$

Proof. Consider the process

$$d\zeta_t = B(t)dW_t'', \quad \zeta_0 = 0. \tag{3.5}$$

By Theorem 7.20 of [5] and comments from subsection 7.6.4 after this theorem in [5], for any $t \geq 0$ the distributions of processes $(X_s^\phi, Y_s^\phi)_{s\leq t}$, $(X_s^\phi, \zeta_s)_{s\leq t}$ are equivalent. Moreover, we have

$$\begin{gathered}\Lambda_t(z,y) = \frac{d\mu^t_{X^\phi,Y^\phi}}{d\mu^t_{X^\phi,\zeta}}(z,y),\\ (z,y) \in \mathcal{C}_{[0,\infty)}(\mathbb{R}^n) \times \mathcal{C}_{[0,\infty)}(\mathbb{R}^m).\end{gathered} \tag{3.6}$$

From this, the following equation is obtained:

$$\int_{\mathcal{C}_{[0,\infty)}(\mathbb{R}^n)} \Lambda_t(z,y)\mu^t_{X^\phi}(dz) = \frac{d\mu^t_{Y^\phi}}{d\mu^t_\zeta}(y), \qquad y \in \mathcal{C}_{[0,\infty)}(\mathbb{R}^m). \tag{3.7}$$

From Theorem 7.23 in [5] and its multidimensional analogue Lemma 2.3 in [12], it is

$$\Psi_t(Y^\phi;\phi) = \int_{\mathcal{C}_{[0,\infty)}(\mathbb{R}^n)} F_t(z)\rho_t(z,Y^\phi)\mu^t_{X^\phi}(dz), \tag{3.8}$$

with

$$\rho_t(z,y) = \frac{d\mu^t_{X^\phi,Y^\phi}}{d\mu^t_{X^\phi,\zeta}}(z,y) \Big/ \frac{d\mu^t_{Y^\phi}}{d\mu^t_\zeta}(y). \tag{3.9}$$

From the expressions of the Radon–Nikodym derivatives, it follows that

$$\begin{gathered}\rho_t(z,y) = \frac{\Lambda_t(z,y)}{\int_{\mathcal{C}_{[0,\infty)}(\mathbb{R}^n)} \Lambda_t(z,y)\mu^t_{X^\phi}(dz)},\\ (z,y) \in \mathcal{C}_{[0,\infty)}(\mathbb{R}^n) \times \mathcal{C}_{[0,\infty)}(\mathbb{R}^m),\end{gathered} \tag{3.10}$$

and from this (3.3) follows. ∎

From assumptions (i)–(iii), there exists a $Q : \mathbb{R}_+ \times \mathbb{R}^n \times \mathcal{C}_{[0,\infty)}(\mathbb{R}^n) \times \mathcal{C}_{[0,\infty)}(\mathbb{R}^m) \mapsto \mathbb{R}^n$, measurable and $\{\mathcal{B}(\mathbb{R}^n) \otimes \mathcal{B}_t^n \otimes \mathcal{B}_t^m\}$-adapted, such that the closed-loop state process can be written as

$$X_t = Q_t(\mathcal{X}, W', Y). \tag{3.11}$$

$Q(\mathcal{X}, W', Y)$ will denote the process $\{Q_s(\mathcal{X}, W', Y), s \in \mathbb{R}_+\}$.

LEMMA 3.3 (Kallianpur–Striebel formula for $\Phi_t(Y)$). *For any $t \geq 0$, the closed-loop filter can be written as*

$$\Phi_t(Y) = \frac{\int_{\mathbb{R}^n \times \mathcal{C}_{[0,t]}(\mathbb{R}^n)} F\big(Q(x,w,Y)\big)\mathfrak{A}_t(x,w,Y)\mu_{\mathcal{X}}(dx)\mu_{W'}^t(dw)}{\int_{\mathbb{R}^n \times \mathcal{C}_{[0,t]}(\mathbb{R}^n)} \mathfrak{A}_t(x,w,Y)\mu_{\mathcal{X}}(dx)\mu_{W'}^t(dw)}, \tag{3.12}$$

where

$$\begin{aligned}\mathfrak{A}_t(x,w,Y) = \exp\Bigg\{ &\int_0^t h^*\ \big(\ s, Q_s(x,w,Y)\big)\mathfrak{D}_s^{-1} dY_s \\ &- \frac{1}{2}\int_0^t \big\|h\big(s, Q_s(x,w,Y)\big)\big\|^2_{\mathfrak{D}_s^{-1}} ds\Bigg\}.\end{aligned} \tag{3.13}$$

Proof. As in the proof of Lemma 3.2, apply Theorem 7.20 of [5] to the processes $\big(\mathcal{X}, W', Y\big)$ and $\big(\mathcal{X}, W', \zeta\big)$: for all $t \geq 0$ the distributions $\mu^t_{\mathcal{X},W',Y}$ and $\mu^t_{\mathcal{X},W',\zeta}$ are equivalent, and the Radon–Nikodym derivative is

$$\begin{aligned}&\frac{d\mu^t_{\mathcal{X},W',Y}}{d\mu^t_{\mathcal{X},W',\zeta}}(x,w,y) = \mathfrak{A}_t(x,w,y), \\ &(x,w,y) \in \mathbb{R}^n \times \mathcal{C}_{[0,\infty)}(\mathbb{R}^n) \times \mathcal{C}_{[0,\infty)}(\mathbb{R}^m),\end{aligned} \tag{3.14}$$

where $\mathfrak{A}_t$ is defined in (3.13). The following equation can be verified:

$$\int_{\mathcal{C}_{[0,\infty)}(\mathbb{R}^n)} \mathfrak{A}_t(x,w,y)\mu_{\mathcal{X}}(dx)\mu_{W'}^t(dw) = \frac{d\mu_Y^t}{d\mu_\zeta^t}(y), \qquad y \in \mathcal{C}_{[0,\infty)}(\mathbb{R}^m). \tag{3.15}$$

Again, using Theorem 7.23 in [5] and its multidimensional analogue Lemma 2.3 in [12], we have

$$\Phi_t(Y) = \int_{\mathbb{R}^n \times \mathcal{C}_{[0,\infty)}(\mathbb{R}^n)} F\big(Q_s(x,w,Y)\big)\gamma_s(x,w,Y)\mu_{\mathcal{X}}(dx)\mu_{W'}^t(dw), \tag{3.16}$$

with

$$\gamma_t(x,w,y) = \frac{d\mu^t_{\mathcal{X},W',Y}}{d\mu^t_{\mathcal{X},,W',\zeta}}(x,w,y) \Big/ \frac{d\mu_Y^t}{d\mu_\zeta^t}(y). \tag{3.17}$$

From these one has

$$\begin{aligned}&\gamma_t(x,w,y) = \frac{\mathfrak{A}_t(x,w,y)}{\int_{\mathbb{R}^n \times \mathcal{C}_{[0,\infty)}(\mathbb{R}^n)} \mathfrak{A}_t(x,w,y)\mu_{\mathcal{X}}(dx)\mu_{W'}^t(dw)}, \\ &(x,w,y) \in \mathbb{R}^n \times \mathcal{C}_{[0,\infty)}(\mathbb{R}^n) \times \mathcal{C}_{[0,\infty)}(\mathbb{R}^m).\end{aligned} \tag{3.18}$$

Equation (3.12) follows. □

Let us define the process Γ as follows:

$$\Gamma_t(X,Y) = \exp\left\{ \int_0^t h^*(s,X_s)\mathfrak{D}_s^{-1} dY_s - \frac{1}{2}\int_0^t \big\|h(s,X_s)\big\|^2_{\mathfrak{D}_s^{-1}} ds \right\}. \tag{3.19}$$

Note that from (3.11) and (3.13) we have

$$\Gamma_t\big(Q(\mathcal{X},W',Y),Y\big) = \mathfrak{A}_t(\mathcal{X},W',Y). \tag{3.20}$$

In the following, we have to rewrite the expressions of the open- and closed-loop filters, given by (3.3) and (3.12), respectively, in a more convenient form related to the underlying probability space. For this purpose, we introduce a copy $(\widetilde{\Omega},\widetilde{\mathcal{F}},\widetilde{P})$ of the original probability space so that all of the processes defined on it are independent copies of the original ones. We also introduce random variables and processes on the product probability space $\big(\Omega\times\widetilde{\Omega},\mathcal{F}\otimes\widetilde{\mathcal{F}},P\times\widetilde{P}\big)$.

Let $Z(\omega,\tilde{\omega})$ be a random variable defined on the product space. Let us define the operator $\widetilde{E}$ as follows:

$$\widetilde{E}(Z)(\omega) = \int_{\widetilde{\Omega}} Z(\omega,\tilde{\omega})P(d\tilde{\omega}). \tag{3.21}$$

For a given process ξ defined on the original space, we shall denote by $\tilde{\xi}$ a process defined on the product space as $\tilde{\xi}(\omega,\tilde{\omega}) = \xi(\tilde{\omega})$. Whenever it does not cause confusion, we shall use the same symbol ξ to denote both the original process and its extension to the product space: $\xi(\omega,\tilde{\omega}) = \xi(\omega)$.

On the product space it is possible to define the process $\widetilde{X}^Y$ as follows:

$$\widetilde{X}^Y_t = Q_t(\widetilde{\mathcal{X}},\widetilde{W}',Y). \tag{3.22}$$

With these positions, recalling also the definition of Γ given in (3.19), we can rewrite expressions (3.3) and (3.12) as follows:

$$\Psi_t(Y^\phi;\phi) = \frac{\widetilde{E}\big\{F_t(\widetilde{X}^\phi)\Lambda_t(\widetilde{X}^\phi,Y^\phi)\big\}}{\widetilde{E}\big\{\Lambda_t(\widetilde{X}^\phi,Y^\phi)\big\}}, \quad P\text{-a.s.}, \tag{3.23}$$

$$\Phi_t(Y) = \frac{\widetilde{E}\big\{F_t(\widetilde{X}^Y)\Gamma_t(\widetilde{X}^Y,Y)\big\}}{\widetilde{E}\big\{\Gamma_t(\widetilde{X}^Y,Y)\big\}}, \quad P\text{-a.s.} \tag{3.24}$$

Now we are in a position to give the proof of Theorem 3.1.

Proof of Theorem 3.1. From expressions (3.23) and (3.24), Theorem 3.1 is proved as soon as it is shown that

$$\Lambda_t(\widetilde{X}^\phi,Y^\phi)\big|_{\phi=Y} = \Gamma_t(\widetilde{X}^Y,Y), \qquad P\times\widetilde{P}\text{-a.s.} \tag{3.25}$$

From definitions (3.4) and (3.19) we have

$$\Lambda_t(\widetilde{X}^\phi,Y^\phi) = \exp\left(\int_0^t h^*(s,\widetilde{X}^\phi_s)\mathfrak{D}_s^{-1} dY^\phi_s - \frac{1}{2}\int_0^t \big\|h(s,\widetilde{X}^\phi_s)\big\|^2_{\mathfrak{D}_s^{-1}} ds \right), \tag{3.26}$$

$$\Gamma_t(\widetilde{X}^Y, Y) = \exp\left(\int_0^t h^*(s, \widetilde{X}_s^Y)\mathfrak{D}_s^{-1} dY_s - \frac{1}{2}\int_0^t \left\| h\big(s, \widetilde{X}_s^Y\big)\right\|^2_{\mathfrak{D}_s^{-1}} ds \right). \tag{3.27}$$

Since the integrals in (3.26) are $\sigma_t(\widetilde{X}^\phi, Y^\phi)$-adapted processes, there exist functions H and $L : \mathcal{C}_{[0,\infty)}(\mathbb{R}^n) \times \mathcal{C}_{[0,\infty)}(\mathbb{R}^m) \mapsto \mathcal{C}_{[0,\infty)}(\mathbb{R})$ that are $\mathcal{B}_t^n \otimes \mathcal{B}_t^m$-adapted and that are such that their values at time t are

$$\begin{aligned} H_t(\widetilde{X}^\phi, Y^\phi) &= \int_0^t h^*(s, \widetilde{X}_s^\phi)\mathfrak{D}_s^{-1} dY_s^\phi, \qquad P \times \widetilde{P}\text{-a.s.}, \\ L_t(\widetilde{X}^\phi, Y^\phi) &= \int_0^t \left\| h(s, \widetilde{X}_s^\phi)\right\|^2_{\mathfrak{D}_s^{-1}} ds, \qquad P \times \widetilde{P}\text{-a.s.} \end{aligned} \tag{3.28}$$

Similarly, the integrals in (3.27) can be written as

$$\begin{aligned} H'_t(\widetilde{X}^Y, Y) &= \int_0^t h^*(s, \widetilde{X}_s^Y)\mathfrak{D}_s^{-1} dY_s, \qquad P \times \widetilde{P}\text{-a.s.}, \\ L'_t(\widetilde{X}^Y, Y) &= \int_0^t \left\| h\big(s, \widetilde{X}_s^Y\big)\right\|^2_{\mathfrak{D}_s^{-1}} ds, \qquad P \times \widetilde{P}\text{-a.s.}, \end{aligned} \tag{3.29}$$

where H' and L' are functions with the same properties of H and L. We can use Lemma 4.10 of [5] to prove that[1]

$$\begin{aligned} H_t(\widetilde{X}^Y, Y) &= H'_t(\widetilde{X}^Y, Y), \\ L_t(\widetilde{X}^Y, Y) &= L'_t(\widetilde{X}^Y, Y), \end{aligned} \qquad P \times \widetilde{P}\text{-a.s.}, \tag{3.30}$$

by showing that the measures $\mu^t_{\widetilde{X}^\phi, Y^\phi}$ and $\mu^t_{\widetilde{X}^Y, Y}$ are equivalent.

As a matter of fact, Theorem 7.19 of [5] guarantees the equivalence of the measures induced by the processes $(\widetilde{X}^\phi, X^\phi, Y^\phi)$ and $(\widetilde{X}^Y, X, Y)$, which are defined on $\big(\Omega \times \widetilde{\Omega}, \mathcal{F} \otimes \widetilde{\mathcal{F}}, P \times \widetilde{P}\big)$ as follows:

$$\begin{aligned} d\widetilde{X}_t^\phi &= f(t, \widetilde{X}_t^\phi, u(t,\phi))dt + b(t, \widetilde{X}_t^\phi)d\widetilde{W}'_t, \quad \widetilde{X}_0^\phi = \widetilde{\mathcal{X}}, \\ dX_t^\phi &= f(t, X_t^\phi, u(t,\phi))dt + b(t, X_t^\phi)dW'_t, \quad X_0^\phi = \mathcal{X}, \\ dY_t^\phi &= h(t, X_t^\phi)dt + B(t)dW''_t, \qquad\qquad\quad Y_0^\phi = 0, \end{aligned} \tag{3.31}$$

$$\begin{aligned} d\widetilde{X}_t^Y &= f(t, \widetilde{X}_t^Y, u(t,Y))dt + b(t, \widetilde{X}_t^Y)d\widetilde{W}'_t, \qquad \widetilde{X}_0^Y = \widetilde{\mathcal{X}}, \\ dX_t &= f(t, X_t, u(t,Y))dt + b(t, X_t)dW'_t, \qquad X_0 = \mathcal{X}, \\ dY_t &= h(t, X_t)dt + B(t)dW''_t, \qquad\qquad Y_0 = 0. \end{aligned} \tag{3.32}$$

Since $\mu^t_{\widetilde{X}^\phi, Y^\phi}$ and $\mu^t_{\widetilde{X}^Y, Y}$ are marginal distributions of $\mu^t_{\widetilde{X}^\phi, X^\phi, Y^\phi}$ and $\mu^t_{\widetilde{X}^Y, X, Y}$, respectively, their equivalence follows as well. □

4. Conclusion. The contribution of this paper is Theorem 3.1, which represents a general property of stochastic systems that can be informally expressed in these words: whenever the optimal filter is available for a given open-loop system, the *same filter* will work optimally on the closed-loop system. One important implication of Theorem 3.1 is that any system admitting a finite-dimensional filter when in open-loop also admits a finite-dimensional filter when in closed-loop.

[1] It may seem that an alternative way of proving (3.30) and, by consequence, (3.25) could be provided by the Karandikar method [9], [3] dealing with the "pathwise definition" of the Itô integral. Our proof of (3.30) permits the traditional definition of the Itô integral and heavily uses only the equivalence of probability measures.

Acknowledgment. The authors gratefully acknowledge the careful review of an anonymous referee, who also pointed out a mistake in the original version. Due to his comments, the paper has been significantly improved.

REFERENCES

[1] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*, Prentice Hall, Englewood Cliffs, NJ, 1979.
[2] A. V. Balakrishnan, *Kalman Filtering Theory*, Optimization Software, New York, 1984.
[3] A. G. Bhatt, A. Budhiraja, and R. L. Karandikar, *Markov property and ergodicity of the nonlinear filter*, SIAM J. Control Optim., 39 (2000) pp. 928–949.
[4] M. Cohen De Lara, *Finite-dimensional filters. Part* I: *The Wei–Norman technique; Finite-dimensional filters. Part* II: *Invariance group techniques*, SIAM J. Control Optim., 35 (1997), pp. 980–1029.
[5] R. S. Liptser and A. N. Shiryayev, *Statistics of Random Processes*, Vol. 1, Springer-Verlag, New York, 1977.
[6] R. S. Liptser and A. N. Shiryayev, *Statistics of Random Processes*, Vol. 2, Springer-Verlag, New York, 1978.
[7] G. Kallianpur and C. Striebel, *Estimation of stochastic systems: Arbitrary system process with additive noise observation errors*, Ann. Math. Statist., 39 (1968), pp. 785–801.
[8] R. E. Kalman, *A new approach to linear filtering and prediction problems*, J. Basic. Engrg., 1 (1960), pp. 35–45.
[9] R. L. Karandikar, *On pathwise stochastic integration*, Stochastic Process. Appl., 57 (1995), pp. 11–18.
[10] W. J. Kolodziej and R. R. Mohler, *State estimation and control of conditionally linear systems*, SIAM J. Control Optim., 24 (1986), pp. 497–508.
[11] W. J. Kolodziej and R. R. Mohler, *On conditionally linear filtering, control, and coding*, in Stochastic Modelling and Filtering (Rome, 1984), Lecture Notes in Control and Inform. Sci. 91, A. Germani, ed., Springer-Verlag, Berlin, 1987, pp. 64–74.
[12] D. Michel, *Régularité des lois conditionelle en théorie du filtrage non-linéaire et calcul des variations stochastique*, J. Funct. Anal., 41 (1981), pp. 8–36.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 5, pp. 1585–1610

ASYMPTOTIC PROPERTIES OF RECEDING HORIZON OPTIMAL CONTROL PROBLEMS*

KAZUFUMI ITO† AND KARL KUNISCH‡

Abstract. The asymptotic behavior of receding horizon optimal control problems with terminal cost chosen as a control Liapunov function is analyzed for regulator as well as disturbance attenuation problems. Both the continuous as well as the discrete time cases are treated. Further, the approximation of the continuous time optimal control problem by the discrete time receding horizon problems is studied.

Key words. receding horizon control, model prediction control, control Liapunov function, stabilizability, Liapunov equation

AMS subject classifications. 49L05, 93D05, 93D15, 93D21

PII. S0363012900369423

1. Introduction.

This research is devoted to the analysis of certain aspects of a receding horizon formulation to optimal control problems. To motivate the approach we consider

$$\inf \int_0^{T_\infty} f^0(x(t), u(t))\, dt \tag{1.1}$$

subject to

$$\begin{cases} \frac{d}{dt}x(t) = f(x(t), u(t)) \ \text{ for } \ t > 0, \\ x(0) = x_0, \quad u(t) \in U. \end{cases} \tag{1.2}$$

We refer to $x(\cdot)$ and $u(\cdot)$ as state and control variables and assume that $x(t) \in R^n$ and $U \subset R^m$. Under appropriate conditions, (1.1)–(1.2) admit a solution which satisfies the minimum principle (e.g., [FR, IK]):

$$\begin{cases} \frac{d}{dt}x(t) = H_p(x(t), u(t), p(t)), \quad x(0) = x_0, \\ \frac{d}{dt}p(t) = -H_x(x(t), u(t), p(t)), \quad p(T_\infty) = 0, \\ u(t) = \arg\min_{u \in U} H(x(t), u, p(t)), \end{cases} \tag{1.3}$$

where $H(x, u, p) = f^0(x, u) + f(x, u) \cdot p$. The coupled system of two-point boundary value problems with initial condition for the primal equation and terminal condition for the adjoint equation represents a significant numerical challenge in the case when T_∞ is large, and it has therefore been the focus of many research efforts. If (1.1)–(1.2) arises as the discretization of an optimal control problem governed by partial differential equations, e.g., from fluid mechanics, then the numerical realization of (1.3) may become infeasible. Alternatively, if the optimal control task is for a smaller

*Received by the editors March 24, 2000; accepted for publication (in revised form) August 13, 2001; published electronically February 6, 2002.
http://www.siam.org/journals/sicon/40-5/36942.html
†Department of Mathematics, North Carolina State University, Raleigh, NC 27695 (kito@eos.ncsu.edu).
‡Institut für Mathematik, Karl-Franzens-Universität Graz, A-8010 Graz, Austria (karl.kunisch@uni-graz.at). The research of this author was partially supported in part by the Fonds zur Förderung der wissenschaftlichen Forschung under SFB 03, "Optimierung und Kontrolle."

problem but requires a fast, possibly real-time, computation, then again it may be inadequate to aim for a solution of (1.3).

Difficulties similar to those explained above occur as well for infinite horizon problems, where $T_\infty = \infty$. Moreover, when expressing the optimality condition by means of the minimum principle (1.3), we focused on open loop solutions. Closed loop solutions for (1.1)–(1.2) are given by the Hamilton–Jacobi–Bellman equation, which is even less tractable numerically.

In view of the difficulties explained above, the question of obtaining suboptimal controls arises. One of the possibilities is given by receding horizon formulations. To briefly explain the concept, let $\{T_i\}_{i\in I}$, with $T_{i-1} < T_i$ and I an index set, be sampling points, and let T stand for the control horizon, satisfying $\max_{i\in I} |T_i - T_{i-1}| \le T \ll T_\infty$. Receding horizon suboptimal solutions to (1.1)–(1.2) are obtained by solving a sequence of auxiliary optimal control problems on the subintervals $[T_i, T_i + T]$ and utilizing the informations obtained on $[T_{i-1}, T_{i-1} + T]$ to initialize the problem on the new horizon $[T_i, T_i + T]$. If $T > T_i - T_{i-1}$, then we have overlapping horizons. If $x(T_i)$ is observed, then the receding horizon control is a state feedback. In fact, the receding horizon optimal control on $[T_i, T_{i+1}]$ is determined as a function of the state $x(T_i)$. It is obtained by solving the two-point boundary value problem (1.3) on the interval $[0, T]$, with x_0 replaced by $x(T_i)$ and an appropriately modified terminal condition for p. If $T > 0$ is small, then the optimality system on the short time interval is better conditioned and easier to solve numerically than (1.3). For surveys of this general concept for continuous as well as for discrete systems, we refer, e.g., to [ABQRW, GPM, SMR].

Receding horizon formulations have proved to be effective numerically both for optimal control problems governed by ordinary (e.g., [CA, JYH, PND]) and for partial differential equations, e.g., in the form of the instantaneous control technique for problems in fluid mechanics [CHK, CTMC, HV] and heat conduction [TU]. For discrete time systems we refer to [NP], for example. The theoretical justification of receding horizon control techniques is commonly addressed by means of the stabilization problem [ABQRW, GPM, SMR]. Assuming that $x = 0$ is a steady state for (1.2) with $u = 0$ which can be stabilized by means of an optimal control formulation (1.1) with $T_\infty = \infty$, the question of whether stabilization can also be achieved by means of a receding horizon technique is addressed. To guarantee that this is the case, different variations of receding horizon formulations were utilized. An important aspect of the analysis in each of the approaches is the monotonicity of the cost functional with respect to the time horizon. In earlier versions of the formulation of receding horizon problems, the terminal condition $x(T_i + T) = 0$ was added to the auxiliary problems; see [KG, K, MM] and the references given there. Later this condition was relaxed to requiring $x(t_i + T)$ to be contained in an appropriate neighborhood of the origin; see, e.g., [ABQRW, SMR].

In this paper, we address the stabilization problem by terminal penalty terms rather than terminal constraints; i.e., we consider a sequence of auxiliary problems of the type

$$\min \int_{T_i}^{T_i+T} f^0(x(t), u(t))\, dt + G(x(T_i + T)) \tag{1.4}$$

subject to

$$\frac{d}{dt}x(t) = f(x(t), u(t)), \quad t \geq T_i$$
$$x(T_i) = \bar{x}(T_i),$$

where $\bar{x}$ is the solution to the auxiliary problem on $[T_{i-1}, T_{i-1} + T]$. The optimal pair for (1.4) has the property that $(\bar{x}(t - T_i), \bar{u}(t - T_i))$, $t \in [T_i, T_{i+1}]$, satisfies the two-point boundary value problem (1.3) on the interval $[0, T]$, with the terminal condition $p(T) = G_x(x(T))$ and the initial condition $x(0) = \bar{x}$. The functional $G\colon R^n \to R$ will be chosen as an appropriately defined control Liapunov function (see Definition 2.1). It will be shown that the addition of the terminal cost G to the cost functional provides asymptotic stability, and hence the receding horizon strategy is a suboptimal synthesis for the optimal control problem (1.1).

Control Liapunov functions received considerable attention as a means of analyzing the stability of the control system (1.1)–(1.2), regardless of issues related to optimal control. We refer to the monograph [FK] and the references given there. The use of control Liapunov functions within the context of receding horizon control is a recent one. In [PND], control Liapunov functions were utilized as explicit constraints in the auxiliary problems to guarantee that the final state $x(T_i + T)$ lies within the level curve of the control Liapunov function that is determined by the trajectory controlled by a minimum norm control. The analysis in [JYH] utilizes control Liapunov functions as a terminal penalty as in (1.4). The stabilizing properties of the resulting receding horizon optimal control strategy are analyzed under the assumption that f possesses an exponentially stabilizable critical point.

Let us now outline the contributions of this paper. Throughout, we use only a penalty term as opposed to terminal constraints to analyze and justify the receding horizon technique. We assume that the sampling points are equidistant and that the sampling rate coincides with the time horizon so that $T = T_i - T_{i-1}$ and $T_i = i\,T$, $i = 0, 1, \ldots$. If G is a control Liapunov function (see Definition 2.1 and Theorem 2.2), then we first establish the monotonicity of the value function $V_T(x_0)$:

$$V_T(x_0) = \inf \left\{ \int_0^T f^0(x(t), u(t))\, dt + G(x(T)) \text{ subject to (1.2)} \right\}$$

with respect to T; i.e., $V_T(x_0) \leq V_{\hat{T}}(x_0) \leq G(x_0)$ for $0 \leq \hat{T} \leq T$ and $x_0 \in R^n$. Thus we have (see Theorems 2.3–2.5)

$$G(x_{i+1}) + \int_{T_i}^{T_{i+1}} f^0(\bar{x}(t), \bar{u}(t))\, dt \leq G(x_i),$$

where $x_i = \bar{x}(T_i)$, $i = 1, 2, \ldots$. This implies that the values x_i are confined to the level set $S_\alpha = \{x \in R^n : G(x) \leq G(x_0) = \alpha\}$. Assume that $f(0, 0) = 0$, $G(0) = 0$, and that $f^0(x, u) > 0$ and $G(x) > 0$ except at $(0, 0)$. Then, we have $G(x_{i+1}) < V_T(x_i) \leq G(x_i)$. Assuming further that S_α is compact, we have the existence of $\rho < 1$ such that $G(x_{i+1}) \leq \rho\, G(x_i)$ for all $x \in S_\alpha$. Hence $G(x_k) \leq \rho^k\, G(x_0) \to 0$ as $k \to \infty$, which implies asymptotic stability. Moreover, if $f^0(x, u) \geq \omega G(x)$ for some $\omega > 0$, then $G(x_{i+1}) \leq e^{-\omega T} G(x_i)$ (see Theorem 2.13). We prove that the quadratic functional $G(x) = \frac{\alpha}{2}\,|x|^2$, $\alpha > 0$, can be chosen as a control Liapunov function if the control system (1.2) is closed loop dissipative (Definition 2.6), and we argue that this is

satisfied for certain classes of dissipative equations. In general, the quadratic terminal penalty is not a Liapunov function. However, we have $V_T(x) \leq \rho_T\, G(x)$ with $\rho_T < 1$ for T sufficiently large (see Theorem 2.7) provided that the value function $V(x)$ of the infinite time horizon problem satisfies $V(x) \leq \frac{\beta}{2}\,|x|^2$ and the corresponding optimal trajectory satisfies $|x^*(t)| \leq M\, e^{-\omega\, t}|x_0|,\ \omega > 0$.

In our discussion above, it is implicitly assumed that the infinite time horizon optimal control problem admits solutions, as, for example, in the case of stabilizable steady states. We refer to this situation as the regulator case, which is discussed in sections 2.1 and 2.2, and distinguish it from the general case, which is analyzed in section 2.3. The latter applies to disturbance attenuation problems and to problems with cost functionals of tracking type. We introduce control λ-Liapunov functions (see Definition 2.8) and extend the previously described analysis to the general case (Theorems 2.10–2.12). Here the positive constant λ represents the attenuation or tracking rate given by $\lim_{T\to\infty} \frac{1}{T} \int_0^T f^0(\bar{x}(t), \bar{u}(t))\, dt$.

While section 2 treats continuous time systems, section 3 is devoted to discrete time problems. Our interest is the case when the discrete time problems arise from a finite difference approximation of (1.1)–(1.2) as, for example, the explicit Euler approximation $x_j = x_{j-1} + \Delta t\, f(x_{j-1}, u_j)$ and general one-step methods. Concepts as well as results paralleling those of the continuous time case are established. In section 4 we investigate the asymptotic behavior of continuous time systems (1.2) if the controls are determined by means of a discrete time synthesis, specifically, the one-step receding horizon formulation. This analysis is of practical importance since numerical methods for solutions of (1.3) commonly use a finite difference approximation of (1.1)–(1.2), and thus the control synthesis $\bar{u}(t)$ is computed in terms of discrete time control systems.

The final section contains examples illustrating the applicability of the concepts and results of this paper.

2. Receding horizon control problems: Continuous case. We consider the control system in R^n

$$\begin{cases} \frac{d}{dt}x(t) = f(x(t), u(t)), & t > 0, \\ x(0) = x_0, \end{cases} \tag{2.1}$$

where $f\colon R^n \times U \to R^n$ is C^2, $f(0,0) = 0$, and U is a closed subset of R^m. It is assumed that for every $x_0 \in R^n$ and $T > 0$ there exists an admissible control $u \in U_{ad} = \{u \in L^1(0,T;R^m)\colon u(t) \in U \text{ a.e.}\}$ such that (2.1) admits a solution $x \in W^{1,1}(0,T;R^n)$. In our notation, we do not indicate the dependence of U_{ad} on T. If $T = \infty$, then $U_{ad} = \{u \in L^1_{loc}(0,\infty;R^n)\colon u(t) \in U \text{ a.e.}\}$. Consider the infinite time horizon optimal control problem

$$\inf_{u\in U_{ad}} \int_0^\infty f^0(x(t), u(t))\, dt \tag{2.2}$$

subject to (2.1). Here $f^0\colon R^n \times U \to R^+$ is assumed to be C^2. We suppose further that for each $(x,p) \in R^n \times R^n$ the functional

$$u \to f^0(x,u) + p \cdot f(x,u)$$

admits a unique minimizer over U denoted by $\Psi(x,p)$ and that Ψ is continuous. A sufficient condition for the continuity of Ψ is that $f(x,u)$ is linear in u and $f^0(x,u)$

is strictly convex in u. Together with (2.2), we consider the finite horizon optimal control problems on $[0,T]$, $T>0$:

$$\inf_{u\in U_{ad}} \int_0^T f^0(x(t),u(t))\,dt + G(x(T))$$

subject to (2.1). Here $G\colon R^n \to R^+$ is a continuous function.

We shall discuss and analyze the receding horizon control strategy: it consists in successively solving the finite horizon optimal control problems on $[(k-1)T, kT]$:

$$\min_{u\in U_{ad}} \int_{(k-1)T}^{kT} f^0(x(t),u(t))\,dt + G(x(kT)) \tag{2.3}$$

subject to

$$\frac{d}{dt}x(t) = f(x(t),\ u(t)),\ x((k-1)T) = x_{k-1}, \tag{2.4}$$

where $(\bar{x}_k, \bar{u}_k)$ is an optimal pair for the kth horizon $[(k-1)T, kT]$ and $x_k = \bar{x}_k(kT)$ for $k = 1, 2, \ldots$.

2.1. Regulator case.

DEFINITION 2.1. *A nonnegative continuous function G with $G(0)=0$ is a control Liapunov function for* (2.1)–(2.2) *if for all $x_0 \in R^n$ and $T>0$ there exists a control $u = u(\cdot; x_0, T) \in U_{ad}$ such that*

$$\int_0^T f^0(x(t),u(t))\,dt + G(x(T)) \le G(x_0), \tag{2.5}$$

where $x(t)$ is a solution to (2.1).

Our definition of a control Liapunov function is adapted to (1.1)–(1.2), and it therefore involves f^0 as well as f, as opposed to the one for control systems (1.2), which only involves f [S]. The following relationship between control Liapunov functions and a certain differential inequality will be useful for providing concrete examples for control Liapunov functions.

THEOREM 2.2. *Assume that G is C^1 with $G(0)=0$ and $\{x\in R^n\colon G(x)\le\alpha\}$ is bounded for every $\alpha \ge 0$.*

(a) *If there exists a locally Lipschitz continuous function $u = \Phi(x) \in U$ such that*

$$f(x,u)\cdot G_x(x) + f^0(x,u) \le 0 \tag{2.6}$$

for all $x\in R^n$, then G is a control Liapunov function for (2.1)–(2.2).

(b) *If U is compact and if for all $x_0 \in R^n$ and $\delta>0$ there exists $\tau>0$ (which may depend on x_0 and $\delta>0$) such that*

$$|x(t)-x_0| \le \delta \ \text{ for all } t\in[0,\tau] \text{ and } u\in\mathcal{U}, \tag{2.7}$$

where $u\in\mathcal{U} = \{u\in U_{ad}\colon \int_0^\tau f^0(x(t),u(t))\,dt \le G(x_0)\}$, then (2.5) *implies that for each $x\in R^n$ there exists $u\in U$ such that* (2.6) *holds.*

Proof. (a) Choose $x\in R^n$, and let $\alpha = 2G(x)$. By assumption, the set $S = \{y\colon G(y)\le\alpha\}$ is compact, and, therefore, Φ is uniformly Lipschitz continuous on S. Hence there exists $\tau = \tau(\alpha)$ such that the closed loop system

$$\frac{d}{dt}x(t) = f(x(t),\Phi(x(t))), \quad x(0)=x_0, \tag{2.8}$$

admits a unique solution on $[0,\tau]$. Thus

$$\frac{d}{dt}G(x(t)) = f(x(t),\Phi(x(t)))\cdot G_x(x(t)) \le -f^0(x(t),u(t))$$

on $[0,\tau]$. Integrating with respect to t, we have

$$G(x(\tau)) + \int_0^\tau f^0(x(t),u(t))\,dt \le G(x_0),$$

and thus $G(x(\tau)) \le G(x_0)$ and $x(\tau)\in S$. By the continuation method, there exists a unique global solution to (2.8), and (2.5) holds.

(b) If the conclusion was false, then, by compactness of U and continuity of f, f^0, and G_x, there exist $x_0 \in R^n$ and $\delta > 0$ such that, for $|x - x_0| \le \delta$,

$$f(x,u)\cdot G_x(x) + f^0(x,u) \ge \delta \tag{2.9}$$

for all $u \in U$. Let $x(0) = x_0$, choose $\tau > 0$ such that (2.7) holds, and let u denote any control satisfying (2.5). Then, in particular, $u \in \mathcal{U}$, and consequently $|x(t) - x_0| \le \delta$ by (2.7). Thus (2.9) implies

$$G(x(\tau)) + \int_0^\tau f^0(x(t),u(t))dt \ge G(x_0) + \delta\,\tau,$$

which contradicts (2.5). □

Remark 2.1. Consider the optimal control problem

$$\begin{cases} \inf \int_0^\infty (\ell(x(t)) + \frac{1}{2}|u(t)|^2)dt, \\ \dot{x}(t) = a(x(t)) + B(x(t))u(t), \quad x(0) = x_0, \end{cases} \tag{2.10}$$

with $u(t) \in R^m$, $x(t) \in R^n$, and $\ell(0) = 0$. The Hamilton–Jacobi–Bellman equation associated to (2.10) is given by

$$a(x)\cdot V_x(x) - \frac{1}{2}|b^T(x)V_x(x)|^2 + \ell(x) = 0, \tag{2.11}$$

with $V(0) = 0$. Assuming that (2.11) admits a solution in C^1, the feedback solution to (2.10) is given by $u(x) = -B^T(x)V_x(x)$. In particular, (2.6) holds as an equality if G is chosen as the minimal value function V associated to (2.10) and $\Phi(x) = -B^T(x)V_x(x)$. In [Lu], the existence of a C^1-solution to (2.11) is proved in a neighborhood N of an equilibrium point $(x_0, 0)$ of (2.10), assuming that the linearized control system of (2.10) at $(x_0, 0)$ is stabilizable and observable. In this case, N is an invariant neighborhood of the closed loop system. In general, a global C^1-solution may not exit. However, consider the special case of system (2.10) with $a(x) = Ax - B\,B^tU_x(x)$, $B(x) = B$, where U is a convex C^1-function satisfying $U_x(0) = 0$. Setting $W = V + U$, we can write the Hamilton–Jacobi–Bellman equation (2.11) as

$$Ax\cdot W_x(x) - \frac{1}{2}\,|b^TW_x(x)|^2 + \ell(x) - Ax\cdot U_x + \frac{1}{2}\,|b^TU_x|^2 = 0.$$

Thus if (A,b) is controllable and the function $x \to \ell(x) - Ax\cdot U_x + \frac{1}{2}\,|b^TU_x|^2$ with $\ell(0) = 0$ is convex, then it can be proved [FS] that W is convex and C^1, and thus $V = W - U$ is C^1. □

For the following discussion, we introduce the value function $V_T(x_0)$ of the finite horizon optimal control problem

$$V_T(x_0) = \inf_{u \in U_{ad}} \int_0^T (f^0(x(t), u(t))\, dt + G(x(T))$$

subject to (2.1) and the value function $V(x_0)$ of the infinite horizon optimal control problem

$$V(x_0) = \inf_{u \in U_{ad}} \int_0^\infty f^0(x(t), u(t))\, dt$$

subject to (2.1).

THEOREM 2.3. *Assume that G is a control Liapunov function. Then $V(x_0) \le V_T(x_0) \le G(x_0)$ for every $T \ge 0$ and $x_0 \in R^n$.*

Proof. The assertion $V_T(x_0) \le G(x_0)$ follows directly from Definition 2.1. By repeated application of (2.5) for some $T > 0$ we construct a control $u \in L^1_{loc}(0, \infty; R^n)$ with $u(t) \in U$ for a.e. $t \in [0, \infty)$ and an associated trajectory $x \in W^{1,1}_{loc}(0, \infty; R^n)$ such that for $k \ge 1$

$$G(x(kT)) + \int_{(k-1)T}^{kT} f^0(x(t), u(t))\, dt \le G((x(k-1)T)).$$

Summation of these inequalities implies for every $k \ge 1$

$$G(x(kT)) + \int_0^{kT} f^0(x(t), u(t))\, dt \le G(x_0).$$

Thus, by the Lebesgue–Fatou lemma,

$$\lim_{k \to \infty} \int_0^{kT} f^0(x(t), u(t))\, dt = \int_0^\infty f^0(x(t), u(t))\, dt \le G(x_0),$$

and hence $V(x_0)$ is finite and $V(x_0) \le G(x_0)$. Next we note that

$$\begin{aligned} &\int_0^T f^0(x(t), u(t))\, dt + G(x(T)) \\ &= \int_0^T f^0(x(t), u(t))\, dt + V(x(T)) + G(x(T)) - V(x(T)) \end{aligned} \tag{2.12}$$

and recall the optimality principle

$$V(x_0) = \inf_{u \in U_{ad}} \left\{ \int_0^T f^0(x(t), u(t))\, dt + V(x(T)) \right\}. \tag{2.13}$$

Since $G(x(T)) \ge V(x(T))$, we have $V(x_0) \le V_T(x_0)$ by (2.12)–(2.13). ∎

THEOREM 2.4 (monotonicity). *Assume that G is a control Liapunov function for (2.1)–(2.2). If $V_T(x_0)$ is the value function of the finite horizon optimal control problem, then $V_{\hat{T}}(x_0) \ge V_T(x_0)$ for $0 \le \hat{T} \le T$.*

Proof. For every $\epsilon > 0$ there exists a pair (x, u) such that (2.1) is satisfied, $u \in U_{ad}$, and

$$V_{\hat{T}}(x_0) + \epsilon \ge \int_0^{\hat{T}} f^0(x(t), u(t))\, dt + G(x(\hat{T})).$$

From (2.5) it follows that there exists $u(t+\hat{T}) = u(t; x(\hat{T}), T-\hat{T})$ for $t \in [0, T-\hat{T}]$ such that

$$\int_{\hat{T}}^{T} f^0(x(t), u(t))\, dt + G(x(T)) \leq G(x(\hat{T})).$$

Concatenation of the solutions on $[0, \hat{T}]$ and $[\hat{T}, T]$ implies that

$$\int_0^{\hat{T}} f^0(x(t), u(t))\, dt + G(x(\hat{T})) \geq \int_0^{T} f^0(x(t), u(t))\, dt + G(x(T)) \geq V_T(x_0),$$

and hence $V_{\hat{T}}(x_0) + \epsilon \geq V_T(x_0)$. Since $\epsilon > 0$ is arbitrary, $V_{\hat{T}}(x_0) \geq V_T(x_0)$ for $0 \leq \hat{T} \leq T$. □

A result related to Theorem 2.4 is contained in [JYH].

Remark 2.2. Theorem 2.4 asserts that a sufficient condition for monotonic decay of $T \to V_T(x)$ is the control Liapunov function property of G. As will be illustrated in section 2.2, this property requires, in some sense, that G be sufficiently large. Note also that αV, with V the value function of the infinite horizon problem, is a control Liapunov function if $\alpha \geq 1$. On the other hand, if $\alpha < 1$, then αV will not be a control Liapunov function in general, as can already be seen for linear-quadratic optimal control problems in dimension one.

For this purpose, consider

$$\text{(2.14)} \qquad \begin{cases} \min \frac{1}{2} \int_0^\infty q\, x^2(t)\, dt + \int_0^\infty u^2(t)\, dt \\ \text{subject to} \\ \dot{x}(t) = ax(t) + u(t), \quad x(0) = x_0, \end{cases}$$

where $q \geq 0$. The associated Riccati equation is given by

$$\text{(2.15)} \qquad 2ap_\infty - p_\infty^2 + q = 0,$$

with solutions $p_\infty^{\pm} = a \pm \sqrt{a^2+q}$. The optimal control in feedback form is given by $u^*(t) = -p_\infty^+ x^*(t)$, where $p_\infty^+ = a + \sqrt{a^2+q}$, and the optimal value function is $V(x) = \frac{1}{2} p_\infty^+ x^2$. Consider also the finite horizon problems

$$\text{(2.16)} \qquad \begin{cases} \min \frac{1}{2} \int_0^T q\, x^2(t)\, dt + \frac{1}{2} \int_0^T u^2(t)\, dt + \frac{1}{2} g x^2(T) \\ \text{subject to} \\ \dot{x}(t) = ax(t) + u(t), \quad x(0) = x_0. \end{cases}$$

The associated Liapunov differential equation is given by

$$\text{(2.17)} \qquad \frac{d}{dt} p(t) + 2ap(t) - p^2(t) + q = 0, \quad p(T) = g.$$

If p_T is a nonnegative solution to (2.17), then $u^*(t) = -p_T(t) x^*(t)$, $t \in [0, T]$, defines a feedback solution to (2.16) and the optimal value is given by $V_T(x) = \frac{1}{2} p_T(0) x^2$. Note that the two solutions to the algebraic equation (2.15) are steady states for (2.17) with p_∞^+ unstable and p_∞^- stable. All trajectories of (2.17) are monotone and nonintersecting. In particular, if $g \geq p_\infty^+$, then $p_\infty^+ \leq p_T(0) < p_{\hat{T}}(0)$ if $0 \leq \hat{T} < T$. If, on the other hand, $g \in (-p_\infty^-, p_\infty^+)$, then $p_{\hat{T}}(0) < p_T(0)$ for $0 \leq \hat{T} < T$. □

The following main result for receding horizon problems can be obtained from the definition of a control Liapunov function and Theorems 2.3–2.4.

THEOREM 2.5. *Assume that G is a control Liapunov function and that $(\bar{x}, \bar{u})$ is a solution to the receding horizon problem* (2.3)–(2.4) *on* $[0, \infty)$. *Then we have*

$$G(x_k) + \int_{T_{k-1}}^{T_k} f^0(\bar{x}(t), \bar{u}(t))\, dt \le G(x_{k-1})$$

and $V_T(\bar{x}(t)) \le V_T(x_{k-1})$ *for* $t \in [T_{k-1}, T_k]$ *and* $k = 1, 2, \ldots$. *Thus*

$$G(x_k) + \int_0^{kT} f^0(\bar{x}(t), \bar{u}(t))\, dt \le G(x_0)$$

for all $k \ge 1$. *Moreover, if we assume that* $V_T(x) \le \rho_T\, G(x)$ *for some* $\rho_T \le 1$ *and* $T \ge 0$ *independently of* $x \in R^n$, *then* $G(x_k) \le \rho_T^k\, G(x_0)$ *for all* $k \ge 1$.

Proof. The second assertion follows from the optimality principle

$$\int_{T_{k-1}}^{t} f^0(t, \bar{x}(t), \bar{u}(t))\, dt + V_{T_k - t}(x(t)) = V_T(x_{k-1}), \quad t \in [T_{k-1}, T_k],$$

and the monotonicity of V_T. If $V_T(x) \le \rho_T\, G(x)$, then $G(x_1) \le V_T(x_0) \le \rho_T\, G(x_0)$ since $f^0 \ge 0$. The final claim then follows by iteration. □

Note that if, in addition to the assumptions of Theorem 2.5, the level-set $S_{x_0} = \{x : g(x) \le G(x_0)\}$ is bounded, $G(x) \ne 0$ for $x \ne 0$, and $\rho_T < 1$, then $\lim_{k\to\infty} x_k = 0$.

From Theorem 2.5 we conclude that the receding horizon trajectory $\bar{x}(t)$ is confined in the set $\{x \in R^n : V_T(x) \le V(x_0)\}$. Concerning the assumption that $\rho_T \le 1$ in Theorem 2.5, note that ρ_T is nonincreasing with respect to T by Theorem 2.4. We refer to the following remark and section 2.2 for a discussion of cases which allow $\rho_T < 1$.

Remark 2.3. Let us briefly comment on a localization of the concept of a control Liapunov function. For $\alpha > 0$, we define the level-set $S_\alpha = \{x\colon G(x) \le \alpha\}$. A nonnegative function G with $G(0) = 0$ is called a local (by α)-control Liapunov function for (2.1)–(2.2) if for all $x \in S_\alpha$ and $T > 0$ there exists a control $u(\cdot; x, T) \in U_{ad}$ such that (2.5) holds. Theorems 2.3–2.5 can be appropriately modified to hold for local-control Liapunov functions if x, respectively, x_0 are chosen in S_α. If S is a compact subset of S_α not containing 0, and $f^0(x, u) > 0$ and $G(x) > 0$ for $x \ne 0$, then there exists $\rho_T < 1$ such that $G(x(T)) \le \rho_T\, G(x_0)$ for every $x_0 \in S$. In fact, if $x_0 \ne 0$, we have $f^0(\bar{x}(t), \bar{u}(t)) \ne 0$ for t in a subinterval of $[0, T]$, and hence $G(x(T)) < G(x_0)$. Compactness of S implies the existence of $\rho_T < 1$ such that $G(x(T)) \le \rho_T\ G(x_0)$ for all $x_0 \in S$.

2.2. Quadratic terminal penalty. In this section, we discuss the case when $G(x) = \frac{\alpha}{2}\,|x|^2,\ \alpha > 0$, serves as a control Liapunov function.

DEFINITION 2.6. *The control system* (2.1)–(2.2) *is called closed loop dissipative if there exists a locally Lipschitzian feedback law* $u = -K(x) \in U$ *such that*

$$f(x, -K(x)) \cdot (\alpha\, x) + f^0(x, -K(x)) \le 0$$

for some $\alpha > 0$ *and all* $x \in R^n$.

If (2.1)–(2.2) is closed loop dissipative, then $\frac{\alpha}{2}|x|^2$ is a control Liapunov function for (2.1)–(2.2) by Theorem 2.2. In the case when (2.1)–(2.2) is not closed loop dissipative, we have the following result, which allows us to use $\frac{\alpha}{2}\,|x|^2$ as a control Liapunov function.

THEOREM 2.7. *Let $G(x) = \frac{\alpha}{2}|x|^2$, and let $V(x)$ and $V_T(x)$ be the infinite and the finite horizon value functionals, respectively. We assume that for every $x \in R^n$ there exists an admissible control $u^*(t) = u^*(t;x)$ such that*

$$V(x^*(T)) + \int_0^T f^0(x^*(t), u^*(t))\, dt = V(x)$$

for all $T \geq 0$ and that the corresponding trajectory $x^*(t)$ satisfies $|x^*(t)| \leq M\, e^{-\omega t}|x|$, with $M \geq 0$, $\omega > 0$, independently of $t \geq 0$ and $x \in R^n$. Then

$$V_T(x) \leq V(x) + \frac{M^2\alpha}{2} e^{-2\omega T}|x|^2.$$

Moreover, if $V(x) \leq \frac{\beta}{2}\,|x|^2$, then

$$V_T(x) \leq \left(\frac{\beta}{2} + \frac{M^2\alpha}{2} e^{-2\omega T}\right)|x|^2 \leq \left(\frac{\beta}{\alpha} + M^2 e^{-2\omega T}\right) G(x).$$

Proof. Note that

$$V_T(x) \leq \int_0^T f^0(x^*(t), u^*(t))\, dt + V(x^*(T)) + G(x^*(T)) - V(x^*(T)) \leq V(x) + G(x^*(T)),$$

which implies the first assertion. The second assertion simply follows from the first one. ∎

Theorem 2.5 implies that for every $\alpha > \beta$ there exists $\bar{T} > 0$ such that for $T \geq \bar{T}$ we have $V_T(x) \leq \rho_T\, G(x)$ with $\rho_T < 1$, and thus Theorem 2.5 applies for $T \geq \bar{T}$.

2.3. General case. In this section, we discuss the case when the value functional $V(x)$ of the infinite horizon optimal control problem (2.1)–(2.2) does not exist for every $x \in X$. This includes, for example, the disturbance attenuation problem: consider the problem (2.1)–(2.2) with the linear control system $f = Ax + Bu + d$ and quadratic performance $f^0 = |x|^2 + |u|^2$, where a nonzero constant vector d denotes the disturbance. In order to deal with this case, the following two approaches suggest themselves. One of them is the introduction of a discounted cost functional

$$\min_{u \in U_{ad}} \quad \int_0^\infty e^{-\rho t}\, f^0(x(t), u(t))\, dt, \quad \rho > 0.$$

The other is the ergodic cost functional

$$\min_{u \in U_{ad}} \quad \lim_{T \to \infty} \frac{1}{T} \int_0^T f^0(x(t), u(t))\, dt.$$

We discuss the ergodic cost functional and return to the discounted cost functional at the end of this subsection.

DEFINITION 2.8. *A nonnegative continuous function G is called a control λ-Liapunov function for (2.1)–(2.2) if for every $x_0 \in R^n$ and $T > 0$ there exists a control $u = u(\cdot; x, T) \in U_{ad}$ satisfying*

$$\int_0^T f^0(x(t), u(t))\, dt + G(x(T)) \leq G(x_0) + \lambda T, \tag{2.18}$$

where $x(\cdot)$ is a solution to (2.1).

We have the following relationship of control λ-Liapunov functions to a differential inequality.

THEOREM 2.9. *Assume that G is a nonnegative C^1-function and that $\{x \in R^n : G(x) \le \alpha\}$ is bounded for every $\alpha > 0$.*

(a) *If there exists a locally Lipschitzian function $u = \Phi(x) \in U$ such that for all $x \in R^n$*

$$f(x,u) \cdot G_x(x) + f^0(x,u) \le \lambda, \tag{2.19}$$

then G is a control λ-Liapunov function for (2.1)–(2.2).

(b) *If U is compact and for all $x_0 \in R^n$ and $\delta > 0$ there exists a $\tau = \tau_{\delta,x_0} > 0$ such that*

$$|x(t) - x_0| \le \delta \quad \text{for all } t \in [0,T] \text{ and } u \in \mathcal{U},$$

where $\mathcal{U} = \{u \in U_{ad} : \int_0^\tau f^0(x(t),u(t))\,dt \le G(x) + \lambda T\}$, then (2.18) *implies that for each $x \in R^n$ there exists $u \in U$ such that* (2.19) *holds.*

The proof is analogous to that for Theorem 2.2, and it is therefore omitted.

THEOREM 2.10 (monotonicity). *Assume that G is a control λ-Liapunov function for* (2.1)–(2.2). *Then $V_{\hat{T}}(x_0) - \lambda T \ge V_T(x_0) - \lambda \hat{T}$ for $0 \le \hat{T} \le T$.*

The proof is analogous to that of Theorem 2.3.

THEOREM 2.11. *Assume that G is a control λ-Liapunov function and $f^0(x,u) \ge \omega G(x)$ for some $\omega > 0$ and all $x \in R^n$ and $u \in U$. Then if $(u(t),x(t))$ minimizes $\int_0^T f^0(x(t),u(t))\,dt + G(x(T))$ over $u \in U_{ad}$ subject to* (2.1), *we have*

$$G(x(T)) \le e^{-\omega T}(G(x_0) + \lambda T) + \lambda\left(\frac{1}{\omega} - e^{-\omega T}\left(T + \frac{1}{\omega}\right)\right).$$

Proof. By the optimality principle,

$$\int_t^\tau f^0(x(s),u(s))\,ds + V_{T-\tau}(x(\tau)) = V_{T-t}(x(t)) \tag{2.20}$$

for every $0 \le t \le \tau \le T$. From (2.20) it follows that $t \to g(t) = V_{T-t}(x(t))$ is a $W^{1,1}$-function. By Definition 2.8 and the lower bound on f^0 it follows that

$$\omega \int_t^\tau V_{T-s}(x(s))\,ds + V_{T-\tau}(x(\tau)) \le V_{T-t}(x(t)) + \omega\lambda \int_t^\tau (T-s)\,ds,$$

and, consequently,

$$\omega g(t) + \frac{d}{dt}\,g(t) \le \omega\lambda(T-t) \text{ for a.e. } t \in [0,T].$$

Multiplying by $e^{\omega t}$ and integrating on $[0,T]$ imply

$$e^{\omega T} g(T) - g(0) \le \lambda\left(\frac{1}{\omega}e^{\omega T} - T - \frac{1}{\omega}\right).$$

We conclude that

$$G(x(T)) \le e^{-\omega T} V_T(x_0) + \lambda\left(\frac{1}{\omega} - e^{-\omega T}\left(T + \frac{1}{\omega}\right)\right),$$

and the claim follows since $V_T(x_0) \le G(x_0) + \lambda T$. □

THEOREM 2.12 (stability). *Assume that $(\bar{x}, \bar{u})$ is a solution to the receding horizon problem* (2.3)–(2.4). *If G is a control λ-Liapunov function, then*

$$G(x_k) + \int_0^{kT} f^0(\bar{x}(t), \bar{u}(t))\, dt - \lambda k T \le G(x_0) \tag{2.21}$$

for all $k = 1, 2, \ldots$. Moreover, if the assumption of Theorem 2.11 *holds, then*

$$\begin{aligned} G(x_k) &\le e^{-k\omega T} G(x_0) \\ &+ \left(\lambda T e^{-\omega T} + \lambda \left(\frac{1}{\omega} - e^{-\omega T} \left(T + \frac{1}{\omega} \right) \right) \right) \frac{1 - e^{-k\omega T}}{1 - e^{-\omega T}}. \end{aligned} \tag{2.22}$$

Note that Theorem 2.12 implies that

$$\lim_{\tau \to \infty} \frac{1}{\tau} \int_0^{\tau} f^0(\bar{x}(t), \bar{u}(t))\, dt \le \lambda$$

and that $\{G(x_k)\}$ is uniformly bounded in k.

Proof of Theorem 2.12. Inequality (2.21) follows by repeated application of Definition 2.8. An induction argument and Theorem 2.11 imply (2.22). □

For the particular case when $\lambda = 0$, we obtain the following theorem as a corollary to Theorems 2.11 and 2.12.

THEOREM 2.13. *Under the assumptions of Theorem* 2.11 *with $\lambda = 0$, we have*

$$G(x(T)) \le e^{-\omega T} V_T(x) \le e^{-\omega T} G(x_0).$$

Moreover, if $(\bar{x}, \bar{u})$ denotes a solution to the receding horizon problem (2.3)–(2.4), *then*

$$G(x_k) \le e^{-k\omega T} G(x_0).$$

Remark 2.4. We discuss the particular case when the control λ-Liapunov function can be taken as a quadratic penalty function $G(x) = \frac{\alpha}{2}|x|^2$, $\alpha > 0$.

Consider (2.1) with right-hand side

$$f(x, u) = \tilde{f}(x, u) + d, \tag{2.23}$$

where $d \in R^n$. Assume that (2.1)–(2.2) with f replaced by $\tilde{f}$ is uniformly closed loop dissipative; i.e., there exist $\alpha > 0$, $\kappa > 0$, and a locally Lipschitz continuous feedback law $u = -K(x) \in U$ such that

$$\tilde{f}(x, -K(x)) \cdot (\alpha x) + f^0(x, -K(x)) \le -\kappa |x|^2$$

for all $x \in R^n$. Then

$$f(x, -K(x)) \cdot (\alpha x) + f^0(x, -K(x)) \le -\kappa |x|^2 + \alpha\, dx \le \frac{\alpha^2}{4\kappa} d^2,$$

and hence $G(x) = \frac{\alpha}{2}|x|^2$ is a control λ-Liapunov function with $\lambda = \frac{\alpha^2}{4\kappa} d^2$ for (2.1), (2.2) with f given by (2.23).

Right-hand sides of the type (2.23) occur in disturbance attenuation and in tracking-type problems. As for the latter, assume that A is exponentially stable, and consider for $z \in R^n$

$$\begin{cases} \inf \frac{1}{2} \int_0^T |y(t) - z|^2 dt + \frac{\gamma}{2} \int_0^T |u(t)|^2 dt, \\ \frac{d}{dt} y(t) = Ay(t) + Bu(t), \quad y(0) = y_0. \end{cases}$$

Then, setting $x = y - z$, we find the equivalent problem

$$\begin{cases} \inf \frac{1}{2}\int_0^T |x(t)|^2 dt + \frac{\gamma}{2}\int_0^T |u(t)|^2 dt, \\ \frac{d}{dt}x(t) = Ax(t) + Bu(t) + Az, \quad x(0) = y_0 + z, \end{cases}$$

where the right-hand side is of the form (2.23).

Remark 2.5. Let us briefly return to the discounted cost functional formulation. In this case, a nonnegative functional G is called a d-control λ-Liapunov function if there exists $\lambda > 0$ such that for every $x \in R^n$ and $T > 0$ there exists a control $u = u(\cdot; x, T) \in U_{ad}$ satisfying

$$\int_0^T e^{-\lambda t} f^0(x(t), u(t))\, dt + e^{-\lambda T} G(x(T)) \leq G(x_0), \tag{2.24}$$

where $x(\cdot)$ is a solution to (2.1). For d-control λ-Liapunov functionals, the analogues of Theorems 2.2–2.4 can readily be derived. The receding horizon control strategy based on the discounted cost functional requires us to successively solve

$$\min_{u \in U_{ad}} \int_{(k-1)T}^{kT} e^{\lambda t} f^0(x(t), u(t))\, dt + e^{-\lambda T} G(x(kT)) \tag{2.25}$$

subject to $\frac{d}{dt}x(t) = f(x(t), u(t)), x((k-1)T) = x_{k-1}$, where $(\bar{x}_k, \bar{u}_k)$ is the optimal pair for the kth horizon $[(k-1)T, kT]$, and $x_k = \bar{x}_k(kT)$. Assuming existence of solutions $(\bar{x}_k, \bar{u}_k)$ to (2.25) it follows from (2.25) that $\int_0^\infty e^{-\lambda t} f^0(\bar{x}(t), \bar{u}(t))\, dt \leq G(x_0)$, provided that G is a d-control λ-Liapunov functional. Here $(\bar{x}, \bar{u})$ arises from concatenation of the solutions $(\bar{x}_k, \bar{u}_k)$ on $[(k-1)T, kT]$. Comparing the estimate $\int_0^\infty e^{-\lambda t} f^0(\bar{x}(t), \bar{u}(t))\, dt \leq G(x_0)$ to the asymptotic properties that were obtained in Theorem 2.12 for the ergodic formulation, the discounted cost formulation gives weaker properties.

3. Receding horizon control problems: Discrete time. In this section, we consider the discrete infinite time horizon problem

$$\inf_{u_j \in U} \sum_{j=1}^{\infty} \Delta t\, f^0(x_{j-1}, u_j) \tag{3.1}$$

subject to

$$x_j = x_{j-1} + \Delta t\, F(x_{j-1}, u_j), \quad x_0 \text{ given}, \tag{3.2}$$

with $\Delta t > 0$. Here (3.2) represents a finite difference discretization to (2.1), where the continuous function $F\colon R^n \times U \to R^n$ may also depend on Δt. For example, in the case of the explicit Euler rule,

$$F(x, u) = f(x, u),$$

and in the case of the implicit midpoint rule,

$$F(x, u) = f(\hat{x}, u),$$

where $\hat{x} \in R^n$ satisfies

$$\hat{x} - x = \frac{\Delta t}{2} f(\hat{x}, u).$$

The minimal value functional associated to (3.1)–(3.2) will be denoted by $V(x)$. Together with (3.1)–(3.2) we consider the finite time problem

$$\inf_{u_j \in U} \sum_{j=1}^{N} \Delta t\ f^0(x_{j-1}, u_j) + G(x_N) \tag{3.3}$$

subject to (3.2). Here $N \in \mathbb{N}$, and $G\colon R^n \to R^+$ is a continuous function. The value function associated to (3.3) is denoted by $V_N(x)$. The receding horizon control approach to (3.1)–(3.2) consists in iteratively solving problem (3.3):

$$\min_{u_j \in U} \sum_{j=(k-1)N+1}^{kN} \Delta t\ f^0(x_{j-1}, u_j) + G(x_{kN}) \tag{3.4}$$

subject to

$$x_j = x_{j-1} + \Delta t\ F(x_{j-1}, u_j),\ x_{(k-1)N} = \bar{x}_{(k-1)N}, \tag{3.5}$$

where $(\bar{x}_{(k-1)N+j}, \bar{u}_{(k-1)N+j})$, $1 \le j \le N$, is the sequence of optimal pairs for the kth optimal control problem, where $k \in \mathbb{N}$.

3.1. Regulator case.

DEFINITION 3.1. *A nonnegative continuous functional G with $G(0) = 0$ is called a discrete control Liapunov functional if for all $x_0 \in R^n$ there exists $u \in U$ such that*

$$\Delta t\ f^0(x_0, u) + G(x_0 + \Delta t\, F(x_0, u)) \le G(x_0). \tag{3.6}$$

If G is a discrete control Liapunov functional, then for every $x_0 \in R^n$ and $N \in \mathbb{N}$ there exists a sequence $\{u_j\}_{j=1}^N$ in U such that

$$\sum_{j=1}^{N} \Delta t\ f^0(x_{j-1}, u_j) + G(x_N) \le G(x_0), \tag{3.7}$$

where $\{x_j\}_{j=1}^N$ satisfies (3.2). In fact, (3.6) implies for each x_{j-1} the existence of $u_j \in U$, $j = 1, \dots, N$, such that

$$\Delta t\ f^0(x_{j-1}, u_j) + G(x_{j-1} + \Delta t\ F(x_{j-1}, u_j)) \le G(x_{j-1}).$$

Summation with respect to j and (3.2) imply (3.7).

The following result is the analogue to Theorem 2.2 for discrete control Liapunov functions.

THEOREM 3.2. *Assume that G is a nonnegative C^1-function with $G(0) = 0$.*

(a) *If G is a convex discrete control Liapunov function, then for each $x \in R^n$ there exists $u \in U$ such that*

$$F(x, u) \cdot G_x(x) + f^0(x, u) \le 0.$$

(b) *Assume that G_x and F are Lipschitz continuous, with $F(0, 0) = 0$, and that there exists a Lipschitz continuous function $u = \phi(x) \in U$, with $\phi(0) = 0$, such that*

$$F(x, \phi(x)) \cdot G_x(x) + f^0(x, \phi(x)) \le -\kappa |x|^2$$

for some $\kappa > 0$ independent of $x \in R^n$. Then there exists $\bar{\Delta} > 0$ such that G is a discrete control Liapunov function for $\Delta t \le \bar{\Delta}$.

Proof. (a) follows directly from (3.6) and the convexity of G. To verify (b), choose $x \in R^n$ and $u = \phi(x)$. Let $K_i, i = 1, 2, 3$, denote the Lipschitz constants of F, ϕ, and G_x. By the Lagrange form of the mean value theorem we find

$$\begin{aligned}
&\Delta t\ f^0 c(x,u) + G(x + \Delta t\ F(x,u)) - G(x) \\
&= \Delta t\ f^0(x,u) + \Delta t\, G_x(x) \cdot F(x,u) \\
&\qquad + \Delta t \int_0^1 (G_x(x + s\Delta t\, F(x,u)) - G_x(x)) F(x,u) ds,
\end{aligned}$$

and, therefore,

$$\begin{aligned}
&\Delta t\ f^0(x,u) + G(x + \Delta t\, F(x,u)) - G(x) \\
&\le -\kappa \Delta t |x|^2 + \frac{(\Delta t)^2}{2} K_1^2 K_3 (|x|^2 + |\phi(x)|^2) \le \Delta t \left(\frac{\Delta t}{2} K_1^2 K_3 (1 + K_2^2) - \kappa \right) |x|^2,
\end{aligned}$$

and the claim follows. ∎

THEOREM 3.3. *Assume that G is a discrete control Liapunov function. Then $V(x_0) \le V_N(x_0) \le V_{\hat{N}}(x_0) \le G(x_0)$ for all $x_0 \in R^n$ and $0 \le \hat{N} \le N$.*

The proof is similar to that of Theorems 2.3 and 2.4, and it is therefore omitted. For the discrete receding horizon problem, we find the following theorem by the argument implying (3.7).

THEOREM 3.4. *Assume that G is a discrete Liapunov function and that $\{(\bar{x}_{j-1}, \bar{u}_j)\}_{j=1}^{kN}$ is a solution to the discrete receding horizon problem. Then we have*

$$\sum_{j=1}^{kN} \Delta t\ f^0(\bar{x}_{j-1}, \bar{u}_j) + G(\bar{x}_{kN}) \le G(x_0).$$

3.2. Quadratic terminal penalty. We discuss situations in which $G(x) = \frac{\alpha}{2}|x|^2, \alpha > 0$, can serve as a discrete control Liapunov functional.

DEFINITION 3.5. *The discrete infinite time horizon problem* (3.1)–(3.2) *is closed loop dissipative if there exist a Lipschitz continuous feedback law $u = -K(x) \in U$, for $x \in R^n$, and $\kappa > 0$, $\alpha > 0$ such that*

$$F(x, -K(x)) \cdot (\alpha x) + f^0(x, -K(x)) \le -\kappa |x|^2$$

for all $x \in R^n$.

If (3.1)–(3.2) is closed loop dissipative, F is Lipschitz continuous, $F(0,0) = 0$, and $K(0) = 0$, then $G(x) = \frac{\alpha}{2}|x|^2$ is a discrete Liapunov functional whenever Δt is sufficiently small. This follows immediately from Theorem 3.2 (b). In the case when (3.1)–(3.2) is not closed loop dissipative, the following result gives sufficient conditions which imply that $\frac{\alpha}{2}|x|^2$ is a discrete control Liapunov function.

THEOREM 3.6. *Let $G(x) = \frac{\alpha}{2}|x|^2$, and assume that there exist $M \ge 1$ and $\rho > 0$ such that for every $x_0 \in R^n$ there is an optimal control $\{u_j^*\}_{j=1}^{\infty}$ with associated states $\{x_j^*\}_{j=1}^{\infty}$ satisfying $|x_j^*| \le M\, e^{-\rho j} |x_0|$ for $j = 1, 2, \ldots$. If, moreover, $V(x) \le \frac{\beta}{2}|x|^2$ for some $\beta > 0$, then $V_N(x_0) \le (\frac{\beta}{2} + M^2 e^{-2\rho N}) G(x_0)$ for every $x_0 \in R^n$ and $N \ge 1$. Further, there exists $\bar{N} > 0$ such that for all $N \ge \bar{N}$*

$$V_N(x_0) \le \rho_N G(x_0)$$

for some $\rho_N < 1$ independent of $x \in R^n$.

Proof. For every N we have

$$\begin{aligned} V_N(x_0) &= \Delta t \sum_{j=1}^{N} f^0(x^*_{j-1}, u^*_j) + G(x^*_N) - V(x^*_N) + V(x^*_N) \\ &\le V(x_0) + G(x^*_N) = V(x_0) + \frac{\alpha}{2} M^2 e^{-2\rho N} |x_0|^2 \le \left(\frac{\beta}{2\alpha} + M^2 e^{-2\rho N} \right) G(x_0). \end{aligned}$$

The first assertion is a consequence of this estimate. The second one follows directly from the first. ∎

3.3. General case. Here we discuss the ergodic case for the discrete receding horizon problem.

DEFINITION 3.7. *A nonnegative continuous function G is called a discrete control λ-Liapunov functional for* (3.1)–(3.2) *if for every $x \in R^n$ there exists $u \in U$ satisfying*

$$\Delta t\, f^0(x,u) + G(x + \Delta t\, F(x,u)) \le G(x) + \lambda\, \Delta t. \tag{3.8}$$

If G is a discrete control λ-Liapunov functional for (3.1)–(3.2), then for every $x \in R^n$ and $N \in \mathbb{N}$ there exists a sequence $\{u_j\}_{j=1}^N$ in U such that

$$\sum_{j=1}^{N} \Delta t\, f^0(x_{j-1}, u_j) + G(x_N) \le G(x_0) + \lambda\, N \Delta t, \tag{3.9}$$

where $\{x_j\}_{j=1}^N$ satisfies (3.2). Moreover, we have

$$V_N(x_0) - \lambda\, \hat{N} \Delta t \le V_{\hat{N}}(x_0) - \lambda\, N \Delta t \tag{3.10}$$

for $0 \le \hat{N} \le N$. Under appropriate conditions, one also obtains the analogue of Theorem 3.2 relating (3.8) to $F(x,u) \cdot G_x(x) + f^0(x,u) \le \lambda$. We shall not pursue this aspect and shall rather turn to the asymptotic behavior of the discrete receding horizon problem. We require the following result.

THEOREM 3.8. *Assume that G is a discrete control λ-Liapunov functional and that $f^0(x,u) \ge \omega G(x)$ for some $\omega \in (0,1]$ independently of $x \in R^n$ and $u \in U$. Then if $\{(x_{j-1}, u_j)\}_{j=1}^N$ is a solution to* (3.3)*, we have*

$$G(x_N) \le e^{-\omega N \Delta t}(G(x_0) + \lambda N \Delta t) + \lambda^{\omega \Delta t} \left[\frac{1}{\omega} - e^{-\omega N \Delta t} \left(N \Delta t + \frac{1}{\omega} \right) \right].$$

Proof. For every $0 < N_1 \le N_2 \le N$ we have

$$\sum_{J=N_1}^{N_2} \Delta t\, f^0(x_{j-1}, u_j) + \sum_{j=N_2+1}^{N} \Delta t\, f^0(x_{j-1}, u_j) + G(x_N) = V_N(x_0),$$

and hence by the optimality principle

$$\sum_{j=N_1}^{N_2} \Delta t\, f^0(x_{j-1}, u_j) + V_{N-N_2}(x_{N_2}) = V_{N-N_1+1}(x_{N_1-1}),$$

and by assumption

$$\omega \sum_{j=N_1}^{N_2} \Delta t\, G(x_{j-1}) + V_{N-N_2}(x_{N_2}) \le V_{N-N_1+1}(x_{N_1-1}).$$

Using (3.9) with $x = x_{j-1}$, we find

$$\omega \sum_{j=N_1}^{N_2} \Delta t\, V_{N-j+1}(x_{j-1}) + V_{N-N_2}(x_{N_2}) \le V_{N-N_1+1}(x_{N_1-1}) + \omega\lambda(\Delta t)^2 \sum_{j=N_1}^{N_2} (N-j+1).$$

Setting $r_j = V_{N-j}(x_j)$ and $N_1 = N_2$ implies that

$$\omega\Delta t\, r_{j-1} + r_j - r_{j-1} \le \omega\lambda(\Delta t)^2(N-j+1).$$

Multiplying with $e^{\omega j\Delta t}$, summing with respect to j, and rearranging terms give

$$\begin{aligned} e^{\omega\Delta t N} r_N \quad &+ \quad (\omega\Delta t - 1)e^{\omega\Delta t} r_0 + (\omega\Delta t + e^{-\omega\Delta t} - 1)\sum_{j=2}^{N} e^{\omega j\Delta t} r_{j-1} \\ &\le \quad \omega\lambda(\Delta t)^2 \sum_{j=1}^{N} e^{\omega j\Delta t}(N-j+1), \end{aligned}$$

and, consequently,

$$\begin{aligned} r_N \quad &\le e^{-\omega(N-1)\Delta t}(1-\omega\Delta t) r_0 + \omega\lambda(\Delta t)^2 e^{-\omega N\Delta t} \sum_{j=1}^{N} e^{\omega j\Delta t}(N-j+1) \\ &\le e^{-\omega N\Delta t} r_0 + \lambda e^{\omega\Delta t}\left[\frac{1}{\omega} - e^{-\omega N\Delta t}\left(N\Delta t + \frac{1}{\omega}\right)\right]. \end{aligned}$$

By the definition of r_j and (3.9), the last inequality implies

$$G(x_N) \le e^{-\omega N\Delta t}(G(x_0) + \lambda N\Delta t) + \lambda e^{\omega\Delta t}\left[\frac{1}{\omega} - e^{-\omega N\Delta t}\left(N\Delta t + \frac{1}{\omega}\right)\right]. \qquad \square$$

THEOREM 3.9. *Let $\{\bar{x}_{j-1}, \bar{u}_j\}_{j=1}^{\infty}$ denote a solution to the discrete receding horizon problem* (3.4)–(3.5)*. Then, under the assumptions of Theorem* 3.8*, we have*

$$G(x_{kN}) + \sum_{j=1}^{kN} \Delta t\; f^0(x_{j-1}, u_j) \le G(x_0) + \lambda kN\Delta t \tag{3.11}$$

and

$$G(x_{kN}) \le e^{-\omega kN\Delta t} G(x_0) + \mu\, \frac{1 - e^{-k\omega N\Delta t}}{1 - e^{-\omega N\Delta t}} \tag{3.12}$$

for all $k = 1, 2, \dots$, where $\mu = \lambda N\Delta t\, e^{-\omega N\Delta t} + \lambda\, e^{\omega N\Delta t}[\frac{1}{\omega} - e^{-\omega N\Delta t}(N\Delta t + \frac{1}{\omega})]$.

Theorem 3.9 implies that

$$\lim_{k\to\infty} \frac{1}{kN} \sum_{j=1}^{kN} f^0(x_{j-1}, u_j) \le G(x) + \lambda$$

and that $\{G(x_{kN})\}_{k=1}^{\infty}$ is bounded.

Proof of Theorem 3.9. Inequality (3.11) follows from (3.9). An induction argument over k utilizing the estimate given in Theorem 3.8 implies (3.12). $\square$

4. Approximation and stabilization by the discrete receding horizon problem. In this section, we therefore analyze some aspects of controlling (2.1) by means of the one-step discrete receding horizon problems

$$\min_{u\in U} \{\Delta t\ f^0(x_0,u) + G(x_0 + \Delta t\, F(x_0,u))\}. \tag{4.1}$$

Here $x_1 = x_0 + \Delta t\, F(x_0,u)$ is a one-step approximation to $\frac{d}{dt}x(t) = f(x(t),u(t))$, $x(0) = x_0$, with step-size Δt. In the notation of section 3, we have $N = 1$. One of the motivations for considering this case is given by large scale optimal control problems; see, e.g., [CHK, CTMC] and the references given there. Computing the solutions to (2.3) or (3.3) can still be very expensive, and one therefore resorts to the extreme case consisting of the one-step receding horizon strategy.

We first consider the case when the full state $x(t)$ of (2.1) can be observed. Let G be a locally Lipschitz continuous control Liapunov function for (2.1)–(2.2) such that level-sets $\{x\colon G(x) \le \alpha\}, \alpha > 0$, are bounded in R^n. We assume that for every $\alpha > 0$ there exists a continuous nondecreasing function ϵ with $\epsilon(0) = 0$ such that for each x_0 with $G(x_0) \le \alpha$

$$|\bar u(t) - u| + |\bar x(t) - x_0| \le \epsilon(\Delta t) \ \text{ for all } \ t \in [0,\Delta t]. \tag{4.2}$$

Here u minimizes (4.1), and $(\bar x, \bar u)$ is an optimal pair for the finite horizon problem

$$\min \int_0^{\Delta t} f^0(x(t),u(t))\,dt + G(x(\Delta t)) \tag{4.3}$$

subject to $\frac{d}{dt}\,x(t) = f(x(t),u(t)), x(0) = x_0$. It is further assumed that there exists $\omega > 0$ such that

$$V_{\Delta t}(x_0) \le (1 - \omega\Delta t)G(x_0) \ \text{ for all } \ x_0 \in R^n \tag{4.4}$$

and all sufficiently small $\Delta t > 0$. Finally, we shall assume that $U \subset R^m$ is bounded. In the two theorems below, this requirement can be replaced by assuming that

$$\big\{u = \arg\min\,[\Delta t f^0(x_0,u) + G(x_0 + \Delta t F(x_0,u))]\colon \Delta t \in [0,1], x_0 \in S\big\} \tag{4.5}$$

is a bounded subset of R^m whenever S is bounded in R^n. Note that (4.5) can be verified, for example, if f^0 is quadratic in x and u, G is quadratic, and F is affine in u.

We are now prepared to analyze the following strategy: given x_{k-1}, solve (4.1) with initial condition $x_0 = x_{k-1}$ for u_k, then advance (2.1) on $[(k-1)\Delta t, k\Delta t]$ by means of

$$\frac{d}{dt}\,x(t) = f(x(t),u_k), \quad x((k-1)\Delta t) = x_{k-1}, \tag{4.6}$$

and set $x_k = x(k\Delta t)$. Henceforth we put $t_k = k\Delta t$.

THEOREM 4.1. *Assume that G is a locally Lipschitz continuous control Liapunov function with bounded level-sets, that* (4.2) *and* (4.4) *hold, and that U is bounded. Then for every $r > 0$ there exists $C > 0$ such that, for all sufficiently small Δt and all k,*

$$G(x_k) \le (1 - \omega\Delta t)^k G(x_0) + \frac{C}{\omega}\epsilon(\Delta t)$$

and

$$\int_0^{k\Delta t} f^0(x(t),u(t))\,dt \le (1-\omega\Delta t)G(x_0) + \frac{C\,\epsilon(\Delta t)}{\omega(1-\omega\Delta t)^{k-1}}$$

whenever $|x_0| \le r$. *Here* $x(\cdot)$ *is the concatenation of the solutions to* (4.6), $u(t) = u_k$ *on* $[t_{k-1}, t_k]$, *and* $u_k \in U$ *minimizes* $[\Delta t\, f^0(x_{k-1},u) + G(x_{k-1} + F(x_{k-1},u))]$.

Proof. Throughout, we assume that Δt is sufficiently small so that $\Delta t < 1$ and $\epsilon(\Delta t) < 1$. Let $r > 0$, and set $\alpha = \max\{G(x)\colon |x| \le r\}$. Then the level-set $\{x\colon G(x) \le \alpha\}$ is contained in a closed ball $B(0,\rho)$ with radius $\rho \ge r$ centered at the origin. Let $\|f\|, \|f^0\|$, and $\|G\|$ denote the Lipschitz constants of f, f^0, and G on $B(0,\rho+1)\times U$ and $B(0,\rho+1)$, respectively. Let $x_0 \in R^n$ be such that $|x_0| \le r$.

First we argue that for x_{k-1} with $G(x_{k-1}) \le \alpha$, $k = 1,\dots,$ (4.6) admits a solution. Let $(\hat{x},\hat{u})$ be a solution to the finite horizon problem (4.3) with initial condition $x_0 = x_{k-1}$, and denote by $(\bar{x},\bar{u})$ its translation from $[0,\Delta t]$ to $[t_{k-1},t_k]$. By Gronwall's lemma we have the a priori estimate

$$|x(t) - \bar{x}(t)| \le \|f\|\Delta t\,\epsilon(\Delta t)\,e^{\|f\|\Delta t} \quad \text{for} \quad t \in [t_{k-1},t_k], \tag{4.7}$$

and, therefore, by (4.2)

$$\begin{aligned} |x(t)| &\le |x_{k-1}| + |x_{k-1} - \bar{x}(t)| + |x(t) - \bar{x}(t)| \\ &\le \rho + \epsilon(\Delta t) + \|f\|\,\Delta t\,\epsilon(\Delta t)\,e^{\|f\|\Delta t} \le \rho + 1 \end{aligned}$$

for all $t \in [t_{k-1},t_k]$, provided that Δt is sufficiently small. In particular, this implies the existence of a solution to (4.6) on $[t_{k-1},t_k]$. To estimate $G(x_k)$, we utilize (4.2), (4.4), and (4.7):

$$\begin{aligned} G(x_k) &+ \int_{t_{k-1}}^{t_k} f^0(x(t),u(t))\,dt = G(x_k) - G(\bar{x}(t_k)) + G(\bar{x}(t_k)) \\ &+ \int_{t_{k-1}}^{t_k} (f^0(x(t),u(t)) - f^0(\bar{x}(t),\bar{u}(t)))\,dt + \int_{t_{k-1}}^{t_k} f^0(\bar{x}(t),\bar{u}(t))\,dt \\ &\le V_{\Delta t}(x_{k-1}) + (\|G\| + \|f^0\|\Delta t)\,\|x - \bar{x}\|_{C([t_{k-1},t_k])} \\ &+ \Delta t\,\|f^0\|\,\|u - \bar{u}\|_{C([t_{k-1},t_k])} \\ &\le (1-\omega\Delta t)G(x_{k-1}) + C\,\Delta t\,\epsilon(\Delta t), \end{aligned}$$

where $C = (\|G\| + \|f^0\|)\|f\|\,e^{\|f\|} + \|f^0\|$. In the above estimate, we used the facts that $|x(t)| \le \rho+1, |\bar{x}(t))| \le \rho+1$, $u_k \in U$, and $u(t) \in U$ for all $t \in [t_{k-1},t_k]$ and the assumption that $\epsilon(\Delta t) \le 1$. Hence

$$G(x_k) + \int_{t_{k-1}}^{t_k} f^0(x(t),u(t))\,dt \le (1-\omega\Delta t)G(x_{k-1}) + C\,\Delta t\,\epsilon(\Delta t), \tag{4.8}$$

and, therefore, $G(x_k) \le \alpha$ for all Δt sufficiently small. Repeated application of (4.8) implies

$$G(x_k) + \sum_{i=0}^{k-1}(1-\omega\Delta t)^i \int_{t_{k-i-1}}^{t_{k-i}} f^0(x(t),u(t))\,dt \le (1-\omega\Delta t)^k G(x_0) + \frac{C\epsilon(\Delta t)}{\omega},$$

and the desired estimates follow. □

Next we consider the case when $x(t)$ is not observed. We assume that $G(x)$ is a discrete control Liapunov function satisfying

$$V_1(x) \leq (1 - \omega\,\Delta t)\, G(x), \tag{4.9}$$

where $\omega > 0$ and V_1 is the optimal value function for (4.1). We analyze the strategy of solving the discrete time optimal control problem (3.4) for $k = 1, \dots$ (and $N = 1$) to obtain a sequence of optimal controls $\{\bar{u}_k\}$ with corresponding states $\{\bar{x}_k\}$ and to utilize these controls in (2.1) with $u(t) = \bar{u}_k$ on (t_{k-1}, t_k). We assume that

$$(f(x_1, u) - f(x_2, u), x_1 - x_2) \leq \beta\, |x_1 - x_2|^2 \tag{4.10}$$

for some $\beta \in R$ and all $x_1, x_2 \in R^n$, and $u \in U$, and we define the linear interpolation

$$\xi(t) = \bar{x}_{k-1} + (t - t_{k-1}) \frac{\bar{x}_k - \bar{x}_{k-1}}{\Delta t} \quad \text{on} \quad (t_{k-1}, t_k). \tag{4.11}$$

Note that

$$\frac{d}{dt}\xi(t) = f(\xi(t), u(t)) + d(t),$$

where

$$d(t) = F(\bar{x}_{k-1}, \bar{u}_k) - f(\xi(t), \bar{u}_k) \quad \text{on} \quad (t_{k-1}, t_k). \tag{4.12}$$

THEOREM 4.2. *Assume that G is a locally Lipschitz continuous discrete control Liapunov function with bounded level-sets, that* (4.9) *and* (4.10) *are satisfied, that U is bounded, and that there exist some $x \in R^n$ and $\hat{u} \in L^1(0, T; R^m)$ with $T = k\,\Delta t$ such that* (2.1) *admits a solution $\hat{x}$ on $[0, T]$. Then for every $r > 0$ there exists a constant $C = C(r, k\Delta t)$ such that*

$$G(x(k\Delta t)) \leq (1 - \omega\Delta t)^k G(x_0) + C\,\tilde{\epsilon}(\Delta t) \tag{4.13}$$

and

$$\int_0^{k\Delta t} f^0(x(t), u(t))\, dt \leq (1 - \omega\Delta t) G(x_0) + C(\tilde{\epsilon}(\Delta t) + k(\Delta t)^2)$$

for every x_0 with $|x_0| \leq r$. Here $x(t)$ is the solution to (2.1) *with $x = x_0$,*

$$u = \bar{u}_j \quad \text{on} \quad (t_{j-1}, t_j), \tag{4.14}$$

with $\bar{u}_j$ the solution of the discrete receding horizon problem (3.4) *with $N = 1$, and $\tilde{\epsilon}(\Delta t) = \int_0^{k\Delta t} d(s)\, ds$.*

Proof. Choose $r > 0$. Since level-sets of G are bounded, there exists $\rho > 0$ such that the set $\{x\colon G(x) \leq r\}$ as well as the trajectory $\{\hat{x}(t)\colon t \in [0, T]\}$ are contained in $B(0, \rho) \subset R^n$. Due to (4.10), the existence of a reference solution $\hat{x}$ on $[0, k, \Delta t]$, and the assumption that U is bounded, there exists, for every $u \in U_{ad}$ and $x_0 \in R^n$ with $|x_0| \leq r$, a solution $x(\cdot; u, x_o)$ to (2.1), and the set $\{x(t; u, x_0)\colon t \in [0, k\Delta t], u \in U_{ad}, |x_0| \leq r\}$ is contained in a ball $B(0, \bar{\rho})$ with $\bar{\rho} \geq \rho$. Let $\|f^0\|$ and $\|G\|$ denote the Lipschitz constants of f^0 and G on $B(0, \bar{\rho}) \times U$ and $B(0, \bar{\rho})$, respectively.

Now fix $x_0 \in R$ with $|x_0| \le r$, and determine the discrete receding horizon sequence $\{(\bar{x}_{j-1}, \bar{u}_j)\}_{j=1}^k$ according to (3.4) with $N = 1$. Due to (4.9), we have $G(\bar{x}_j) \le G(x_0)$ for all $j = 1, \dots, k$ (in fact all $j \ge 1$), and hence $\{(\bar{x}_{j-1}, \bar{u}_j)\}_{j=1}^k \subset B(0, \bar{\rho}) \times U$. Let x denote the solution to (2.1) with $x(0) = x_0$ and $u = \bar{u}_j$ on $[t_{j-1}, t_j], j = 1, \dots, r$, and note that $x(t) \in B(0, \bar{\rho})$ for $t \in [0, k\Delta t]$. For the interpolation ξ, according to (4.11) we find

$$\frac{d}{dt}(\xi(t) - x(t)) = f(\xi(t), u(t)) - f(x(t), u(t)) + d(t),$$

with d given in (4.12). It thus follows from (4.10) that

$$|\xi(t) - x(t)| \le \int_0^t (\beta|\xi(s) - x(s)| + d(s))\, ds,$$

and by Gronwall's inequality

$$|\xi(t) - x(t)| \le e^{\beta t} \int_0^t |d(s)| ds \quad \text{for} \quad t \in [0, k\Delta t]. \tag{4.15}$$

From (4.9) we deduce that

$$G(\bar{x}_k) + \Delta t \sum_{j=0}^{k-1} \gamma^{k-j} f^0(\bar{x}_{j-1}, \bar{u}_j) \le \gamma^k G(x_0), \tag{4.16}$$

where we put $\gamma = 1 - \omega\Delta t$. This implies

$$\begin{aligned} &G(x(k\Delta t)) \le \gamma^k G(x_0) + \|G\|\, |x(k\Delta t) - \bar{x}_k| \\ &\le \gamma^k G(x_0) + \|G\|\, |x(k\Delta t) - \xi(k\Delta t)| \le \gamma^k G(x_0) + \|G\| \exp(\beta k\Delta t)\, \tilde{\epsilon}(\Delta t), \end{aligned}$$

and (4.13) follows with $C = \|G\| \exp(\beta k\Delta t)$. We utilize (4.16) once again to obtain

$$\begin{aligned} &\int_0^{t_k} f^0(x(s), u(s))\, ds \le \gamma G(x_0) + \sum_{j=0}^{k-1} \int_{t_{j-1}}^{t_j} |f^0(x(s), u(s)) - f^0(\bar{x}_{j-1}, \bar{u}_j)|\, ds \\ &\le \gamma G(x_0) + \|f^0\| \int_0^{t_k} |x(s) - \xi(x)|\, ds + \|f^0\| \sum_{j=0}^{k-1} \int_0^{\Delta t} s \frac{|\bar{x}_j - \bar{x}_{j-1}|}{\Delta t}\, ds. \end{aligned}$$

By the continuity of F and the boundedness of $\{(\bar{x}_{j-1}, \bar{u}_j)\}_{j=1}^k$, it follows that $\{\frac{1}{\Delta t}(\bar{x}_j - \bar{x}_{j-1})\}_{j=1}^k$ is bounded by a constant c; hence

$$\int_0^{t_k} f^0(x(x), u(s))\, ds \le \gamma G(x_0) + \|f^0\| \left(\tilde{\epsilon}(\Delta t) + \frac{k(\Delta t)^2 c}{2}\right),$$

and (4.14) follows.

5. Applications. In this section, we demonstrate the applicability of the concepts of this paper to certain classes of control systems. Most of the examples that we consider are motivated by nonlinear dissipative control systems with finite or infinite dimensional dynamics (e.g., [T]). In the case of infinite dimensional systems, we assume that an appropriate discretization has been carried out. A full account of infinite dimensional systems will be given independently.

We consider control systems of the general form

$$\frac{d}{dt}x(t) = A_0x(t) + F(x(t)) + B(x(t))u(t) + d, \tag{5.1}$$

where $A_0 \in R^{n\times n}$, $F : R^n \to R^n$, with $F(0) = 0$, $d \in \mathbb{R}^n$, and $B : R^n \to R^{n\times m}$ are given. Many physical processes modeled by (5.1) have a natural dissipative mechanism when $u = 0$, and thus it can be shown that a certain quadratic form can serve as a control Liapunov function.

5.1. Gradient dynamics. Consider control systems of the form

$$\frac{d}{dt}x(t) = A_0x(t) - \Psi_x(x(t)) + Bu(t), \tag{5.2}$$

where $A_0 \in R^{n\times n}$ and $B \in R^{n\times m}$. Further, Ψ is a convex C^1-functional on R^n with $\Psi_x(0) = 0$ and the property that there exists a positive definite symmetric matrix $P \in R^{n\times n}$ such that

$$\Psi_x(x) \cdot (Px) \geq 0 \text{ for all } x \in R^n. \tag{5.3}$$

We assume that there exist positive constants β and c_1 such that with $u = -\beta\, B^TPx$

$$(A_0x - \Psi_x(x) + Bu) \cdot (\,Px) \leq -c_1\, x^TPx \text{ for all } x \in R^n. \tag{5.4}$$

Let $f^0(x,u) = \frac{1}{2}|x|^2 + \frac{\gamma}{2}|u|^2$, and let $\gamma > 0$ be given. Then by Theorem 2.2 the quadratic functional $G(x) = \frac{\alpha}{2}x^TPx$ is a control Liapunov function for all α sufficiently large.

If $P = I$, then the uncontrolled dynamics ($u = 0$) consists of linear as well as gradient dynamics. Due to the convexity of Ψ and $\Psi_x(0) = 0$, we have $\Psi_x(x) \cdot x \geq 0$, and hence (5.3) holds. In this case, condition (5.4) is equivalent to assuming that

$$W(x) = -(A_0x, x) + (\Psi_x(x), x) + \beta\,|B^Tx|^2$$

is positive for $x \neq 0$.

Example. As a specific case for (5.2), consider the control system (motivated by the Schrödinger equation):

$$\frac{d}{dt}z(t) = i\,Az(t) - |z(t)|^{p-2}z(t) + B_0v(t),$$

where $z(t) \in C^n$, $v \in C^m$, and A is a symmetric nonnegative definite matrix on R^n. For $z = x_1 + i\,x_2$ with $x = (x_1, x_2) \in R^{2n}$ and $v = u_1 + i\,u_2$ with $u = (u_1, u_2) \in R^{2m}$, the above equation can be written as (5.2) with

$$A_0 = \begin{pmatrix} 0 & A \\ -A & 0 \end{pmatrix}, \quad B = \begin{pmatrix} B_0 & 0 \\ 0 & B_0 \end{pmatrix}, \text{ and } \Psi(x) = \frac{1}{p}\,|x|^p.$$

Example (damped vibration dynamics). System (5.2) is also motivated by the damped wave equation. We consider the second order system

$$\frac{d^2}{dt^2}z(t) + \psi_z\left(\frac{d}{dt}z(t)\right) + \Lambda z(t) = B_0u(t),$$

where $z(t) \in R^n, B_0 \in R^{n\times m}$, and $\Lambda \in R^{n\times n}$ is positive definite. Further, ψ is a convex functional on R^n with $\psi_x(0) = 0$ and $|\psi_x(z)| \le c_1|z|$ for a constant c_1 independent of $z \in R^n$. In the case of Coulomb-like friction $\psi(z) = \sum_{i=1}^n \gamma_i \,|\sqrt{|z_i|^2 + \epsilon}|$ with constants $\gamma_i \ge 0$ and $\epsilon > 0$. If we define $x_1(t) = z(t)$ and $x_2(t) = \frac{d}{dt}z(t)$, then $x = col(x_1, x_2)$ satisfies (5.2) with

$$A_0 = \begin{pmatrix} 0 & I \\ -\Lambda & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ B_0 \end{pmatrix}, \quad \text{and} \quad \Psi(x) = \begin{pmatrix} 0 \\ \psi(x_2) \end{pmatrix}.$$

Setting

$$P = \begin{pmatrix} \Lambda & 0 \\ 0 & I \end{pmatrix},$$

we find

$$(A_0x - \Psi_x(x)) \cdot (Px) = -(\psi_x(x_2), x_2) \le 0.$$

Also note that

$$(A_0x - \Psi_x(x)) \cdot Q_x = |x_2|^2 - x_1^T\Lambda x_1 - x_1^T\psi_x(x_2),$$

where $Q(x) = x_1^Tx_2$. Let us define

$$G(x) = \frac{\alpha}{2}x^TPx + x_1^Tx_2.$$

Then for $u = -B_0^Tx_2$ we find

$$(A_0x - \Psi_x(x) + Bu) \cdot G_x = -x_1^T\Lambda x_1 - \alpha\,(|B_0^Tx_2|^2 + (\psi_x(x_2), x_2)) + |x_2|^2 - x_1^T\psi_x(x_2).$$

Thus if $|B_0^Tx_2|^2 + (\psi_z(x_2), x_2) \ge c_2\,|x_2|^2$ for some $c_2 > 0$ and α is sufficiently large, then $G(x)$ is a control Liapunov function for $f^0(x,u) = \frac{c_3}{2}\,(x^TPx + |u|^2)$ for $0 \le c_3 \le 2$.

We return to the general discussion at the beginning of this subsection and replace condition (5.4) by the following assumptions:

$$A_0x \cdot Px \le 0 \ \text{ for all } \ x \in \mathbb{R}^n, \ \text{ and } \ (A_0 - BB^TP) \ \text{ is exponentially stable.}$$

Let Π denote the solution to the Liapunov equation

$$(A_0 - BB^TP)^T\Pi + \Pi(A_0 - BB^TP) + P = 0.$$

We choose the control Liapunov function $G(x) = \frac{1}{2}x^T\Pi x + \alpha x^TPx$, with $\alpha \ge 0$ and the feedback control law $u = -B^TPx$. Then we have with $G_x(x) = \Pi x + \alpha Px$

$$\begin{aligned}
&(A_0x - \Psi_x(x) - BB^TPx) \cdot (\Pi x + \alpha Px) = (A_0x - BB^TPx) \cdot \Pi x \\
&-\Psi_x(x) \cdot \Pi x + \alpha A_0x \cdot Px - \alpha\Psi_x(x) \cdot Px - \alpha|P^TPx|^2 \\
&\le -\frac{1}{2}\,x^TPx - \Psi_x(x) \cdot \Pi x - \alpha|B^TPx|^2 - \alpha\Psi_x(x) \cdot Px,
\end{aligned}$$

where we recall that $\Psi_x(x) \cdot Px \ge 0$. Assume that there exists $\bar\alpha$ such that

$$\Psi_x(x) \cdot \Pi x \le \bar\alpha\,\Psi_x(x) \cdot Px + \frac{1}{4}\,x^TPx + \frac{\bar\alpha}{2}\,|B^TPx|^2.$$

With this choice of $\bar\alpha$, the functional $G(x)$ is a control Liapunov function for the control system (5.2) with cost given by $f^\circ(x,u) = \frac{1}{4}\,x^TPx + \frac{\bar\alpha}{2}\,|B^TPx|^2$.

5.2. Dissipative systems. The second class of dissipative nonlinear systems that we consider here is motivated by the incompressible Navier–Stokes equations (e.g., with homogeneous Dirichlet boundary conditions),

$$\frac{d}{dt}x(t) = A_0x(t) + F(x(t)) + Bu(t) + d, \tag{5.5}$$

where $F : R^n \to R^n$ is a locally Lipschitz function satisfying $(F(x), x)_{R^n} = 0$, A_0 is a nonpositive definite symmetric matrix on R^n, $B \in R^{n\times m}$, and $d \in R^n$. Then for $u \in L^2_{loc}(0,\infty, R^m)$ there exists a globally defined unique solution to (5.5). In fact, suppose x is a solution defined on $[0,\tau)$. Then

$$\frac{1}{2}\frac{d}{dt}|x(t)|^2 = (A_0x(t) + F(x(t)) + Bu(t) + d)\cdot x(t) \le \frac{1}{2}|x(t)|^2 + |B|^2|u(t)|^2 + |d|^2$$

for $t \in [0,\tau)$. Thus we have the a priori bound

$$|x(t)|^2 \le e^t\,|x_0|^2 + 2\int_0^t e^{t-s}(|B|^2|u(t)|^2 + |d|^2)\,ds$$

for $t \in [0,\tau)$. By the continuation method there exists a unique global solution to (5.5). Let us assume that

$$\ker(A_0) \subset \text{range}(B).$$

Further, let $G(x) = \frac{\alpha}{2}|x|^2$, $\alpha > 0$, and set $u = -B^Tx$. Then there exists $\omega_1 > 0$ such that

$$(A_0x + F(x) + Bu)\cdot(\alpha x) = x^TA_0x - |B^Tx|^2 \le -\alpha\,\omega_1\,|x|^2$$

for all $x \in R^n$. Consequently, G is a control Liapunov function for $f^0(x,u) = \frac{1}{2}|x|^2 + \frac{\gamma}{2}|u|^2$, $\gamma > 0$, and $\alpha = 0$, whenever α is sufficiently large. Moreover, if $(F(x), A_0x) \ge 0$ (for example, in the case of the incompressible Navier–Stokes equation with periodic boundary condition $(F(x), A_0x) = 0$ [T]), then $G(x) = -\frac{\alpha}{2}x^TA_0x$ is a control Liapunov function for all α sufficiently large and f^0 as above. In fact, for $u = B^TA_0x$, we have

$$(A_0x + F(x) + Bu)\cdot(-A_0x) = -|A_0x|^2 - |B^TA_0x|^2 \le -\omega_2\,|x|^2$$

for some $\omega_2 > 0$, independently of $x \in R^n$, and the claim follows.

Example (Lorenz system). The following system, proposed by Lorenz as an evidence of the limits of predictability in weather prediction, is also of the form (5.5). It is a three-mode Galerkin approximation of the Boussinesq equations for fluid convection in a two-dimensional layer heated from below and is given by

$$\frac{d}{dt}\begin{pmatrix} x_1\\ x_2\\ x_3\end{pmatrix} = \begin{pmatrix} -\sigma & 0 & 0\\ 0 & -1 & 0\\ 0 & 0 & -b\end{pmatrix}x(t) + \begin{pmatrix} \sigma x_2\\ -\sigma x_1 - x_1x_3\\ x_1x_2\end{pmatrix} - \begin{pmatrix} 0\\ 0\\ b(r+\sigma)\end{pmatrix} + \begin{pmatrix} 0\\ 0\\ 1\end{pmatrix}u(t).$$

5.3. Localized control Liapunov function. We discuss the local control Liapunov function in the sense of Remarks 2.3 and 2.5 for systems of the form

$$\frac{d}{dt}x(t) = Ax(t) + F(x(t)) + Bu(t) + d, \tag{5.6}$$

where $F(0) = 0$. Suppose that (A, B) is controllable. Then there exists a unique positive definite symmetric matrix solution Π to the Riccati equation $A^T\Pi + \Pi A - \Pi BB^T\Pi + I = 0$. We shall argue that $G(x) = \frac{1}{2}\, x^T\Pi x$ is a local λ-control Liapunov function for the feedback law $u = -B^T\Pi x$ provided that there exist $\alpha > 0$ and $\omega > 0$ such that

$$-\frac{1}{2}\left(|B^T\Pi x|^2 + |x|^2\right) + f^0(x, -B^T\Pi x) + (F(x), \Pi x) \le -\frac{\omega}{2}\,|x|^2 \tag{5.7}$$

for all $x \in S_\alpha = \{x : G(x) \le \alpha\}$. In fact,

$$f(x, -B^T\Pi)\cdot(\Pi x) + f^0(x, -B^T\Pi x) \le -\frac{\omega}{2}\,|x|^2 + (d, x) \le \frac{|d|^2}{2\omega},$$

and hence $G(x) = \frac{1}{2}\, x^T\Pi x$ is a local λ-control Liapunov function with $\lambda = \frac{|d|^2}{2\omega}$. For example, consider system (5.6) with

$$A = \begin{pmatrix} 0 & 1 \\ \\ 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ \\ 1 \end{pmatrix}, \quad F(x) = \begin{pmatrix} 0 \\ \\ -U_{x_1}(x_1) \end{pmatrix},$$

where U is a potential function on R and satisfies $U_{x_1}(0) = 0$. In the case of the inverted pendulum, $U_{x_1}(x_1) = x_1 - \sin(x_1)$. Then $\Pi = \begin{pmatrix} 1+\gamma & \gamma \\ \gamma & \gamma \end{pmatrix}$ with $\gamma = 1 + \sqrt{2}$ solves the Riccati equation. Condition (5.7) is equivalent to

$$-\frac{1}{2}\left(\gamma^2\,|x_1 + x_2|^2 + |x|^2\right) - \gamma\,(U_{x_1}(x_1), x_1 + x_2) + f^0(x, -\gamma\,(x_1 + x_2)) \le -\frac{\omega}{2}\,|x|^2.$$

In the case of the inverted pendulum, it can be proved that this holds for all $x \in R^2$, where $f^\circ(x, u) = \frac{c_1}{2}\,|x|^2 + \frac{c_2}{2}\,|u|^2$ for appropriate choice of $c_1 > 0$, $c_2 > 0$.

REFERENCES

[ABQRW] F. Allgöwer, T. Badgwell, J. Qin, J. Rawlings, and S. Wright, *Nonlinear predictive control and moving horizon estimation—an introductory overview*, in Advances in Control, P. Frank, ed., Springer-Verlag, New York, 1999, pp. 391–449.

[CA] H. Chen and F. Allgöwer, *A quasi–infinite horizon nonlinear model predictive control scheme with guaranteed stability*, Automatica J. IFAC, 34 (1998), pp. 1205–1217.

[CHK] H. Choi, M. Hinze, and K. Kunisch, *Instantaneous control of backward facing step flow*, Appl. Numer. Math., 31 (1999), pp. 133–158.

[CTMC] H. Choi, R. Temam, P. Moin, and J. Kim, *Feedback control for unsteady flow and its application to the stochastic Burgers equation*, J. Fluid Mech., 253 (1993), pp. 509–543.

[FK] R. A. Freeman and P. V. Kokotovič, *Robust Nonlinear Control Design, State–Space and Lyapunov Techniques*, Birkhäuser Boston, Boston, 1996.

[FR] W. H. Fleming and R. W. Rishel, *Deterministic and Stochastic Control*, Springer-Verlag, New York, 1975.

[FS] W. H. Fleming and M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[GPM] C. E. Garcia, D. M. Prett, and M. Morari, *Model predictive control: Theory and practice—a survey*, Automatica J. IFAC, 25 (1989), pp. 335–348.

[HV] M. Hinze and S. Volkwein, *Analysis of instantaneous control for the Burgers equation*, Nonlinear Anal., to appear.

[IK] K. Ito and K. Kunisch, *Optimal Control*, Encyclopedia of Electrical and Electronics Engineering 15, John Wiley, New York, 1999, pp. 364–379.

[JYH] A. Jadababaie, J. Yu, and J. Hauser, *Unconstrained Receding Horizon Control of Nonlinear Systems*, preprint.

[KG] S. S. Keerthi and E. G. Gilbert, *Optimal infinite-horizon feedback laws for a general class of constrained discrete-time systems: Stability and moving-horizon approximations*, J. Optim. Theory Appl., 57 (1988), pp. 265–293.

[K] D. L. Kleinman, *An easy way to stabilize a linear constant system*, IEEE Trans. Automat. Control, 15 (1970), pp. 692–712.

[Lu] D. L. Lukes, *Optimal regulation of nonlinear dynamical systems*, SIAM J. Control, 7 (1969), pp. 75–100.

[MM] D. Q. Mayne and H. Michalska, *Receding horizon control of nonlinear systems*, IEEE Trans. Automat. Control, 35 (1990), pp. 814–824.

[NP] V. Nevistič and J. A. Primbs, *A framework for robustness analysis of constrained finite receding horizon control*, IEEE Trans. Automat. Control, 45 (2000), pp. 1828–1838.

[PND] J. A. Primbs, V. Nevistič, and J. C. Doyle, *A receding horizon generalization of pointwise min–norm controllers*, IEEE Trans. Automat. Control, 45 (2000), pp. 898–909.

[S] E. D. Sontag, *Mathematical Control Theory, Deterministic Finite Dimensional Systems*, Springer-Verlag, New York, 1990.

[SMR] P. Scokaert, D. Q. Mayne, and J. B. Rawlings, *Suboptimal predictive control (feasibility implies stability)*, IEEE Trans. Automat. Control, 44 (1999), pp. 648–654.

[T] R. Temam, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, Springer-Verlag, New York, 1988.

[TU] F. Tröltzsch and A. Unger, *Fast solution of optimal control problems in the selective cooling of steel*, ZAMM Z. Angew. Math. Mech., 81 (2001), pp. 447–456.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 5, pp. 1611–1627

OVERLAPPING QUADRATIC OPTIMAL CONTROL OF LINEAR TIME-VARYING COMMUTATIVE SYSTEMS*

LUBOMÍR BAKULE†, JOSÉ RODELLAR‡, AND JOSEP M. ROSSELL§

Abstract. Overlapping quadratic optimal control of linear time-invariant continuous-time systems by using *generalized selection of complementary matrices* has been recently developed as a powerful and effective means of decentralized control design of linear time-invariant systems. In this paper, it is shown that similar generalizations exist for linear time-varying systems. The results presented here concern implicit conditions for a general form of the transition matrices and explicit conditions for a commutative class of linear time-varying systems. Several important large classes of complementary matrices are selected to offer computationally attractive results. The effectiveness of this generalized selection scheme is illustrated by a numerical example of overlapping decentralized control design.

Key words. linear time-varying continuous-time systems, large scale systems, commutative systems, inclusion principle, overlapping decomposition, optimal control, decentralization, suboptimality

AMS subject classifications. 93A14, 93A15, 93B17, 93C05, 93C50, 93C35

PII. S0363012900367643

1. Introduction.

In a large variety of physical, natural, and man-made systems, subsystems share common parts. It is useful to recognize this reality, which is usually determined by either system structure or computational requirements, in proposing decentralized control schemes that use overlapping information sets. In certain control problems appearing in areas such as traffic systems, large space structures, power systems, or data communication networks, this approach is the only effective way to proceed. Decentralized control strategies offer satisfactory performance at minimum communication cost. The designer of overlapping decentralized control first expands the system into a larger space where the subsystems are disjoint, then designs decentralized controllers in the expanded space by using standard weak coupling disjoint control design methods, and finally contracts the system and local control laws into the original space to implement such controllers.

This paper addresses the problem of overlapping decentralized control design via state linear quadratic (LQ) optimal control for a commutative class of continuous-time linear time-varying (LTV) systems.

1.1. Relevant references.

The mathematical framework for expansion-contraction relations and conditions became known as the inclusion principle [11], [12], [13], [22]. This principle defines a framework for two dynamic systems with different dimensions, in which solutions of the system with larger dimension include solutions of

*Received by the editors February 11, 2000; accepted for publication (in revised form) July 3, 2001; published electronically February 6, 2002. This work was supported in part by the MEC under grant DGESIC-SAB1999-0114, by the ASCR under grant A2075802, by the CICYT under grant TAP99-1079-C03-02, and by the UPC under grant PR99-08.

http://www.siam.org/journals/sicon/40-5/36764.html

†Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, 182 08 Prague 8, Czech Republic (bakule@utia.cas.cz).

‡Department of Applied Mathematics III, Technical University of Catalonia (UPC), Campus Nord, C-2, 08034 Barcelona, Spain (jose.rodellar@upc.es).

§Department of Applied Mathematics III, Technical University of Catalonia (UPC), 08240 Manresa, Spain (josep.maria.rossell@upc.es).

the system with smaller dimension. The relation between both systems is usually constructed on the basis of appropriate linear transformations between the corresponding systems in the original and expanded spaces, where a key role in the selection of appropriate structure of all matrices in the expanded space is played by the so-called *complementary matrices* [11], [22]. The conditions given in these works on the complementary matrices have a fundamental, implicit nature, in the sense that it is not easy to select specific values for these matrices. In fact, only two particular forms of aggregations and restrictions have been commonly adopted in the literature for numerical computations [2], [15], [22], [24]. A new characterization of the complementary matrices for linear time-invariant (LTI) systems has been recently presented in [3], [19], which gives a more explicit method for their selection and includes aggregations and restrictions as particular cases. It relies on a new constructive way of approaching the concept of canonical form within the inclusion principle previously proposed in [13], [22]. This structural characterization has been used to develop the strategy of *generalized selection of complementary matrices* in [4] to find both their structure and the optimal values of their free elements with respect to suboptimality when considering the problem of overlapping decentralized state LQ control design for LTI systems. The importance of the inclusion principle is underlined by promising applications in such diverse areas as applied mathematics [5], [6], [21], [26], automated highway systems [10], [23], flexible structures [2], data communication networks [9], and electric power systems [12], [22], [24].

One of the open research issues surrounding the inclusion principle is the extension of the results available for LTI systems to LTV systems. To the authors' knowledge, the only available results in this direction are in [14], where overlapping decentralized state LQ control of LTV systems is considered. However, the results in [14] are restricted to the use of standard forms of complementary matrices, i.e., aggregations and restrictions.

The present paper differs from all the cited references in the application of the strategy of *generalized selection of complementary matrices* to LTV systems. It extends both the results in [14] as well as those in [3], [4], [19].

1.2. Outline of the paper. When abstracting the problem of quadratic optimal control, the influence of complementary matrices on suboptimality is an important issue. The strategy of *generalized selection of complementary matrices* has been developed as an effective tool to find both structure and optimal values of free elements of complementary matrices for LTI systems. We devote the main part of this paper to an extension of this strategy for overlapping state LQ optimal control from LTI systems to a class of LTV systems possessing the commutativity property, including contractibility conditions.

The paper is organized as follows. The problem is formulated in section 2. The main results are presented in section 3, identifying a new block structure of the complementary matrices that generalizes well-known results for expansion-contraction of pairs of systems and optimal control criteria. Subsection 3.1 presents the general conditions on the complementary matrices in the LQ control of LTV systems. Because of their implicit dependence on the transition matrix, they cannot be used to derive explicit conditions. The explicit conditions are derived for LTV systems possessing the commutativity property in subsection 3.2. Subsection 3.3 presents the derivation of structural properties of the complementary matrices. From this structure, subsection 3.4 outlines a selection procedure for these matrices. In section 4, this procedure is used in an overlapping state LQ optimal control problem, which is illustrated by a

numerical example.

2. Problem formulation.

2.1. Preliminaries. Consider the optimal control problems

(2.1)
$$\min_{u(t)} \quad J(x_0,u) = x^T(t_f)\Pi\, x(t_f) + \int_{t_0}^{t_f} \left[x^T(t)Q^*(t)x(t) + u^T(t)R^*(t)u(t)\right]\, dt,$$
$$\text{subject to (s.t.) } \mathbf{S}: \quad \dot{x}(t) = A(t)x(t) + B(t)u(t),$$

(2.2)
$$\min_{\tilde{u}(t)} \quad \tilde{J}(\tilde{x}_0,\tilde{u}) = \tilde{x}^T(t_f)\tilde{\Pi}\,\tilde{x}(t_f) + \int_{t_0}^{t_f} \left[\tilde{x}^T(t)\tilde{Q}^*(t)\tilde{x}(t) + \tilde{u}^T(t)\tilde{R}^*(t)\tilde{u}(t)\right]\, dt,$$
$$\text{s.t. } \tilde{\mathbf{S}}: \quad \dot{\tilde{x}}(t) = \tilde{A}(t)\tilde{x}(t) + \tilde{B}(t)\tilde{u}(t),$$

where $x(t) \in \mathbb{R}^{\mathbf{n}}$, $u(t) \in \mathbb{R}^{\mathbf{m}}$ are the state and input of $\mathbf{S}$ at time t for $t \in [t_0, t_f]$; t_0 and t_f are the initial and the terminal time, respectively; $\tilde{x}(t) \in \mathbb{R}^{\tilde{\mathbf{n}}}$ and $\tilde{u}(t) \in \mathbb{R}^{\tilde{\mathbf{m}}}$ are the state and input of $\tilde{\mathbf{S}}$. The matrices $A(t)$, $B(t)$ and $\tilde{A}(t)$, $\tilde{B}(t)$ are continuous in t of dimensions $n \times n$, $n \times m$ and $\tilde{n} \times \tilde{n}$, $\tilde{n} \times \tilde{m}$, respectively. $Q^*(t)$, $\tilde{Q}^*(t)$ are symmetric, nonnegative definite matrices, continuous in t, of dimensions $n \times n$, $\tilde{n} \times \tilde{n}$, respectively. $R^*(t)$, $\tilde{R}^*(t)$ are symmetric, positive definite matrices, continuous in t, of dimensions $m \times m$, $\tilde{m} \times \tilde{m}$, respectively. Π, $\tilde{\Pi}$ are constant, symmetric, nonnegative definite matrices of dimensions $n \times n$, $\tilde{n} \times \tilde{n}$, respectively. In problems (2.1) and (2.2), the final time t_f is fixed, and $x(t_f)$ is free. The minimization of $J(x_0,u)$ searches for a control $u(t)$ able without an excessive effort to maintain the state vector $x(t)$ close to the zero required state at any time $t \in [t_0, t_f]$, with particular emphasis at the terminal time t_f as weighted by matrix Π. It is well known that the solution of (2.1) exists, is unique, and is given in the form $u(t) = -K(t)x(t) = -(R^*)^{-1}(t)B^T(t)P(t)x(t)$, where $P(t)$ is the nonnegative definite symmetric solution of the corresponding Riccati equation [1]. If t_f is finite, this control law ensures a bounded state, and the stability issues are absent. If t_f is infinite (with $\Pi = 0$), the question of stability becomes important. It has been shown that this control guarantees that the closed-loop system is exponentially stable under certain conditions related to controllability and observability [14], [18]. We assume that the system $\mathbf{S}$ satisfies such conditions. Similar comments hold for (2.2). Suppose that the dimensions of the state and input vectors $x(t)$, $u(t)$ of $\mathbf{S}$ are smaller than (or at most equal to) those of $\tilde{x}(t)$, $\tilde{u}(t)$ of $\tilde{\mathbf{S}}$. Denote by $x(t; x_0, u)$ the solution of $\mathbf{S}$ for a fixed input $u(t)$ and an initial state $x(0) = x_0$. Analogously, $\tilde{x}(t; \tilde{x}_0, \tilde{u})$ is used for the system $\tilde{\mathbf{S}}$. In order to simplify the notation, define $x(t; x_0, u) = x(t)$ and $\tilde{x}(t; \tilde{x}_0, \tilde{u}) = \tilde{x}(t)$. It is well known that

(2.3)
$$x(t) = \Phi(t, t_0)\, x_0 + \int_{t_0}^{t} \Phi(t,\tau)B(\tau)u(\tau)\, d\tau,$$

(2.4)
$$\tilde{x}(t) = \tilde{\Phi}(t, t_0)\, \tilde{x}_0 + \int_{t_0}^{t} \tilde{\Phi}(t,\tau)\tilde{B}(\tau)\tilde{u}(\tau)\, d\tau$$

are the unique, continuously differentiable solutions of the systems in (2.1) and (2.2), respectively. The *transition matrices* $\Phi(t, t_0)$, $\tilde{\Phi}(t, t_0)$ are given by the Peano–Baker series [20].

The systems $\mathbf{S}$ and $\tilde{\mathbf{S}}$ are related by the following transformations:

(2.5)
$$\tilde{x}(t) = Vx(t), \quad x(t) = U\tilde{x}(t), \quad \tilde{u}(t) = Ru(t), \quad u(t) = Q\tilde{u}(t),$$

where V, U, R, and Q are constant matrices of appropriate dimensions and full ranks.

DEFINITION 2.1. *Consider* $\mathbf{S}$ *and* $\tilde{\mathbf{S}}$ *given in* (2.1) *and* (2.2)*, respectively. We say that a system* $\tilde{\mathbf{S}}$ *includes the system* $\mathbf{S}$*—that is,* $\tilde{\mathbf{S}} \supset \mathbf{S}$*—if there exists a quadruplet of constant matrices* (U, V, Q, R) *such that* $UV = I_n$*,* $QR = I_m$*, and for any initial state* x_0 *and any fixed input* $u(t)$ *of* $\mathbf{S}$*,* $x(t; x_0, u) = U\tilde{x}(t; Vx_0, Ru)$ *for all* $t \in [t_0, t_f]$*.*

DEFINITION 2.2. *A pair* $(\tilde{\mathbf{S}}, \tilde{J})$ *includes a pair* $(\mathbf{S}, J)$ *if* $\tilde{\mathbf{S}} \supset \mathbf{S}$ *and* $J(x_0, u) = \tilde{J}(Vx_0, Ru)$*.*

DEFINITION 2.3. *If* $(\tilde{\mathbf{S}}, \tilde{J}) \supset (\mathbf{S}, J)$*, then* $(\tilde{\mathbf{S}}, \tilde{J})$ *is said to be an* expansion *of* $(\mathbf{S}, J)$*, and* $(\mathbf{S}, J)$ *is called a* contraction *of* $(\tilde{\mathbf{S}}, \tilde{J})$*.*

In other words, this inclusion means that the problem $(\mathbf{S}, J)$ can be solved by solving the problem $(\tilde{\mathbf{S}}, \tilde{J})$.

DEFINITION 2.4. *Consider* $\mathbf{S}$ *and* $\tilde{\mathbf{S}}$ *given in* (2.1) *and* (2.2)*, respectively, such that* $\tilde{\mathbf{S}} \supset \mathbf{S}$*. Then a control law* $\tilde{u}(t) = -\tilde{K}(t)\,\tilde{x}(t)$ *for* $\tilde{\mathbf{S}}$ *is contractible to the control law* $u(t) = -K(t)\,x(t)$ *for* $\mathbf{S}$ *if the choice* $\tilde{x}_0 = Vx_0$ *and* $\tilde{u}(t) = Ru(t)$ *implies* $K(t)x(t; x_0, u) = Q\tilde{K}(t)\tilde{x}(t; Vx_0, Ru)$ *for all* $t \in [t_0, t_f]$*, for any initial state* x_0 *and any fixed input* $u(t)$ *of* $\mathbf{S}$*.*

We may also say that the gain matrix $\tilde{K}(t)$ is contractible to the gain matrix $K(t)$. Contractibility implies that the expanded closed-loop system $\dot{\tilde{x}}(t) = [\tilde{A}(t) - \tilde{B}(t)\tilde{K}(t)]\tilde{x}(t)$ includes the closed-loop system $\dot{x}(t) = [A(t) - B(t)K(t)]x(t)$.

In order to obtain conditions for expansions and contractions between problems (2.1) and (2.2) as well as conditions for contractibility of control laws, the following matrix relations are introduced:

$$
\begin{aligned}
\tilde{A}(t) &= VA(t)U + M(t), & \tilde{B}(t) &= VB(t)Q + N(t),\\
\tilde{\Pi} &= U^T \Pi U + M_\Pi, & \tilde{Q}^*(t) &= U^T Q^*(t) U + M_{Q^*}(t),\\
\tilde{R}^*(t) &= Q^T R^*(t) Q + N_{R^*}(t), & \tilde{K}(t) &= RK(t)U + F(t),
\end{aligned}
\tag{2.6}
$$

where $M(t)$, $N(t)$, M_Π, $M_{Q^*}(t)$, $N_{R^*}(t)$, and $F(t)$ are called *complementary matrices*.

Usually, the transformations (U, V) and (Q, R) are selected a priori to define structural relations between the state and control variables in both systems $\mathbf{S}$ and $\tilde{\mathbf{S}}$. Given these transformations, the choice of the complementary matrices gives degrees of freedom to complete the definition of the expansion/contraction framework involving problems $(\mathbf{S}, J)$ and $(\tilde{\mathbf{S}}, \tilde{J})$ to meet some design requirements.

2.2. The problem. The motivation of this work is to systematically extend the strategy of *generalized selection of complementary matrices* for overlapping state linear quadratic optimal control from LTI systems to LTV systems. The specific objectives are the following:

- To give implicit conditions on the complementary matrices for the inclusion $(\tilde{\mathbf{S}}, \tilde{J}) \supset (\mathbf{S}, J)$ and the contractibility of the gain matrix for general LTV systems.
- To give explicit conditions on the complementary matrices for the inclusion $(\tilde{\mathbf{S}}, \tilde{J}) \supset (\mathbf{S}, J)$ and the contractibility of the gain matrix for LTV commutative systems and to derive a systematic procedure for their selection, involving overlapping decentralized state LQ optimal control design.
- To illustrate the derived results on a numerical example.

3. Main results. This section gives the results covering mainly the expansion-contraction process for continuous-time LTV systems. Subsection 3.1 includes the general results. Subsection 3.2 presents the results for LTV systems possessing the commutativity property. Subsection 3.3 characterizes the expansion-contraction process presented at a general level by using a new basis and including contractibility

conditions. Subsection 3.4 offers the results of this process for selecting particular transformation matrices.

3.1. General LTV systems. For $(\tilde{\mathbf{S}}, \tilde{J})$ to be an expansion of $(\mathbf{S}, J)$ and to ensure contractibility, we must impose some conditions on the complementary matrices given in (2.6). This is provided by the following theorems.

THEOREM 3.1. *Consider the problems given in* (2.1) *and* (2.2). $(\tilde{\mathbf{S}}, \tilde{J}) \supset (\mathbf{S}, J)$ *if and only if*

$$\begin{aligned} &U\tilde{\Phi}(t,t_0)V = \Phi(t,t_0), \qquad U\tilde{\Phi}(t,\tau)N(\tau)R = 0, \qquad V^T M_\Pi V = 0, \\ &V^T M_{Q^*}(t)V = 0, \qquad\qquad R^T N_{R^*}(t)R = 0 \end{aligned} \tag{3.1}$$

for all $t \in [t_0, t_f]$ *and all* $\tau \in [t_0, t]$.

Proof. By Definition 2.2, suppose $x(t) = U\tilde{x}(t)$. Substituting (2.3) and (2.4) into this relation and comparing both sides, we obtain (1) $U\tilde{\Phi}(t,t_0)V = \Phi(t,t_0)$ and (2) $\int_{t_0}^{t} U\tilde{\Phi}(t,\tau)\tilde{B}(\tau)\tilde{u}(\tau)\,d\tau = \int_{t_0}^{t} \Phi(t,\tau)B(\tau)u(\tau)\,d\tau$ for all $t \in [t_0, t_f]$. From (2.6), relation (2) is equivalent to $\int_{t_0}^{t} U\tilde{\Phi}(t,\tau)N(\tau)Ru(\tau)\,d\tau = 0$ for all $t \in [t_0, t_f]$, and consequently to $U\tilde{\Phi}(t,\tau)N(\tau)R = 0$ for all $t \in [t_0, t_f]$ and all $\tau \in [t_0, t]$. The remaining conditions can be obtained easily from $J(x_0, u) = \tilde{J}(Vx_0, Ru)$ and by using (2.6). ∎

THEOREM 3.2. *Consider the problems given in* (2.1) *and* (2.2). $(\tilde{\mathbf{S}}, \tilde{J}) \supset (\mathbf{S}, J)$ *if* $V^T M_\Pi V = 0$, $V^T M_{Q^*}(t)V = 0$, $R^T N_{Q^*}(t)R = 0$, *and either*

$$\begin{aligned} &\text{(a)} \quad M(t)V = 0, \qquad N(t)R = 0 \ \textit{or} \\ &\text{(b)} \quad UM(t) = 0, \quad UN(t)R = 0 \end{aligned} \tag{3.2}$$

for all $t \in [t_0, t_f]$.

Proof. We can view the transition matrix $\tilde{\Phi}(t,t_0)$ as a function of two variables defined by the Peano–Baker series

$$\begin{aligned} \tilde{\Phi}(t,t_0) = I &+ \int_{t_0}^{t} \tilde{A}(\sigma_1)\,d\sigma_1 + \int_{t_0}^{t} \tilde{A}(\sigma_1) \int_{t_0}^{\sigma_1} \tilde{A}(\sigma_2)\,d\sigma_2 d\sigma_1 \\ &+ \int_{t_0}^{t} \tilde{A}(\sigma_1) \int_{t_0}^{\sigma_1} \tilde{A}(\sigma_2) \int_{t_0}^{\sigma_2} \tilde{A}(\sigma_3)\,d\sigma_3 d\sigma_2 d\sigma_1 + \cdots, \end{aligned} \tag{3.3}$$

where $\tilde{A}(\sigma_i) = VA(\sigma_i)U + M(\sigma_i)$, $i = 1, 2, \ldots$. Multiplying both sides of $\tilde{\Phi}(t,t_0)$ by U and V, respectively, and comparing to $\Phi(t,t_0)$, the condition $M(t)V = 0$ implies $U\tilde{\Phi}(t,t_0)V = \Phi(t,t_0)$. Obviously, $N(t)R = 0$ implies $U\tilde{\Phi}(t,\tau)N(\tau)R = 0$ for all $t \in [t_0, t_f]$ and all $\tau \in [t_0, t]$. This proves part (a). Following a similar process for part (b), $UM(t) = 0$ and $UN(t)R = 0$ are sufficient conditions in order for $U\tilde{\Phi}(t,t_0)V = \Phi(t,t_0)$ and $U\tilde{\Phi}(t,\tau)N(\tau)R = 0$ to hold, respectively. ∎

The conditions (a) and (b) are two independent sets of sufficient conditions for $(\tilde{\mathbf{S}}, \tilde{J})$ to be an expansion of $(\mathbf{S}, J)$. Condition (a) corresponds to a restriction, whereas condition (b) corresponds to an aggregation [14].

THEOREM 3.3. *Consider* $\mathbf{S}$ *and* $\tilde{\mathbf{S}}$ *given in* (2.1) *and* (2.2), *respectively, such that* $\tilde{\mathbf{S}} \supset \mathbf{S}$. *A control law* $\tilde{u}(t) = -\tilde{K}(t)\,\tilde{x}(t)$ *for* $\tilde{\mathbf{S}}$ *is contractible to the control law* $u(t) = -K(t)\,x(t)$ *for* $\mathbf{S}$ *if and only if*

$$QF(t)\left[\tilde{\Phi}(t,t_0)Vx_0 + \int_{t_0}^{t} \tilde{\Phi}(t,\tau)\tilde{B}(\tau)\tilde{u}(\tau)\,d\tau\right] = 0 \tag{3.4}$$

for all $t \in [t_0, t_f]$.

Proof. By Definition 2.4, a control law $\tilde{u}(t)$ is contractible to the control law $u(t)$ when $K(t)x(t; x_0, u) = Q\tilde{K}(t)\tilde{x}(t; Vx_0, Ru)$. Then, substituting $\tilde{K}(t)$ given in (2.6) and $\tilde{x}(t)$ given in (2.4) into the above equation and comparing both sides, the proof is straightforward. □

THEOREM 3.4. *Consider* $\mathbf{S}$ *and* $\tilde{\mathbf{S}}$ *given in* (2.1) *and* (2.2)*, respectively, such that* $\tilde{\mathbf{S}} \supset \mathbf{S}$. *A control law* $\tilde{u}(t) = -\tilde{K}(t)\,\tilde{x}(t)$ *for* $\tilde{\mathbf{S}}$ *is contractible to the control law* $u(t) = -K(t)\,x(t)$ *for* $\mathbf{S}$ *if either*

$$\begin{aligned} &\text{(a)} \quad M(t)V = 0, \qquad N(t)R = 0, \quad QF(t)V = 0 \ \text{ or} \\ &\text{(b)} \quad UM(t) = 0, \quad UN(t)R = 0, \quad QF(t) = 0 \end{aligned} \tag{3.5}$$

for all $t \in [t_0, t_f]$.

Proof. Substituting the transition matrix $\tilde{\Phi}(t, t_0)$ given in (3.3) into (3.4) and using (2.6), the sufficient conditions (a) and (b) independently imply Theorem 3.3. □

Theorems 3.2 and 3.4 do not require that we know the transition matrices. However, the selections of $M(t)$, $N(t)$, $F(t)$ are constrained only to restrictions and aggregations.

The transition matrix $\tilde{\Phi}(t, t_0)$ appears in the conditions given by Theorems 3.1 and 3.3. Since it depends on the system matrix $\tilde{A}(t)$, $\tilde{\Phi}(t, t_0)$ implicitly depends on the complementary matrix $M(t)$ as given in (2.6). On the other hand, it is very difficult, if not impossible, to obtain expressions for the transition matrices except for some particular classes of systems. The computation of the solutions via the Peano–Baker series can be a complicated task excluding trivial cases. Thus the task of obtaining explicit conditions for complementary matrices satisfying Theorems 3.1 and 3.3 is practically unsolvable for general time-varying systems. Therefore, we focus our attention on a particular but sufficiently large and important class of time-varying systems characterized by possessing the commutativity property.

3.2. Commutative systems. Let us start the presentation for this class of systems.

DEFINITION 3.5. *An LTV system* $\mathbf{S}$ *such as* (2.1) *is a commutative system if and only if* $A(t)$ *satisfies* $A(t)(\int_{t_0}^{t} A(\tau)\,d\tau) = (\int_{t_0}^{t} A(\tau)\,d\tau)A(t)$ *for all* $t \in [t_0, t_f]$. *In such a case, the matrix* $\Phi(t, t_0)$ *is given by*

$$\Phi(t, t_0) = e^{\int_{t_0}^{t} A(\tau)\,d\tau} = \sum_{k=0}^{\infty} \frac{1}{k!} \left(\int_{t_0}^{t} A(\tau)\,d\tau \right)^k .$$

In particular, this is the case for any $A(t)$ given by $A(t) = \sum_{i=1}^{r} f_i(t)A_i$, where $f_i(t)$ are arbitrary real-valued functions of t and A_i are arbitrary constant $n \times n$ matrices which satisfy the commutativity conditions $A_iA_j = A_jA_i$ for all integers $1 \le i,\ j \le r$. A linear system is called *exponential* when its state-transition matrix can be written in the matrix exponential form $\Phi(t, t_0) = e^{\Gamma(t, t_0)}$, where $\Gamma(t, t_0)$ is an $n \times n$ matrix function of t and t_0. Any commutative system is exponential. In such a case, $\Gamma(t, t_0) = \int_{t_0}^{t} A(\tau)\,d\tau$. If $A(t)$ is a triangular matrix, then the solution can be reduced to a readily solvable set of scalar differential equations. When $A(t)$ is a diagonal or a constant matrix, then it meets the commutative property and the results are well known. Summarizing, the class of systems for which $A(t)$ commutes with its integral is actually fairly large [16], [20], [25].

We need to know the conditions ensuring the commutativity property of an expanded system $\tilde{\mathbf{S}}$ when assuming the commutativity of the initial system $\mathbf{S}$. This result is given by the following proposition.

PROPOSITION 3.6. *Consider* $\mathbf{S}$ *and* $\tilde{\mathbf{S}}$ *given in* (2.1) *and* (2.2)*, respectively, such that* $\tilde{\mathbf{S}} \supset \mathbf{S}$*. Suppose* $\mathbf{S}$ *a commutative system. Then* $\tilde{\mathbf{S}}$ *is a commutative system if and only if*

$$\begin{aligned} &VA(t)U\left(\int_{t_0}^{t} M(\tau)\,d\tau\right) + M(t)V\left(\int_{t_0}^{t} A(\tau)\,d\tau\right)U + M(t)\left(\int_{t_0}^{t} M(\tau)\,d\tau\right) \\ &= V\left(\int_{t_0}^{t} A(\tau)\,d\tau\right)UM(t) + \left(\int_{t_0}^{t} M(\tau)\,d\tau\right)VA(t)U + \left(\int_{t_0}^{t} M(\tau)\,d\tau\right)M(t) \end{aligned} \tag{3.6}$$

for all $t \in [t_0, t_f]$.

Proof. The relation $\tilde{A}(t)(\int_{t_0}^{t} \tilde{A}(\tau)\,d\tau) = (\int_{t_0}^{t} \tilde{A}(\tau)\,d\tau)\tilde{A}(t)$ for all $t \in [t_0, t_f]$ holds when considering $\tilde{A}(t) = VA(t)U + M(t)$ given in (2.6) together with (3.6). ∎

The remainder of this subsection specifies Theorems 3.1 and 3.3 for the class of commutative systems.

THEOREM 3.7. *Consider that* $\mathbf{S}$ *and* $\tilde{\mathbf{S}}$ *given in* (2.1) *and* (2.2)*, respectively, are commutative systems.* $(\tilde{\mathbf{S}}, \tilde{J}) \supset (\mathbf{S}, J)$ *if and only if*

$$\begin{aligned} &U\left(\int_{t_0}^{t} M(\tau)\,d\tau\right)^{i} V = 0, \quad U\left(\int_{\tau}^{t} M(\beta)\,d\beta\right)^{i-1} N(\tau)R = 0, \quad V^T M_{\Pi} V = 0, \\ &V^T M_{Q^*}(t)V = 0, \qquad\qquad R^T N_{R^*}(t)R = 0 \end{aligned} \tag{3.7}$$

for $i = 1, \ldots, \tilde{n}$*, all* $t \in [t_0, t_f]$*, and all* $\tau \in [t_0, t]$.

Proof. Since $\tilde{\mathbf{S}}$ is a commutative system, the expanded transition matrix is given by $\tilde{\Phi}(t, t_0) = \sum_{k=0}^{\infty} \frac{1}{k!}(\int_{t_0}^{t} \tilde{A}(\tau)\,d\tau)^k$. Substituting $\tilde{\Phi}(t, t_0)$ into the corresponding requirements given in Theorem 3.1, the proof is concluded. ∎

THEOREM 3.8. *Consider that* $\mathbf{S}$ *and* $\tilde{\mathbf{S}}$ *given in* (2.1) *and* (2.2)*, respectively, are commutative systems such that* $\tilde{\mathbf{S}} \supset \mathbf{S}$*. A control law* $\tilde{u}(t) = -\tilde{K}(t)\,\tilde{x}(t)$ *for* $\tilde{\mathbf{S}}$ *is contractible to the control law* $u(t) = -K(t)\,x(t)$ *for* $\mathbf{S}$ *if and only if*

$$QF(t)\left(\int_{t_0}^{t} M(\tau)\,d\tau\right)^{i-1} V = 0, \qquad QF(t)\left(\int_{\tau}^{t} M(\beta)\,d\beta\right)^{i-1} N(\tau)R = 0 \tag{3.8}$$

for $i = 1, \ldots, \tilde{n}$*, all* $t \in [t_0, t_f]$*, and all* $\tau \in [t_0, t]$.

Proof. By Theorem 3.3 with $\tilde{\Phi}(t, t_0) = \sum_{k=0}^{\infty} \frac{1}{k!}(\int_{t_0}^{t} \tilde{A}(\tau)\,d\tau)^k$ and by using relations (2.6), the theorem is proved. ∎

3.3. Expansion-contraction process.

Change of basis. Since the inclusion principle does not depend on the specific basis used in the state, input, and output spaces, we may introduce convenient changes of basis in $\tilde{\mathbf{S}}$ [3], [4], [19]. Thus the expansion-contraction process between systems $\mathbf{S}$ and $\tilde{\mathbf{S}}$ can be illustrated in the form

$$\begin{array}{ccccccccc} \mathbf{S} & \longrightarrow & \tilde{\mathbf{S}} & \longrightarrow & \bar{\tilde{\mathbf{S}}} & \longrightarrow & \tilde{\mathbf{S}} & \longrightarrow & \mathbf{S}, \\ \mathbb{R}^{\mathbf{n}} & \xrightarrow{V} & \mathbb{R}^{\tilde{\mathbf{n}}} & \xrightarrow{T_A^{-1}} & \bar{\mathbb{R}}^{\tilde{\mathbf{n}}} & \xrightarrow{T_A} & \mathbb{R}^{\tilde{\mathbf{n}}} & \xrightarrow{U} & \mathbb{R}^{\mathbf{n}}, \\ \mathbb{R}^{\mathbf{m}} & \xrightarrow{R} & \mathbb{R}^{\tilde{\mathbf{m}}} & \xrightarrow{T_B^{-1}} & \bar{\mathbb{R}}^{\tilde{\mathbf{m}}} & \xrightarrow{T_B} & \mathbb{R}^{\tilde{\mathbf{m}}} & \xrightarrow{Q} & \mathbb{R}^{\mathbf{m}}, \end{array} \tag{3.9}$$

where $\tilde{\tilde{\mathbf{S}}}$ denotes the expanded system with the new basis. The idea of using changes of basis in the expansion-contraction process was already introduced by Ikeda, Šiljak, and White [13] to represent $\tilde{\mathbf{S}}$ in a canonical form. Given V and R, we define their pseudoinverses as $U = (V^TV)^{-1}V^T$ and $Q = (R^TR)^{-1}R^T$, respectively. Let us consider the changes of basis

$$T_A = (V \;\; W_A), \qquad T_B = (R \;\; W_B), \tag{3.10}$$

where W_A, W_B are chosen such that $Im\, W_A = Ker\, U$, $Im\, W_B = Ker\, Q$. Using these transformations, it is easy to verify the conditions $\bar{U}\bar{V} = I_n$, $\bar{V}\bar{U} = \left(\begin{smallmatrix} I_n & 0 \\ 0 & 0 \end{smallmatrix}\right)$, and $\bar{Q}\bar{R} = I_m$, $\bar{R}\bar{Q} = \left(\begin{smallmatrix} I_m & 0 \\ 0 & 0 \end{smallmatrix}\right)$, where $\bar{V} = T_A^{-1}V = \left(\begin{smallmatrix} I_n \\ 0 \end{smallmatrix}\right)$, $\bar{U} = UT_A = (I_n \; 0)$ and $\bar{R} = T_B^{-1}R = \left(\begin{smallmatrix} I_m \\ 0 \end{smallmatrix}\right)$, $\bar{Q} = QT_B = (I_m \; 0)$. In fact, obtaining these conditions is the motivating factor for defining T_A and T_B in (3.10). These conditions will be crucial to obtaining explicit block structures (with zero blocks) of the complementary matrices and, further, to giving a general strategy for their selection.

Expansion-contraction in the new basis. For simplicity, we will consider the system $\mathbf{S}$ having the following structure:

(3.11)

$$\begin{pmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \dot{x}_3(t) \end{pmatrix} = \left(\begin{array}{c|c|c} A_{11}(t) & A_{12}(t) & A_{13}(t) \\ \hline & --- & --- \\ A_{21}(t) & A_{22}(t) & A_{23}(t) \\ --- & --- & \\ \hline A_{31}(t) & A_{32}(t) & A_{33}(t) \end{array}\right) \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{pmatrix} + \left(\begin{array}{c|c|c} B_{11}(t) & B_{12}(t) & B_{13}(t) \\ \hline & --- & --- \\ B_{21}(t) & B_{22}(t) & B_{23}(t) \\ --- & --- & \\ \hline B_{31}(t) & B_{32}(t) & B_{33}(t) \end{array}\right) \begin{pmatrix} u_1(t) \\ u_2(t) \\ u_3(t) \end{pmatrix},$$

where $A_{ii}(t)$, $B_{ii}(t)$, $i = 1, 2, 3$, are $n_i \times n_i$, $n_i \times m_i$ matrices, respectively. This system is composed of two subsystems with one overlapped part, but it is well known that it can be easily generalized for any number of interconnected overlapped subsystems. This structure has been extensively adopted as a prototype structure in the literature [12], [14], [22].

Consider $(\tilde{\tilde{\mathbf{S}}}, \tilde{\tilde{J}})$ defined by the problem

(3.12)

$$\min_{\tilde{\tilde{u}}(t)} \; \tilde{\tilde{J}}\big(\tilde{\tilde{x}}_0, \tilde{\tilde{u}}(t)\big) = \tilde{\tilde{x}}^T(t_f)\tilde{\tilde{\Pi}}\,\tilde{\tilde{x}}(t_f) + \int_{t_0}^{t_f} \left[\tilde{\tilde{x}}^T(t)\tilde{\tilde{Q}}^*(t)\tilde{\tilde{x}}(t) + \tilde{\tilde{u}}^T(t)\tilde{\tilde{R}}^*(t)\tilde{\tilde{u}}(t)\right] dt$$
$$\text{s.t. } \tilde{\tilde{\mathbf{S}}}: \quad \dot{\tilde{\tilde{x}}}(t) = \tilde{\tilde{A}}(t)\,\tilde{\tilde{x}}(t) + \tilde{\tilde{B}}(t)\,\tilde{\tilde{u}}(t),$$

where $\tilde{\tilde{A}}(t)$, $\tilde{\tilde{B}}(t)$, $\tilde{\tilde{\Pi}}$, $\tilde{\tilde{Q}}^*(t)$, and $\tilde{\tilde{R}}^*(t)$ denote the matrices in the system $\tilde{\tilde{\mathbf{S}}}$ of appropriate dimensions. The vectors $\tilde{\tilde{x}}(t)$ and $\tilde{\tilde{u}}(t)$ are defined as $\tilde{\tilde{x}}(t) = T_A^{-1}Vx(t) = \bar{V}x(t)$, $\tilde{\tilde{u}}(t) = T_B^{-1}Ru(t) = \bar{R}u(t)$. Now, analogously to $\tilde{\mathbf{S}}$, denote the relations for the system $\tilde{\tilde{\mathbf{S}}}$ as

$$\begin{aligned} \tilde{\tilde{A}}(t) &= \bar{V}A(t)\bar{U} + \bar{M}(t), & \tilde{\tilde{B}}(t) &= \bar{V}B(t)\bar{Q} + \bar{N}(t), \\ \tilde{\tilde{\Pi}} &= \bar{U}^T\Pi\bar{U} + \bar{M}_\Pi, & \tilde{\tilde{Q}}^*(t) &= \bar{U}^TQ^*(t)\bar{U} + \bar{M}_{Q^*}(t), \\ \tilde{\tilde{R}}^*(t) &= \bar{Q}^TR^*(t)\bar{Q} + \bar{N}_{R^*}(t), \end{aligned} \tag{3.13}$$

where the new complementary matrices are

$$\begin{aligned} \bar{M}(t) &= T_A^{-1}M(t)T_A, & \bar{N}(t) &= T_A^{-1}N(t)T_B, & \bar{M}_\Pi &= T_A^TM_\Pi T_A, \\ \bar{M}_{Q^*}(t) &= T_A^TM_{Q^*}(t)T_A, & \bar{N}_{R^*}(t) &= T_B^TN_{R^*}(t)T_B. \end{aligned} \tag{3.14}$$

Note. Since changes of basis do not affect the commutativity property, the system $\tilde{\tilde{\mathbf{S}}}$ is commutative if $\tilde{\mathbf{S}}$ is commutative.

First, we analyze the structure of the matrices $\bar{M}(t)$, $\bar{N}(t)$, $\bar{M}_{\Pi}$, $\bar{M}_{Q^*}(t)$, and $\bar{N}_{R^*}(t)$ in the expanded system. Consider the complementary matrices of $\tilde{\mathbf{S}}$ having the form $M(t) = (M_{ij}(t))$, $N(t) = (N_{ij}(t))$, $M_{\Pi} = (M_{\Pi_{ij}})$, $M_{Q^*}(t) = (M_{Q^*_{ij}}(t))$, $N_{R^*}(t) = (N_{R^*_{ij}}(t))$ for $i, j = 1, \ldots, 4$, with $M_{\Pi_{ij}} = M^T_{\Pi_{ji}}$, $M_{Q^*_{ij}}(t) = M^T_{Q^*_{ji}}(t)$, $N_{R^*_{ij}}(t) = N^T_{R^*_{ji}}(t)$, where each matrix has appropriate dimensions corresponding to the initial structure given in (3.11). It is convenient to deal with matrix blocks when using the matrices T_A, T_B, T_A^{-1}, T_B^{-1}. Suppose the matrices

$$\bar{M}(t) = \begin{pmatrix} \bar{M}_{11}(t) & \bar{M}_{12}(t) \\ \bar{M}_{21}(t) & \bar{M}_{22}(t) \end{pmatrix}, \quad \bar{N}(t) = \begin{pmatrix} \bar{N}_{11}(t) & \bar{N}_{12}(t) \\ \bar{N}_{21}(t) & \bar{N}_{22}(t) \end{pmatrix}, \quad \bar{M}_{\Pi} = \begin{pmatrix} \bar{M}_{\Pi_{11}} & \bar{M}_{\Pi_{12}} \\ \bar{M}^T_{\Pi_{12}} & \bar{M}_{\Pi_{22}} \end{pmatrix},$$

$$\bar{M}_{Q^*}(t) = \begin{pmatrix} \bar{M}_{Q^*_{11}}(t) & \bar{M}_{Q^*_{12}}(t) \\ \bar{M}^T_{Q^*_{12}}(t) & \bar{M}_{Q^*_{22}}(t) \end{pmatrix}, \quad \bar{N}_{R^*}(t) = \begin{pmatrix} \bar{N}_{R^*_{11}}(t) & \bar{N}_{R^*_{12}}(t) \\ \bar{N}^T_{R^*_{12}}(t) & \bar{N}_{R^*_{22}}(t) \end{pmatrix},$$

where $\bar{M}_{11}(t)$, $\bar{M}_{22}(t)$ are $n \times n$, $(\tilde{n} - n) \times (\tilde{n} - n)$ matrices, respectively. $\bar{N}_{11}(t)$, $\bar{N}_{22}(t)$ are $n \times m$, $(\tilde{n} - n) \times (\tilde{m} - m)$ matrices, respectively. $\bar{M}_{\Pi_{11}}$, $\bar{M}_{\Pi_{22}}$ are $n \times n$, $(\tilde{n} - n) \times (\tilde{n} - n)$ matrices, respectively. $\bar{M}_{Q^*_{11}}(t)$, $\bar{M}_{Q^*_{22}}(t)$ are $n \times n$, $(\tilde{n} - n) \times (\tilde{n} - n)$ matrices, respectively. $\bar{N}_{R^*_{11}}(t)$, $\bar{N}_{R^*_{22}}(t)$ are $m \times m$, $(\tilde{m} - m) \times (\tilde{m} - m)$ matrices, respectively. We need to know the form of the submatrices $\bar{M}_{ij}(t)$, $\bar{N}_{ij}(t)$, $\bar{M}_{\Pi_{ij}}$, $\bar{M}_{Q^*_{ij}}(t)$, and $\bar{N}_{R^*_{ij}}(t)$ for $i, j = 1, 2$. This is given in the following propositions.

PROPOSITION 3.9. *Consider that* $\mathbf{S}$ *and* $\tilde{\tilde{\mathbf{S}}}$ *given in* (2.1) *and* (3.12)*, respectively, are commutative systems such that* $\tilde{\tilde{\mathbf{S}}} \supset \mathbf{S}$*. Then* $\bar{M}(t) = \left(\begin{smallmatrix} 0 & \bar{M}_{12}(t) \\ \bar{M}_{21}(t) & \bar{M}_{22}(t) \end{smallmatrix}\right)$*, where* (0) *denotes a matrix of order* n*, and the other blocks satisfy* $\int_{t_0}^t \bar{M}_{12}(\tau)\, d\tau (\int_{t_0}^t \bar{M}_{22}(\tau)\, d\tau)^{i-2} \int_{t_0}^t \bar{M}_{21}(\tau)\, d\tau = 0$ *for* $i = 2, \ldots, \tilde{n}$ *and all* $t \in [t_0, t_f]$*.*

Proof. Imposing the first condition given in Theorem 3.7, $\bar{U}(\int_{t_0}^t \bar{M}(\tau)\, d\tau)^i \bar{V} = 0$ for $i = 1$, we get $\bar{U}(\int_{t_0}^t \bar{M}(\tau)\, d\tau)\bar{V} = 0$ and consequently $\int_{t_0}^t \bar{M}_{11}(\tau)\, d\tau = 0$ for all $t \in [t_0, t_f]$. Then $\bar{M}_{11}(t) = 0$ for all $t \in [t_0, t_f]$. For $\mu = 2, \ldots, \tilde{n}$ we obtain

$$\left(\int_{t_0}^t \bar{M}(\tau)\, d\tau\right)^{\mu} = \begin{pmatrix} \int_{t_0}^t \bar{M}_{12}(\tau)\, d\tau \Big(\int_{t_0}^t \bar{M}_{22}(\tau)\, d\tau\Big)^{\mu-2} \int_{t_0}^t \bar{M}_{21}(\tau)\, d\tau & | \\ & | \\ \Big(\int_{t_0}^t \bar{M}_{22}(\tau)\, d\tau\Big)^{\mu-1} \int_{t_0}^t \bar{M}_{21}(\tau)\, d\tau & | \end{pmatrix}$$

$$\begin{pmatrix} | & \int_{t_0}^t \bar{M}_{12}(\tau)\, d\tau \Big(\int_{t_0}^t \bar{M}_{22}(\tau)\, d\tau\Big)^{\mu-1} \\ | & \\ | & \Big(\int_{t_0}^t \bar{M}_{22}(\tau)\, d\tau\Big)^{\mu} + \sum_{j=0}^{j=\mu-2} \Big(\int_{t_0}^t \bar{M}_{22}(\tau)\, d\tau\Big)^j \int_{t_0}^t \bar{M}_{21}(\tau)\, d\tau \int_{t_0}^t \bar{M}_{12}(\tau)\, d\tau \Big(\int_{t_0}^t \bar{M}_{22}(\tau)\, d\tau\Big)^{\mu-2-j} \end{pmatrix}. \tag{3.15}$$

Then, for $i = k \geq 2$, $\bar{U}(\int_{t_0}^t \bar{M}(\tau)\, d\tau)^k \bar{V} = 0$ implies $\int_{t_0}^t \bar{M}_{12}(\tau)\, d\tau\, (\int_{t_0}^t \bar{M}_{22}(\tau)\, d\tau)^{k-2} \int_{t_0}^t \bar{M}_{21}(\tau)\, d\tau = 0$. Repeating this process for $i = \tilde{n}$, $\bar{U}(\int_{t_0}^t \bar{M}(\tau)\, d\tau)^{\tilde{n}} \bar{V} = 0$ leads to $\int_{t_0}^t \bar{M}_{12}(\tau)\, d\tau (\int_{t_0}^t \bar{M}_{22}(\tau)\, d\tau)^{\tilde{n}-2} \int_{t_0}^t \bar{M}_{21}(\tau)\, d\tau = 0$ for all $t \in [t_0, t_f]$. ∎

PROPOSITION 3.10. *Consider that* $\mathbf{S}$ *and* $\tilde{\tilde{\mathbf{S}}}$*, given in* (2.1) *and* (3.12)*, respectively, are commutative systems such that* $\tilde{\tilde{\mathbf{S}}} \supset \mathbf{S}$*. Then* $\bar{N}(t) = \left(\begin{smallmatrix} 0 & \bar{N}_{12}(t) \\ \bar{N}_{21}(t) & \bar{N}_{22}(t) \end{smallmatrix}\right)$*, where* (0) *is an* $n \times m$ *matrix and the other blocks satisfy* $\int_{t_0}^t \bar{M}_{12}(\tau)\, d\tau \, (\int_{\tau}^t \bar{M}_{22}(\beta)\, d\beta)^{i-2} \, \bar{N}_{21}(\tau) = 0$ *for* $i = 2, \ldots, \tilde{n}$*, all* $t \in [t_0, t_f]$*, and all* $\tau \in [t_0, t]$*.*

Proof. This proof is similar to Proposition 3.9 from the condition $\bar{U}\left(\int_\tau^t \bar{M}(\beta)\,d\beta\right)^{i-1}\bar{N}(\tau)\bar{R}=0$, $i=1,\dots,\tilde{n}$, given in Theorem 3.7 and using Proposition 3.9. ∎

Note. The conditions imposed by Theorem 3.7 on matrices $M(t)$ and $N(t)$ in order to verify $\tilde{\bar{\mathbf{S}}}\supset\mathbf{S}$ have been reduced to conditions on submatrices, that is, $\int_{t_0}^t \bar{M}_{12}(\tau)\,d\tau\left(\int_{t_0}^t \bar{M}_{22}(\tau)\,d\tau\right)^{i-2}\int_{t_0}^t \bar{M}_{21}(\tau)\,d\tau=0$ and $\int_{t_0}^t \bar{M}_{12}(\tau)\,d\tau\left(\int_\tau^t \bar{M}_{22}(\beta)\,d\beta\right)^{i-2}\bar{N}_{21}(\tau)=0$ for $i=2,\dots,\tilde{n}$, all $t\in[t_0,t_f]$, and all $\tau\in[t_0,t]$, where $\bar{M}(t)=\left(\begin{smallmatrix}0 & \bar{M}_{12}(t)\\ \bar{M}_{21}(t) & \bar{M}_{22}(t)\end{smallmatrix}\right)$ and $\bar{N}(t)=\left(\begin{smallmatrix}0 & \bar{N}_{12}(t)\\ \bar{N}_{21}(t) & \bar{N}_{22}(t)\end{smallmatrix}\right)$.

THEOREM 3.11. *Consider that* $\mathbf{S}$ *and* $\tilde{\bar{\mathbf{S}}}$, *given in* (2.1) *and* (3.12), *respectively, are commutative systems.* $(\tilde{\bar{\mathbf{S}}},\tilde{\bar{J}})\supset(\mathbf{S},J)$ *if* $\bar{M}_\Pi=\left(\begin{smallmatrix}0 & \bar{M}_{\Pi_{12}}\\ \bar{M}_{\Pi_{12}}^T & \bar{M}_{\Pi_{22}}\end{smallmatrix}\right)$, $\bar{M}_{Q^*}(t)=\left(\begin{smallmatrix}0 & \bar{M}_{Q_{12}^*}(t)\\ \bar{M}_{Q_{12}^*}^T(t) & \bar{M}_{Q_{22}^*}(t)\end{smallmatrix}\right)$, $\bar{N}_{R^*}(t)=\left(\begin{smallmatrix}0 & \bar{N}_{R_{12}^*}(t)\\ \bar{N}_{R_{12}^*}^T(t) & \bar{N}_{R_{22}^*}(t)\end{smallmatrix}\right)$, *and either*

$$\begin{aligned}&\text{(a)}\quad \bar{M}(t)=\begin{pmatrix}0 & \bar{M}_{12}(t)\\ 0 & \bar{M}_{22}(t)\end{pmatrix},\quad \bar{N}(t)=\begin{pmatrix}0 & \bar{N}_{12}(t)\\ 0 & \bar{N}_{22}(t)\end{pmatrix}\quad \textit{or}\\ &\text{(b)}\quad \bar{M}(t)=\begin{pmatrix}0 & 0\\ \bar{M}_{21}(t) & \bar{M}_{22}(t)\end{pmatrix},\quad \bar{N}(t)=\begin{pmatrix}0 & \bar{N}_{12}(t)\\ \bar{N}_{21}(t) & \bar{N}_{22}(t)\end{pmatrix}\end{aligned}\tag{3.16}$$

for all $t\in[t_0,t_f]$.

Proof. Considering the conditions (a) and (b) given by Theorem 3.2, respectively, in the new expanded system $\tilde{\bar{\mathbf{S}}}$, the proof is straightforward. ∎

Contractibility. The idea is to design control laws in the expanded system $\tilde{\mathbf{S}}$ so that we can contract and implement them into the original system $\mathbf{S}$. Now, we want to determine the conditions under which a control law designed in $\tilde{\bar{\mathbf{S}}}$ can be contracted into the system $\mathbf{S}$ in terms of the complementary matrices.

Suppose that the complementary matrix $F(t)$ has the form $F(t)=(F_{ij}(t))$, $i,j=1,\dots,4$, where $F_{11}(t)$, $F_{22}(t)$, $F_{33}(t)$, and $F_{44}(t)$ are $m_1\times n_1$, $m_2\times n_2$, $m_2\times n_2$, and $m_3\times n_3$ matrices, respectively. Define $\bar{F}(t)=\left(\begin{smallmatrix}\bar{F}_{11}(t) & \bar{F}_{12}(t)\\ \bar{F}_{21}(t) & \bar{F}_{22}(t)\end{smallmatrix}\right)$, where $\bar{F}_{11}(t)$ and $\bar{F}_{22}(t)$ are $m\times n$ and $(\tilde{m}-m)\times(\tilde{n}-n)$ matrices, respectively. Similarly, denote $K(t)=(K_{ij}(t))$, $i,j=1,\dots,3$, where $K_{11}(t)$, $K_{22}(t)$, $K_{33}(t)$ are $m_i\times n_i$ matrices, $i=1,\dots,3$, respectively. The gain matrix $\tilde{\bar{K}}(t)$ for the system $\tilde{\bar{\mathbf{S}}}$ has the form $\tilde{\bar{K}}(t)=\bar{R}K(t)\bar{U}+\bar{F}(t)$, where $\tilde{\bar{K}}(t)=T_{_B}^{-1}\tilde{K}(t)T_{_A}$ and $\bar{F}(t)=T_{_B}^{-1}F(t)T_{_A}$. By Definition 2.4, $\tilde{\bar{u}}(t)=-\tilde{\bar{K}}(t)\tilde{\bar{x}}(t)$ of $\tilde{\bar{\mathbf{S}}}$ is contractible to the control law $u(t)=-K(t)x(t)$ of $\mathbf{S}$ whenever $K(t)x(t;x_0,u)=\bar{Q}\tilde{\bar{K}}(t)\tilde{\bar{x}}(t;\bar{V}x_0,\bar{R}u)$ for all $t\in[t_0,t_f]$.

So far we do not know the form of the complementary matrix $F(t)$ and the conditions which must be satisfied in order to get a contractible control law. The following theorem answers this question.

THEOREM 3.12. *Consider that* $\mathbf{S}$ *and* $\tilde{\bar{\mathbf{S}}}$ *given in* (2.1) *and* (3.12), *respectively, are commutative systems such that* $\tilde{\bar{\mathbf{S}}}\supset\mathbf{S}$. *A control law* $\tilde{\bar{u}}(t)=-\tilde{\bar{K}}(t)\tilde{\bar{x}}(t)$ *in the expanded system* $\tilde{\bar{\mathbf{S}}}$ *is contractible to the control law* $u(t)=-K(t)x(t)$ *of* $\mathbf{S}$ *if and only if* $\bar{F}=\left(\begin{smallmatrix}0 & \bar{F}_{12}(t)\\ \bar{F}_{21}(t) & \bar{F}_{22}(t)\end{smallmatrix}\right)$ *and satisfies*

$$\begin{aligned}\bar{F}_{12}(t)\left(\int_{t_0}^t \bar{M}_{22}(\tau)\,d\tau\right)^{i-1}\int_{t_0}^t \bar{M}_{21}(\tau)\,d\tau=0,\\ \bar{F}_{12}(t)\left(\int_\tau^t \bar{M}_{22}(\beta)\,d\beta\right)^{j-1}\bar{N}_{21}(\tau)=0\end{aligned}\tag{3.17}$$

for $i = 1, \dots, \tilde{n}-1$, $j = 1, \dots, \tilde{n}$, *all* $t \in [t_0, t_f]$, *and all* $\tau \in [t_0, t]$.

Proof. Let $\bar{F}(t)$ be $\bar{F}(t) = \left(\begin{smallmatrix} \bar{F}_{11}(t) & \bar{F}_{12}(t) \\ \bar{F}_{21}(t) & \bar{F}_{22}(t) \end{smallmatrix}\right)$. Imposing the first condition given in Theorem 3.8, that is, $\bar{Q}\bar{F}(t)\left(\int_{t_0}^{t} \bar{M}(\tau)\,d\tau\right)^{i-1}\bar{V} = 0$ for $i = 1$, we obtain $\bar{Q}\bar{F}(t)\bar{V} = 0$ and thus that $\bar{F}_{11}(t) = 0$. For $i = k \geq 2$, $\bar{Q}\bar{F}(t)\left(\int_{t_0}^{t} \bar{M}(\tau)\,d\tau\right)^{k-1}\bar{V} = 0$ implies $\bar{F}_{12}(t)\left(\int_{t_0}^{t} \bar{M}_{22}(\tau)\,d\tau\right)^{k-2}\int_{t_0}^{t} \bar{M}_{21}(\tau)\,d\tau = 0$. For $i = \tilde{n}$, $\bar{Q}\bar{F}(t)\left(\int_{t_0}^{t} \bar{M}(\tau)\,d\tau\right)^{\tilde{n}-1}\bar{V} = 0$ implies that $\bar{F}_{12}(t)\left(\int_{t_0}^{t} \bar{M}_{22}(\tau)\,d\tau\right)^{\tilde{n}-2}\int_{t_0}^{t} \bar{M}_{21}(\tau)\,d\tau = 0$ holds for all $t \in [t_0, t_f]$. This proves the first equation in (3.17). Similarly, $\bar{Q}\bar{F}(t)\left(\int_{\tau}^{t} \bar{M}(\beta)\,d\beta\right)^{i-1}\bar{N}(\tau)\bar{R} = 0$ for $i = 1, \dots, \tilde{n}$, all $t \in [t_0, t_f]$, and all $\tau \in [t_0, t]$; this equality leads to $\bar{F}_{12}(t)\left(\int_{\tau}^{t} \bar{M}_{22}(\beta)\,d\beta\right)^{j-1}\cdot \bar{N}_{21}(\tau) = 0$ for $j = 1, \dots, \tilde{n}$, all $t \in [t_0, t_f]$, and all $\tau \in [t_0, t]$. ∎

3.4. Selection of complementary matrices. The above results do not depend on the selection of the matrices V and R, and thus they can be applied to any expansion-contraction process. To use these results in a practical scheme, we start by defining specific transformations V and R to expand a given problem (2.1). Here we consider the following expansion transformation matrices:

$$V = \begin{pmatrix} I_{n_1} & 0 & 0 \\ 0 & I_{n_2} & 0 \\ 0 & I_{n_2} & 0 \\ 0 & 0 & I_{n_3} \end{pmatrix}, \qquad R = \begin{pmatrix} I_{m_1} & 0 & 0 \\ 0 & I_{m_2} & 0 \\ 0 & I_{m_2} & 0 \\ 0 & 0 & I_{m_3} \end{pmatrix}. \tag{3.18}$$

These usual transformations are chosen to lead, in a simple natural way, to an expanded system where the state vector $x_2(t)$ and the control vector $u_2(t)$ appear repeated in $\tilde{x}(t) = (x_1^T(t), x_2^T(t), x_2^T(t), x_3^T(t))^T$ and $\tilde{u}(t) = (u_1^T(t), u_2^T(t), u_2^T(t), u_3^T(t))^T$, respectively. According to (3.10), the changes of basis to define the system $\tilde{\bar{\mathbf{S}}}$ for matrices (3.18) are given by

$$T_A = \begin{pmatrix} I_{n_1} & 0 & 0 & 0 \\ 0 & I_{n_2} & 0 & I_{n_2} \\ 0 & I_{n_2} & 0 & -I_{n_2} \\ 0 & 0 & I_{n_3} & 0 \end{pmatrix}, \qquad T_A^{-1} = \begin{pmatrix} I_{n_1} & 0 & 0 & 0 \\ 0 & \frac{1}{2}I_{n_2} & \frac{1}{2}I_{n_2} & 0 \\ 0 & 0 & 0 & I_{n_3} \\ 0 & \frac{1}{2}I_{n_2} & -\frac{1}{2}I_{n_2} & 0 \end{pmatrix}. \tag{3.19}$$

Analogously, they are given by T_B, T_B^{-1}. The following theorems express the structure of the complementary matrices $M(t)$, $N(t)$, M_Π, $M_{Q^*}(t)$, $N_{R^*}(t)$, and $F(t)$ in the initial basis.

THEOREM 3.13. *Consider that* $\mathbf{S}$ *and* $\tilde{\mathbf{S}}$, *given in* (2.1) *and* (2.2), *respectively, are commutative systems.* $\tilde{\mathbf{S}} \supset \mathbf{S}$ *if and only if*

(3.20)

$$\begin{pmatrix} \int_{t_0}^{t} M_{12}(\tau)\,d\tau \\ \int_{t_0}^{t} (M_{23}(\tau)+M_{33}(\tau))\,d\tau \\ \int_{t_0}^{t} M_{42}(\tau)\,d\tau \end{pmatrix}\left(\int_{t_0}^{t} \left(M_{22}(\tau)+M_{33}(\tau) \right) d\tau\right)^{i-2}\int_{t_0}^{t} \left(M_{21}(\tau)M_{22}(\tau)+M_{23}(\tau)M_{24}(\tau) \right) d\tau = 0,$$

$$\begin{pmatrix} \int_{t_0}^{t} M_{12}(\tau)\,d\tau \\ \int_{t_0}^{t} (M_{23}(\tau)+M_{33}(\tau))\,d\tau \\ \int_{t_0}^{t} M_{42}(\tau)\,d\tau \end{pmatrix}\left(\int_{\tau}^{t} \left(M_{22}(\beta)+M_{33}(\beta) \right) d\beta\right)^{i-2}\begin{pmatrix} N_{21}(\tau)N_{22}(\tau)+N_{23}(\tau)N_{24}(\tau) \end{pmatrix} = 0$$

for $i = 2, \dots, \tilde{n}$, *all* $t \in [t_0, t_f]$, *and all* $\tau \in [t_0, t]$, *where*

$$M(t) = \begin{pmatrix} 0 & M_{12}(t) & -M_{12}(t) & 0 \\ M_{21}(t) & M_{22}(t) & M_{23}(t) & M_{24}(t) \\ -M_{21}(t) & -(M_{22}(t)+M_{23}(t)+M_{33}(t)) & M_{33}(t) & -M_{24}(t) \\ 0 & M_{42}(t) & -M_{42}(t) & 0 \end{pmatrix}. \tag{3.21}$$

The matrix $N(t)$ has the same structure as $M(t)$.

Proof. Consider $\bar{M}(t) = T_A^{-1} M(t) T_A$, given in (3.14), where T_A and T_A^{-1} are given in (3.19). From Proposition 3.9, $\bar{M}_{11}(t) = 0$ and the matrix blocks $\bar{M}_{ij}(t)$, $i, j = 1, 2$, can be identified. Consequently, we obtain the structure of the complementary matrix $M(t)$ given in (3.21). We proceed analogously for the matrix $N(t)$ by using $\bar{N}(t) = T_A^{-1} N(t) T_B$ given in (3.14) and Proposition 3.10. Now, imposing $\int_{t_0}^{t} \bar{M}_{12}(\tau)\, d\tau$ $(\int_{t_0}^{t} \bar{M}_{22}(\tau)\, d\tau)^{i-2}$ $\int_{t_0}^{t} \bar{M}_{21}(\tau)\, d\tau = 0$ and $\int_{t_0}^{t} \bar{M}_{12}(\tau)\, d\tau$ $(\int_{\tau}^{t} \bar{M}_{22}(\beta)\, d\beta)^{i-2}$ $\bar{N}_{21}(\tau) = 0$ for $i = 2, \ldots, \tilde{n}$, given by Propositions 3.9 and 3.10, respectively, we get (3.20). ∎

PROPOSITION 3.14. *Consider that* $\mathbf{S}$ *and* $\tilde{\mathbf{S}}$*, given in* (2.1) *and* (2.2)*, respectively, are commutative systems such that* $\tilde{\mathbf{S}} \supset \mathbf{S}$*. Then the corresponding expanded matrix* $\tilde{A}(t)$ *has the form*

$$\tilde{A}(t) = \begin{pmatrix} A_{11}(t) & \frac{1}{2}A_{12}(t)+M_{12}(t) & \frac{1}{2}A_{12}(t)-M_{12}(t) & A_{13}(t) \\ A_{21}(t)+M_{21}(t) & \frac{1}{2}A_{22}(t)+M_{22}(t) & \frac{1}{2}A_{22}(t)+M_{23}(t) & A_{23}(t)+M_{24}(t) \\ A_{21}(t)-M_{21}(t) & \frac{1}{2}A_{22}(t)-(M_{22}(t)+M_{23}(t)+M_{33}(t)) & \frac{1}{2}A_{22}(t)+M_{33}(t) & A_{23}(t)-M_{24}(t) \\ A_{31}(t) & \frac{1}{2}A_{32}(t)+M_{42}(t) & \frac{1}{2}A_{32}(t)-M_{42}(t) & A_{33}(t) \end{pmatrix}.$$

A similar structure can be presented for the expanded control matrix $\tilde{B}(t)$*.*

Proof. The proof is straightforward by substituting the matrix $M(t)$ given in (3.21) into $\tilde{A}(t) = VA(t)U + M(t)$. The same holds for $\tilde{B}(t) = VB(t)Q + N(t)$. ∎

THEOREM 3.15. *Consider that* $\mathbf{S}$ *and* $\tilde{\mathbf{S}}$*, given in* (2.1) *and* (2.2)*, respectively, are commutative systems.* $(\tilde{\mathbf{S}}, \tilde{J}) \supset (\mathbf{S}, J)$ *if*

(3.22)

$$M_\Pi = \begin{pmatrix} 0 & M_{\Pi_{12}} & -M_{\Pi_{12}} & 0 \\ M_{\Pi_{12}}^T & -M_{\Pi_{23}} - M_{\Pi_{23}}^T - M_{\Pi_{33}} & M_{\Pi_{23}} & M_{\Pi_{24}} \\ -M_{\Pi_{12}}^T & M_{\Pi_{23}}^T & M_{\Pi_{33}} & -M_{\Pi_{24}} \\ 0 & M_{\Pi_{24}}^T & -M_{\Pi_{24}}^T & 0 \end{pmatrix},$$

$$M_{Q^*}(t) = \begin{pmatrix} 0 & M_{Q_{12}^*}(t) & -M_{Q_{12}^*}(t) & 0 \\ M_{Q_{12}^*}^T(t) & -M_{Q_{23}^*}(t) - M_{Q_{23}^*}^T(t) - M_{Q_{33}^*}(t) & M_{Q_{23}^*}(t) & M_{Q_{24}^*}(t) \\ -M_{Q_{12}^*}^T(t) & M_{Q_{23}^*}^T(t) & M_{Q_{33}^*}(t) & -M_{Q_{24}^*}(t) \\ 0 & M_{Q_{24}^*}^T(t) & -M_{Q_{24}^*}^T(t) & 0 \end{pmatrix},$$

$$N_{R^*}(t) = \begin{pmatrix} 0 & N_{R_{12}^*}(t) & -N_{R_{12}^*}(t) & 0 \\ N_{R_{12}^*}^T(t) & -N_{R_{23}^*}(t) - N_{R_{23}^*}^T(t) - N_{R_{33}^*}(t) & N_{R_{23}^*}(t) & N_{R_{24}^*}(t) \\ -N_{R_{12}^*}^T(t) & N_{R_{23}^*}^T(t) & N_{R_{33}^*}(t) & -N_{R_{24}^*}(t) \\ 0 & N_{R_{24}^*}^T(t) & -N_{R_{24}^*}^T(t) & 0 \end{pmatrix}$$ *and either*

(a) $$M(t) = \begin{pmatrix} 0 & M_{12}(t) & -M_{12}(t) & 0 \\ 0 & M_{22}(t) & -M_{22}(t) & 0 \\ 0 & M_{32}(t) & -M_{32}(t) & 0 \\ 0 & M_{42}(t) & -M_{42}(t) & 0 \end{pmatrix}, \quad N(t) = \begin{pmatrix} 0 & N_{12}(t) & -N_{12}(t) & 0 \\ 0 & N_{22}(t) & -N_{22}(t) & 0 \\ 0 & N_{32}(t) & -N_{32}(t) & 0 \\ 0 & N_{42}(t) & -N_{42}(t) & 0 \end{pmatrix}$$ *or*

(b) $$M(t) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ M_{21}(t) & M_{22}(t) & M_{23}(t) & M_{24}(t) \\ -M_{21}(t) & -M_{22}(t) & -M_{23}(t) & -M_{24}(t) \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$N(t) = \begin{pmatrix} 0 & N_{12}(t) & -N_{12}(t) & 0 \\ N_{21}(t) & N_{22}(t) & N_{23}(t) & N_{24}(t) \\ -N_{21}(t) & -(N_{22}(t)+N_{23}(t)+N_{33}(t)) & N_{33}(t) & -N_{24}(t) \\ 0 & N_{42}(t) & -N_{42}(t) & 0 \end{pmatrix}$$

for all $t \in [t_0, t_f]$*.*

Proof. By using the equations (3.14) together with Theorem 3.11, the proof is concluded. ∎

THEOREM 3.16. *Consider that* $\mathbf{S}$ *and* $\tilde{\mathbf{S}}$*, given in* (2.1) *and* (2.2)*, respectively, are commutative systems such that* $\tilde{\mathbf{S}} \supset \mathbf{S}$*. A control law* $\tilde{u}(t) = -\tilde{K}(t)\tilde{x}(t)$ *in the expanded system* $\tilde{\mathbf{S}}$ *is contractible to the control law* $u(t) = -K(t)x(t)$ *of the system* $\mathbf{S}$ *if and only if*

(3.23)
$$\begin{pmatrix} F_{12}(t) \\ F_{23}(t)+F_{33}(t) \\ F_{42}(t) \end{pmatrix} \left(\int_{t_0}^{t} \left(M_{22}(\tau)+M_{33}(\tau) \right) d\tau \right)^{i-1} \int_{t_0}^{t} \left(M_{21}(\tau) \;\; M_{22}(\tau)+M_{23}(\tau) \;\; M_{24}(\tau) \right) d\tau = 0,$$
$$\begin{pmatrix} F_{12}(t) \\ F_{23}(t)+F_{33}(t) \\ F_{42}(t) \end{pmatrix} \left(\int_{\tau}^{t} \left(M_{22}(\beta)+M_{33}(\beta) \right) d\beta \right)^{j-1} \begin{pmatrix} N_{21}(\tau) \;\; N_{22}(\tau)+N_{23}(\tau) \;\; N_{24}(\tau) \end{pmatrix} = 0$$

for $i = 1, \ldots, \tilde{n}-1$*,* $j = 1, \ldots, \tilde{n}$*, all* $t \in [t_0, t_f]$*, and all* $\tau \in [t_0, t]$*, where the matrix* $F(t)$ *has the form*

$$F(t) = \begin{pmatrix} 0 & F_{12}(t) & -F_{12}(t) & 0 \\ F_{21}(t) & F_{22}(t) & F_{23}(t) & F_{24}(t) \\ -F_{21}(t) & -(F_{22}(t)+F_{23}(t)+F_{33}(t)) & F_{33}(t) & -F_{24}(t) \\ 0 & F_{42}(t) & -F_{42}(t) & 0 \end{pmatrix}. \tag{3.24}$$

Proof. Consider $\bar{F}(t) = T_B^{-1} F(t) T_A$ with T_B^{-1}, T_A given by (3.19). From Theorem 3.12, $\bar{F}_{11}(t) = 0$ and the other matrix blocks $\bar{F}_{ij}(t)$, $i, j = 1, 2$, can be identified. Thus we obtain the structure of the complementary matrix $F(t)$ given in (3.24). We get (3.23) by imposing (3.17). □

Note. From (3.20), we may identify some important cases in order to obtain possible structures of the complementary matrices $M(t)$ and $N(t)$. They are

$$\begin{aligned} &(a) \quad M_{12}(t) = 0, \quad M_{23}(t) + M_{33}(t) = 0, \quad M_{42}(t) = 0, \\ &(b) \quad M_{21}(t) = 0, \quad M_{22}(t) + M_{23}(t) = 0, \quad M_{24}(t) = 0, \quad N_{21}(t) = 0, \\ &\qquad N_{22}(t) + N_{23}(t) = 0, \quad N_{24}(t) = 0, \\ &(c) \quad \text{others} \end{aligned} \tag{3.25}$$

for all $t \in [t_0, t_f]$.

We focus our attention on the case (c) because it offers more freedom to select the corresponding complementary matrices than do cases (a) and (b). Now we assume that $M_{22}(t)+M_{33}(t) = 0$. Then (3.20) holds for $i > 2$, all $t \in [t_0, t_f]$, and all $\tau \in [t_0, t]$. If $i = 2$, choosing $M_{23}(t)+M_{33}(t) = 0$ or $M_{22}(t)+M_{23}(t) = 0$ in (3.20), two subcases of case (c) are obtained.

Subcase (c_1): $M_{23}(t) + M_{33}(t) = 0$. Then, relations (3.20) become

$$\begin{aligned} &\begin{pmatrix} M_{12}(t) \\ 0 \\ M_{42}(t) \end{pmatrix} \begin{pmatrix} \int_{t_0}^{t} M_{21}(\tau)\, d\tau & \int_{t_0}^{t} M_{22}(\tau)\, d\tau & \int_{t_0}^{t} M_{24}(\tau)\, d\tau \end{pmatrix} = 0, \\ &\begin{pmatrix} M_{12}(t) \\ 0 \\ M_{42}(t) \end{pmatrix} \begin{pmatrix} N_{21}(\tau) & N_{22}(\tau)+N_{23}(\tau) & N_{24}(\tau) \end{pmatrix} = 0. \end{aligned} \tag{3.26}$$

Subcase (c_2): $M_{22}(t) + M_{23}(t) = 0$. Then (3.20) becomes

$$\begin{aligned} &\begin{pmatrix} M_{12}(t) \\ M_{22}(t) \\ M_{42}(t) \end{pmatrix} \begin{pmatrix} \int_{t_0}^{t} M_{21}(\tau)\, d\tau & 0 & \int_{t_0}^{t} M_{24}(\tau)\, d\tau \end{pmatrix} = 0, \\ &\begin{pmatrix} M_{12}(t) \\ M_{22}(t) \\ M_{42}(t) \end{pmatrix} \begin{pmatrix} N_{21}(\tau) & N_{22}(\tau)+N_{23}(\tau) & N_{24}(\tau) \end{pmatrix} = 0 \end{aligned} \tag{3.27}$$

for all $t \in [t_0, t_f]$ and all $\tau \in [t_0, t]$.

The above cases give possible structures for choosing $M(t)$ and $N(t)$. A similar track should be followed to obtain structures for the other complementary matrices [3], [4], [19]. Finally, the designer can select specific values of free elements of the block complementary matrices in (3.26) or (3.27) according to given design requirements. Simultaneously, it should be checked that the matrix $M(t)$ satisfies condition (3.6) in Proposition 3.6. In the next section, this procedure is illustrated in the context of the design of overlapping decentralized controllers.

Note. It is important to recognize that it is not necessary to know the transition matrices explicitly in order to select the complementary matrices satisfying the required conditions.

4. Example.

Objective. Consider problem (2.1) for system (3.11) with the specific matrices

$$A(t) = \left(\begin{array}{ccc:c} -t & 0 & 0 & 0 \\ \hdashline 0 & 0 & 0 & -t \\ 0 & 0 & 0 & 0 \\ \hdashline 0 & 0 & 0 & -t \end{array}\right), \qquad B(t) = \left(\begin{array}{ccc:c} e^{-t} & 0 & 0 & 0 \\ \hdashline 0 & 0 & 0 & -0.5 \\ -1 & 0 & 0 & e^{-t} \\ \hdashline -3 & 0 & 0 & 0 \end{array}\right), \tag{4.1}$$

$\Pi = Q^* = diag(1,1,1,1)$, and $R^* = diag(1,1,1)$. The overlapping decomposition is determined by dashed lines. Consider the initial and the terminal time as $t_0 = 0$ and $t_f = 3$, respectively. System (4.1) is a commutative system by Definition 3.5.

The objective is to show the potential advantages offered by the characterization of the presented complementary matrices for an overlapping decentralized state LQ optimal control design.

We consider the following scheme as in [14].

(1) The pair $(\mathbf{S}, J)$ in (2.1) is expanded to $(\tilde{\mathbf{S}}, \tilde{J})$. The system $\tilde{\mathbf{S}}$ can be represented as

$$\tilde{\mathbf{S}}: \quad \dot{\tilde{x}}(t) = \tilde{A}_{D}(t)\,\tilde{x}(t) + \tilde{B}_{D}(t)\,\tilde{u}(t) + \tilde{A}_{C}(t)\,\tilde{x}(t) + \tilde{B}_{C}(t)\,\tilde{u}(t), \tag{4.2}$$

where $\tilde{A}_{D}(t)$, $\tilde{B}_{D}(t)$ are the block diagonal matrices and $\tilde{A}_{C}(t)$, $\tilde{B}_{C}(t)$ the corresponding interconnection matrices. Check the controllability of the pair $(\tilde{A}_{D}(t), \tilde{B}_{D}(t))$.

(2) A decentralized control law $\tilde{u}_{D}(t)$ is designed for the decoupled expanded system

$$\tilde{\mathbf{S}}_{\mathbf{D}}: \quad \dot{\tilde{x}}(t) = \tilde{A}_{D}(t)\,\tilde{x}(t) + \tilde{B}_{D}(t)\,\tilde{u}(t), \tag{4.3}$$

where $\tilde{u}_{D}(t) = -\tilde{K}_{D}(t)\,\tilde{x}(t)$ and $\tilde{K}_{D}(t) = (\tilde{R}^*)^{-1}(t)\tilde{B}_{D}^{T}(t)\tilde{P}_{D}(t)$. The matrix $\tilde{P}_{D}(t)$ is the symmetric, nonnegative definite solution of the Riccati equation

$$\dot{\tilde{P}}_{D}(t) = -\tilde{A}_{D}^{T}(t)\tilde{P}_{D}(t) - \tilde{P}_{D}(t)\tilde{A}_{D}(t) + \tilde{P}_{D}(t)\tilde{B}_{D}(t)(\tilde{R}^*)^{-1}(t)\tilde{B}_{D}^{T}(t)\tilde{P}_{D}(t) - \tilde{Q}^*(t) \tag{4.4}$$

with the boundary condition $\tilde{P}_{D}(t_f) = \tilde{\Pi}_{D}$.

(3) This control is contracted to $u_{D}(t) = -Q\tilde{K}_{D}(t)Vx(t) = -K_{D}(t)x(t)$, to be implemented into the original system $\mathbf{S}$. The evaluation of this control law is made

by means of the concept of suboptimality, which is determined by the value of the cost function in (2.1) for this controller. It is known that this value is given by

$$J^{\oplus}(x_0) = x_0^T H(t_0)\, x_0, \tag{4.5}$$

where $H(t)$ is the nonnegative definite solution of the differential Lyapunov equation

$$\begin{aligned} \dot{H}(t) = &- \left[A(t) - B(t)K_{D}(t)\right]^T H(t) - H(t)\left[A(t) - B(t)K_{D}(t)\right] \\ &- \left[Q^*(t) + K_{D}^T(t)R^*(t)K_{D}(t)\right] \end{aligned} \tag{4.6}$$

satisfying the boundary condition $H(t_f) = \Pi$.

In order to eliminate the dependence of $J^{\oplus}$ on the initial state, it is possible to assume x_0 as a random variable uniformly distributed over a dimensional unit sphere. Then the expected value of the performance criterion can be evaluated as $J^{\oplus} = \mathrm{tr}\{H(t_0)\}$, where $\mathrm{tr}\{\cdot\}$ denotes the trace operator [7], [17].

Consider the complementary matrices $N(t) = 0$, $M_{\Pi} = 0$, $M_{Q^*}(t) = 0$, $N_{R^*}(t) = 0$, and $F(t) = 0$, which are particular simple cases that verify Theorems 3.15 and 3.16. The complementary matrix $M(t)$ is selected within the cases described in the previous section, and $J^{\oplus}$ is computed. To evaluate and compare the results obtained by the proposed method, we will also consider the selection of $M(t)$ corresponding to the cases of aggregations and restrictions.

Results. *Overlapping decomposition using an aggregation.* Choosing a typical matrix M used in the literature [22], we obtain

$$M(t) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ A_{21} & \frac{1}{2}A_{22} & -\frac{1}{2}A_{22} & -A_{23} \\ -A_{21} & -\frac{1}{2}A_{22} & \frac{1}{2}A_{22} & A_{23} \\ 0 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & t \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -t \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{4.7}$$

with the corresponding expanded system matrix

$$\tilde{A}(t) = \begin{pmatrix} A_{11} & \frac{1}{2}A_{12} & \frac{1}{2}A_{12} & A_{13} \\ 2A_{21} & A_{22} & 0 & 0 \\ 0 & 0 & A_{22} & 2A_{23} \\ A_{31} & \frac{1}{2}A_{32} & \frac{1}{2}A_{32} & A_{33} \end{pmatrix} = \begin{pmatrix} -t & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -2t \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -t \end{pmatrix}. \tag{4.8}$$

Overlapping decomposition using a restriction. Another frequent choice of the matrix $M(t)$ [8], [12], [14], [22] is given by

$$M(t) = \begin{pmatrix} 0 & \frac{1}{2}A_{12} & -\frac{1}{2}A_{12} & 0 \\ 0 & \frac{1}{2}A_{22} & -\frac{1}{2}A_{22} & 0 \\ 0 & -\frac{1}{2}A_{22} & \frac{1}{2}A_{22} & 0 \\ 0 & -\frac{1}{2}A_{32} & \frac{1}{2}A_{32} & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{4.9}$$

with the corresponding expanded system matrix

$$\tilde{A}(t) = \begin{pmatrix} A_{11} & A_{12} & 0 & A_{13} \\ A_{21} & A_{22} & 0 & A_{23} \\ A_{21} & 0 & A_{22} & A_{23} \\ A_{31} & 0 & A_{32} & A_{33} \end{pmatrix} = \begin{pmatrix} -t & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -t \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -t \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -t \end{pmatrix}. \tag{4.10}$$

Overlapping decomposition using the proposed method. Considering subcase (c_1) with the purpose of maximizing the number of zeros in the off-diagonal blocks of

the expanded matrix $\tilde{A}(t)$, we select the submatrices of $M(t)$ as follows: $M_{12}(t) = \frac{1}{2}A_{12}(t) = \begin{pmatrix} 0 & 0 \end{pmatrix}$, $M_{21}(t) = A_{21}(t) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $M_{24}(t) = -A_{23}(t) = \begin{pmatrix} t \\ 0 \end{pmatrix}$. We also select $M_{42}(t) = \begin{pmatrix} 0 & t \end{pmatrix}$. From (3.26), if $M_{42}(t)M_{22}(t) = 0$, then the conditions for subcase (c_1) are satisfied. Consider $M_{22}(t) = \begin{pmatrix} 0 & m_{23} \\ 0 & 0 \end{pmatrix}$. Then the complementary matrix $M(t)$ has the form

$$M(t) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & m_{23} & 0 & m_{23} & t \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -m_{23} & 0 & -m_{23} & -t \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & t & 0 & -t & 0 \end{pmatrix} \tag{4.11}$$

with the degree of freedom to select the value of m_{23}. Notice that $M(t)$ satisfies condition (3.6), ensuring that $\tilde{\mathbf{S}}$ is a commutative system.

Following the steps given above and using an algorithm to minimize $\mathrm{tr}\{H(t_0)\}$ with respect to m_{23}, we can summarize the obtained results as follows.

Proposed method	Aggregation	Restriction	Centralized case
$J^{\oplus} = 7.95$ $m_{23} = -0.67$	$J^{\oplus} = 15.30$	$J^{\oplus} = 14.80$	$J^{\circ} = 5.39$

Note that J° is the cost for the centralized optimal control solving (2.1). Since a goal of a decentralized control is to drive the system as close as possible to the (ideal) centralized control, we may observe the best performance when using the proposed method for the selection of the complementary matrices. This method represents a reduction of 48% and 46.3% in the $J^{\oplus}$ with respect to the aggregations and restrictions, respectively. These results illustrate the freedom introduced by this approach in selecting the complementary matrices to minimize the cost function in overlapping decentralized control design. The minimization can be considered for more elements of $M(t)$ and also for other complementary matrices.

5. Conclusion. The inclusion principle has been specialized for a quadratic optimal control design for both general and commutative continuous-time LTV systems. The strategy of *generalized selection of complementary matrices* has been developed for commutative continuous-time LTV systems. It includes the presentation of a general structure of complementary matrices, including explicit conditions on them as well as optimization of their free elements. This structure offers flexibility in selection of complementary matrices, resulting in more appropriate costs when designing quadratic optimal control via overlapping decompositions for this class of systems. The efficiency of the proposed method has been demonstrated on an illustrative example.

REFERENCES

[1] M. Athans and P. L. Falb, *Optimal Control*, McGraw-Hill, New York, 1966.
[2] L. Bakule and J. Rodellar, *Decentralized control and overlapping decomposition of mechanical systems. Part* 1: *System decomposition. Part* 2: *Decentralized stabilization*, Internat. J. Control, 61 (1995), pp. 559–587.
[3] L. Bakule, J. Rodellar, and J. M. Rossell, *Structure of expansion-contraction matrices in the inclusion principle for dynamic systems*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1136–1155.
[4] L. Bakule, J. Rodellar, and J. M. Rossell, *Generalized selection of complementary matrices in the inclusion principle*, IEEE Trans. Automat. Control, 45 (2000), pp. 1237–1243.

[5] B. Barán, E. Kaszkurewicz, and A. Bhaya, *Parallel asynchronous team algorithms: Convergence and performance analysis*, IEEE Trans. Parallel and Distributed Systems, 7 (1996), pp. 677–688.
[6] Z. Drici, *New directions in the method of vector Lyapunov functions*, J. Math. Anal. Appl., 184 (1994), pp. 317–325.
[7] A. İftar, *Decentralized optimal control with extensions*, in Proceedings of the First IFAC Symposium on Design Methods of Control Systems, Zurich, Switzerland, 1991, pp. 747–752.
[8] A. İftar, *Overlapping decentralized dynamic optimal control*, Internat. J. Control, 58 (1993), pp. 187–209.
[9] A. İftar and E. J. Davison, *A decentralized discrete–time controller for dynamic routing*, Internat. J. Control, 69 (1998), pp. 599–632.
[10] A. İftar and Ü. Özgüner, *Overlapping decompositions, expansions, contractions, and stability of hybrid systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 1040–1055.
[11] M. Ikeda and D. D. Šiljak, *Overlapping decompositions, expansions and contractions of dynamic systems*, Large Scale Systems, 1 (1980), pp. 29–38.
[12] M. Ikeda, D. D. Šiljak, and D. E. White, *Decentralized control with overlapping information sets*, J. Optim. Theory Appl., 34 (1981), pp. 279–310.
[13] M. Ikeda, D. D. Šiljak, and D. E. White, *An inclusion principle for dynamic systems*, IEEE Trans. Automat. Control, 29 (1984), pp. 244–249.
[14] M. Ikeda, D. D. Šiljak, and D. E. White, *Overlapping decentralized control of linear time-varying systems*, in Advances in Large Scale Systems, Vol. 1, JAI Press, Greenwich, CT, 1984, pp. 93–116.
[15] H. Ito, *Overlapping decomposition for multirate decentralized control*, in Preprints of the 8th IFAC/IFORS/IMACS/IFIP Symposium on Large Scale Systems: Theory and Applications, Patras, Greece, 1998, pp. 259–264.
[16] E. W. Kamen, *Fundamentals of linear time-varying systems*, in The Control Handbook, CRC Press, Boca Raton, FL, 1996, pp. 451–467.
[17] W. S. Levine, T. L. Johnson, and M. Athans, *Optimal limited state variable feedback controllers for linear systems*, IEEE Trans. Automat. Control, 16 (1971), pp. 785–793.
[18] R. Ravi, K. M. Nagpal, and P. P. Khargonekar, *H^∞ control of linear time-varying systems: A state-space approach*, SIAM J. Control Optim., 29 (1991), pp. 1394–1413.
[19] J. M. Rossell, *Contribution to Decentralized Control of Large-Scale Systems via Overlapping Models*, Ph.D. thesis, Technical University of Catalunya, Barcelona, Spain, 1998 (in Spanish).
[20] W. J. Rugh, *Linear System Theory*, 2nd ed., Prentice–Hall, Upper Saddle River, NJ, 1996.
[21] M. E. Sezer and D. D. Šiljak, *Nested epsilon decompositions of linear systems: Weakly coupled and overlapping blocks*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 521–533.
[22] D. D. Šiljak, *Decentralized Control of Complex Systems*, Academic Press, New York, 1991.
[23] D. D. Šiljak, S. M. Mladenović, and S. Stanković, *Overlapping decentralized observation and control of a platoon of vehicles*, in Proceedings of the American Control Conference, San Diego, CA, 1999, pp. 4522–4526.
[24] S. S. Stanković, X. B. Chen, and D. D. Šiljak, *Stochastic inclusion principle applied to decentralized automatic generation control*, Internat. J. Control, 72 (1999), pp. 276–288.
[25] M. Y. Wu and A. Sherif, *On the commutative class of linear time-varying systems*, Internat. J. Control, 23 (1976), pp. 433–444.
[26] A. I. Zečević and D. D. Šiljak, *A block-parallel Newton method via overlapping epsilon decompositions*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 824–844.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 5, pp. 1628–1661

FINITE TIME–HORIZON RISK-SENSITIVE CONTROL AND THE ROBUST LIMIT UNDER A QUADRATIC GROWTH ASSUMPTION*

FRANCESCA DA LIO† AND WILLIAM M. MCENEANEY‡

Abstract. The finite time–horizon risk-sensitive limit problem for continuous nonlinear systems is considered. Previous results are extended to cover more typical examples. In particular, the cost may grow quadratically, and the diffusion coefficient may depend on the state. It is shown that the risk-sensitive value function is the solution of the corresponding dynamic programming equation. It is also shown that this value converges to the value of the robust control problem as the cost becomes infinitely risk averse, with corresponding scaling of the diffusion coefficient.

Key words. risk-sensitive control, robust, H_∞, viscosity solutions, nonlinear HJB equations, nonlinear Isaacs equations

AMS subject classifications. 35B37, 49L25, 90D25, 93B36, 93C10, 93E05, 93E20

PII. S0363012998345159

1. Introduction.

The nonlinear, finite time–horizon risk-sensitive limit problem is considered. It is, by now, well known that the value functions of risk-sensitive stochastic control problems tend to converge to the value functions of the corresponding robust/H_∞ control problems as one approaches infinite risk aversion. This was addressed first in the LEQG (linear-exponential-quadratic-Gaussian) case in which, in fact, one does not need to take the risk averse limit [28], [51], [6], [12]. In the nonlinear case, results were first developed for finite time–horizon control problems [52], [29], [17], [42], [43]. Further studies considered nonlinear, infinite time–horizon problems (in which case one gets the H_∞ limit) [18], [43], [19], [16], [44], [23], [47], [38], [31], [5] and nonlinear escape problems [37], [9]. Other studies have involved discrete systems [48], [13], [15] and the partial observations case [7], [30].

The results for both the finite time–horizon problem and the infinite time–horizon problem have mainly been obtained under assumptions which preclude quadratic cost criteria. In other words, the LEQG results were not subsumed by the nonlinear results. Since control systems for nonlinear systems are often designed by analogy with the linear-quadratic theory, it would be most desirable for the nonlinear theory to subsume the LEQG theory. For the infinite time–horizon problem, this is addressed in a preliminary fashion in [27] and [38] (with some results taken from [39], [40] in this latter case). The infinite time–horizon case presents certain technical difficulties which are not present in the finite time–horizon case. In particular, there is a certain lack of uniqueness of solutions to the dynamic programming equation (DPE) in the infinite time–horizon case which is not a problem in the finite time–horizon case. That is, there may be an infinite number of viscosity solutions and multiple classical solutions (when they exist) to the DPE. (This is in addition to the multiplicity incurred by scaling with an additive constant.) This question is addressed in [39], [40].

*Received by the editors September 25, 1998; accepted for publication (in revised form) July 31, 2001; published electronically February 6, 2002.
http://www.siam.org/journals/sicon/40-5/34515.html

†Department of Mathematics, University of Torino, via Carlo Alberto, 10 10123 Torino, Italy (dalio@dm.unito.it).

‡Department of Mathematics, North Carolina State University, Raleigh, NC 27695–8205 (wmm@math.ncsu.edu). Current address: Department of Mathematics, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0112 (wmceneaney@ucsd.edu). The research of this author was partially supported by AFOSR grant F49620-95-1-0296.

One approach to the finite time–horizon problem is to prove that the value function is a viscosity (or classical) solution to the associated DPE for the risk-sensitive problem. This is the approach taken by most of the work to date, and is the approach taken here as well. More specifically, let the parameter associated with risk-sensitivity be ε (described more accurately below). This parameter will also be used to scale the diffusion term. One wishes to show that as $\varepsilon \downarrow 0$ (i.e., as the problem becomes infinitely risk averse), the value function of the risk-sensitive problem converges to that of a robust (H_∞) control problem. It is easily seen that the DPE for the risk-sensitive problem formally converges to the DPE for the robust control problem as $\varepsilon \downarrow 0$. One may also show that the value of the robust control problem is a viscosity solution of this limit DPE. If one proves uniform boundedness and equicontinuity results for the prelimit value functions, then by the stability of viscosity solutions these converge (subsequentially) to viscosity solutions of the limit DPE. A key step is to prove a uniqueness result for viscosity solutions of this limit DPE. Then, one can assert that the solutions of the prelimit DPE (the risk-sensitive values) converge to the unique solution of the limit DPE, which must also then be the value of the limit robust control problem.

One of the key difficulties encountered in earlier attempts (cf. [17], [43]) employing the plan outlined in the previous paragraph was that the viscosity solution uniqueness results did not cover the limit DPEs except under rather strict conditions such as globally Lipschitz conditions on the cost and dynamics. In [42], viscosity solution uniqueness results were extended to cover certain cases in which the cost could grow quadratically. However, it was still required that one prove a certain Lipschitz condition with Lipschitz constant growing linearly with the norm of the state, uniformly for the prelimit solutions as well as the limit. Some recent results (obtained independently and concurrently; [3], [26]) have extended the uniqueness results in ways that do not require this Lipschitz condition. These results allow one to extend the previous risk-sensitive limit results to a larger class which subsumes a large variety of LEQG problems. (However, it does not subsume all, and in particular, the control set must be bounded.) Such extension is the subject of this paper. Consequently, the uniqueness result of Theorem 4.1 is a central topic of the paper. Note that the uniqueness results in [3], [26] do not cover the particular limit PDE encountered here. In section 4 of the present paper, the results of [3] are extended to yield the uniqueness result needed here.

Some other technical difficulties arise in this case as well. In particular, there is some difficulty with a Novikov condition that was not present previously (see the discussion prior to Lemma 2.2). Some difficulties also arise in the proof of the equicontinuity of the value functions. These last technical difficulties were overcome in this paper by consideration of the risk-sensitive value function as the value function for a stochastic game, and the use of known results showing that this game value is the limit of the value of a discrete game as the time-step size goes to zero. The delicateness of the required estimates is not surprising given the possible finite time blow-up property of the Riccati equations associated with linear-quadratic problems.

In sections 2 and 3, a representation result will be obtained that states that any classical solution of the risk-sensitive DPE must be the value function of the risk-sensitive problem. The assumptions will also be listed in that section. Further, in section 3, the existence of the above classical solution will be proved, and the uniform boundedness and equicontinuity results will be obtained along the way. Section 4 contains the limit result. Of course these results require the uniqueness result dis-

cussed above. This uniqueness result is stated and proved there. At the same time, it is shown that this unique solution is the value of the robust control problem, thus making the connection between the risk averse limit of the risk-sensitive problem and the robust control problem.

2. Risk-sensitive representation result. In the stochastic (prelimit) risk-sensitive problem, we consider a system of the form

$$\begin{aligned} dy_t^\varepsilon &= f(y_t^\varepsilon, u_t)\,dt + \sqrt{\frac{\varepsilon}{\gamma^2}}\sigma(y_t^\varepsilon)\,dB_t, \\ y_s^\varepsilon &= x, \end{aligned} \tag{1}$$

where y_t^ε is the state at time t taking values in $\Re^n$, x is the (known) initial state at time $s \geq 0$, f represents the nominal dynamics with control u taking values in $U \in \Re^l$, and $\{B_\cdot, \mathcal{F}_\cdot\}$ is an m-dimensional Brownian motion on the probability space $(\Omega, \mathcal{F}, P)$, where $\mathcal{F}_0$ contains all the P-negligible elements of $\mathcal{F}$ and σ is an $n \times m$-valued diffusion coefficient. The role of the parameters $\varepsilon, \gamma > 0$ will become more clear subsequently. However, perhaps it should be remarked here that ε will be a measure of the risk-sensitivity and scales the diffusion term in (1) above so that the cost below will remain bounded (for each x, as a function of ε).

We consider a cost criterion of the form

$$J^\varepsilon(s, x, u_\cdot) = \mathrm{E} \exp\left\{ \frac{1}{\varepsilon} \left[\int_s^T \ell(y_t^\varepsilon, u_t)\,dt + \psi(y_T^\varepsilon) \right] \right\}, \tag{2}$$

where $0 \leq s \leq T < \infty$, T is a fixed terminal time, ℓ is the running cost, and ψ is the terminal cost. The value function is

$$\begin{aligned} V^\varepsilon(s, x) &= \inf_{u \in \mathcal{U}_s} \varepsilon \log J^\varepsilon(s, x, u_\cdot) \\ &= \varepsilon \log \inf_{u \in \mathcal{U}_s} J^\varepsilon(s, x, u_\cdot), \end{aligned} \tag{3}$$

where $\mathcal{U}_s$ is the set of U-valued, $\mathcal{F}_t$-progressively measurable controls such that there exists a strong solution to (1).

We make the following assumptions throughout the paper.

(A0) U is a compact subset of $\Re^l$,

(A1)
$$\begin{cases} \text{i)} & f \in C^2(\Re^n \times U, \Re^n), & \\ \text{ii)} & |f_x(x,u)| \leq K & \forall x \in \Re^n,\ u \in U, \\ \text{iii)} & |f(x,u)| \leq K(1+|x|) & \forall x \in \Re^n,\ u \in U, \\ \text{iv)} & |f_{xx}(x,u)| \leq K/(1+|x|) & \forall x \in \Re^n,\ u \in U, \end{cases}$$

(A2)
$$\begin{cases} \text{i)} & \sigma \in C^2(\Re^n, \mathcal{L}(\Re^m, \Re^n)), & \\ \text{ii)} & |\sigma(x)| \leq M & \forall x \in \Re^n, \\ \text{iii)} & \xi^T \sigma(x) \sigma^T(x) \xi \geq \eta |\xi|^2 & \forall \xi, x \in \Re^n, \\ \text{iv)} & |\sigma_x(x)| \leq L_\sigma/(1+|x|) & \forall x \in \Re^n, \\ \text{v)} & |\sigma_{xx}(x)| \leq L_\sigma/(1+|x|^2) & \forall x \in \Re^n, \end{cases}$$

(A3)
$$\begin{cases} \text{i)} & \ell \in C^\infty(\Re^n \times U, \Re), \quad |\ell_{xx}(x,u)| \leq \overline{C} & \forall (x,u) \in \Re^n \times U, \\ \text{ii)} & \ell(x,u) \geq 0 & \forall (x,u) \in \Re^n \times U, \\ \text{iii)} & |\ell(x,u)| \leq C(1+|x|^2) & \forall (x,u) \in \Re^n \times U, \\ \text{iv)} & |\ell(x,u) - \ell(y,u)| \leq C(1+|x|+|y|)|x-y| & \forall x, y \in \Re^n, \forall u \in U, \end{cases}$$

(A4)

$$\begin{cases} \text{i)} & \psi \in C^\infty(\Re^n, \Re), \qquad |\psi_{xx}(x)| \le \overline{C} & \forall x \in \Re^n, \\ \text{ii)} & \psi(x) \ge 0 & \forall x \in \Re^n, \\ \text{iii)} & |\psi(x)| \le C(1+|x|^2) & \forall x \in \Re^n, \\ \text{iv)} & |\psi(x)-\psi(y)| \le C(1+|x|+|y|)|x-y| & \forall x, y \in \Re^n, \\ \text{v)} & \text{higher (above second) derivatives of } \psi \text{ uniformly bounded,} & \end{cases}$$

where $L_\sigma, C, K, M, \eta \in (0,\infty)$, and $\mathcal{L}(\Re^m, \Re^n)$ appearing in (A2.i) represents the space of $n \times m$ matrices. There is some intended redundancy in the above assumptions; for instance, (A1.iii) is included to indicate that the same constant, K, will be used as in (A1.ii), even though the existence of some constant for which (A1.iii) holds follows from (A0), (A1.i), (A1.ii). Note that many of the assumptions are needed only for the results of section 3, and in particular, are not needed for the uniqueness result of section 4. Specifically in section 4 we use (A1.iii), (A2.ii), (A3.ii)–(A3.iv), (A4.ii)–(A4.iv), while (A1.iv), (A2.v), and (A4.v) are used only to obtain the results in section 3. This is also true of the requirement of continuous, bounded second-derivatives in (A3.i) and (A4.i). It seems that many of the above smoothness assumptions, which are required in the prelimit analysis but not in the limit uniqueness result, might be removable; however, the prelimit analysis is already quite technical, and so no attempt is made to reduce the assumptions further. There will be one additional assumption in section 3.

In section 3, it will be shown that there exists a classical solution $\widetilde{V}^\varepsilon \in C^{1,2}$ to the PDE

$$\begin{aligned} 0 = {} & V_s + \frac{\varepsilon}{2\gamma^2} \sum_{i,j=1}^n a_{i,j}(x) V_{x_i x_j} + \frac{1}{2\gamma^2} \nabla V^T a(x) \nabla V \\ & + \min_{u \in U} [f(x,u) \cdot \nabla V + \ell(x,u)], \qquad 0 \le s \le T, \\ & V(T,x) = \psi(x), \end{aligned} \tag{4}$$

where $a \doteq \sigma\sigma^T$. For the purposes of this section, assume that the classical solution exists; we postpone the proof of this fact to the next section. Note that

$$\frac{1}{2\gamma^2} \nabla V^T a \nabla V = \max_{w \in \Re^m} \left[(\sigma w) \cdot \nabla V - \frac{\gamma^2}{2} \|w\|^2 \right]. \tag{5}$$

Define a measurable (see, for instance, [21]) feedback control

$$\overline{u}(t,x) \in \operatorname{argmin}[f(x,u) \cdot \nabla \widetilde{V}^\varepsilon(t,x) + \ell(x,u)]. \tag{6}$$

Using [50], one proves the existence of a strong solution, $\overline{y}^\varepsilon$, to (1) with feedback $\overline{u}$ (see [43]; also see [4] for a quicker proof of a weak solution). Then define the $\mathcal{F}_t$-progressively measurable control $\widetilde{u}_\cdot$ given by $\widetilde{u}_t(\omega) \doteq \overline{u}(t, \overline{y}^\varepsilon_t(\omega))$ for all $\omega \in \Omega$.

The first lemma is rather easily proved.

LEMMA 2.1.

$$\varepsilon \log J^\varepsilon(s,x,\widetilde{u}) \le \widetilde{V}^\varepsilon(s,x) \qquad \forall\, s \in [0,T], x \in \Re^n.$$

Proof. By Ito's rule and (1),

$$\widetilde{V}^\varepsilon(T,y_T^\varepsilon)-\widetilde{V}^\varepsilon(s,x)=\int_s^T\left[f(\overline{y}_t^\varepsilon,\widetilde{u}_t)\cdot\nabla\widetilde{V}^\varepsilon+\widetilde{V}_t^\varepsilon+\frac{\varepsilon}{2\gamma^2}\sum_{i,j=1}^n a_{i,j}(\overline{y}_t^\varepsilon)\widetilde{V}^\varepsilon_{x_ix_j}\right]dt$$
$$+\int_s^T\sqrt{\frac{\varepsilon}{\gamma^2}}(\nabla\widetilde{V}^\varepsilon)^T\sigma\cdot dB_t.$$

Using (4) and (6), one finds

$$\int_s^T\ell(\overline{y}_t^\varepsilon,\widetilde{u}_t)\,dt+\psi(\overline{y}_T^\varepsilon)=\widetilde{V}^\varepsilon(s,x)+\sqrt{\frac{\varepsilon}{\gamma^2}}\int_s^T(\nabla\widetilde{V}^\varepsilon)^T\sigma\,dB_t-\frac{1}{2\gamma^2}\int_s^T(\nabla\widetilde{V}^\varepsilon)^Ta\nabla\widetilde{V}^\varepsilon\,dt,$$

which yields

$$\mathrm{E}\exp\frac{1}{\varepsilon}\left[\int_s^T\ell(\overline{y}_t^\varepsilon,\widetilde{u}_t)\,dt+\psi(\overline{y}_T^\varepsilon)\right]$$
$$=e^{\frac{1}{\varepsilon}\widetilde{V}^\varepsilon(s,x)}\cdot\mathrm{E}\exp\left[\frac{1}{\sqrt{\gamma^2\varepsilon}}\int_s^T(\nabla\widetilde{V}^\varepsilon)^T\sigma\,dB_t-\frac{1}{2\gamma^2\varepsilon}\int_s^T(\nabla\widetilde{V}^\varepsilon)^Ta\nabla\widetilde{V}^\varepsilon\,dt\right].$$

By [25] (see also [46]), this last term is a supermartingale, and thus one has

$$\mathrm{E}\exp\frac{1}{\varepsilon}\left[\int_s^T\ell(\overline{y}_t^\varepsilon,\widetilde{u}_t)\,dt+\psi(\overline{y}_T^\varepsilon)\right]\le e^{\frac{1}{\varepsilon}\widetilde{V}^\varepsilon(s,x)}.\qquad\square$$

For the approach that we will use to prove a reverse inequality for (not necessarily optimal) controls, one needs to use Girsanov's theorem. Consequently, one is interested in controls for which a weak form of the Novikov condition holds. More specifically, we will be interested in controls for which there exists $\tau>0$, $m<\infty$ such that

$$\mathrm{E}\exp\frac{1}{2\gamma^2\varepsilon}\left[\int_t^{t+\tau}(\nabla\widetilde{V}^\varepsilon(t,y_t^\varepsilon))^Ta(y_t^\varepsilon)\nabla\widetilde{V}^\varepsilon(t,y_t^\varepsilon)\,dt\right]\le m\qquad\forall\,0\le s\le t\le t+\tau\le T. \tag{7}$$

The following lemma applies only to controls satisfying (7), but this lemma will also be useful in a generalization to appear at a later point in the paper.

LEMMA 2.2. *Let $0\le s\le T$ and $x\in\Re^n$. Let $u\in\mathcal{U}_s$ be a control such that* (7) *holds. Then*

$$\varepsilon\log J^\varepsilon(s,x,u)\ge\widetilde{V}^\varepsilon(s,x).$$

Proof. Let y^ε be the solution to (1) corresponding to the initial conditions and control in the lemma statement. Consider the change of drift given by

$$y_t^\varepsilon=x+\int_s^t f(y_r^\varepsilon,u_r)\,dr+\sqrt{\frac{\varepsilon}{\gamma^2}}\int_s^t\sigma(y_r^\varepsilon)\,dB_r \tag{8}$$
$$=x+\int_s^t[f(y_r^\varepsilon,u_r)+\sigma(y_r^\varepsilon)w_r]\,dr+\sqrt{\frac{\varepsilon}{\gamma^2}}\int_s^t\sigma(y_r^\varepsilon)\,d\widehat{B}_r,$$

where w is an adapted process to be given below. By (7) (see, for instance, [32]), $\widehat{B}$ is a Brownian motion under new probability measure, $\widehat{P}$, where

$$P(d\omega)=\exp\Bigg[-\frac{1}{\sqrt{\gamma^2\varepsilon}}\int_s^t(\nabla\widetilde{V}^\varepsilon(r,y_r^\varepsilon))^T\sigma(y_r^\varepsilon)\,d\widehat{B}_r$$
$$-\frac{1}{2\gamma^2\varepsilon}\int_s^t(\nabla\widetilde{V}^\varepsilon(r,y_r^\varepsilon))^Ta(y_r^\varepsilon)\nabla\widetilde{V}^\varepsilon(r,y_r^\varepsilon)\,dr\Bigg]\widehat{P}(d\omega).$$

Consequently,

$$\begin{aligned}
&\mathrm{E}\exp\frac{1}{\varepsilon}\left[\int_s^T \ell(y_t^\varepsilon,u_t)\,dt+\psi(y_T^\varepsilon)\right]\\
&\quad=\widehat{\mathrm{E}}\exp\frac{1}{\varepsilon}\Bigg[\int_s^T\left(\ell(y_t^\varepsilon,u_t)-\frac{1}{2\gamma^2}(\nabla\widetilde{V}^\varepsilon(t,y_t^\varepsilon))^T a(y_t^\varepsilon)\nabla\widetilde{V}^\varepsilon(t,y_t^\varepsilon)\right)dt\\
&\qquad\qquad-\sqrt{\frac{\varepsilon}{\gamma^2}}\int_s^T(\nabla\widetilde{V}^\varepsilon(t,y_t^\varepsilon))^T\sigma(y_t^\varepsilon)\,d\widehat{B}_t+\psi(y_T^\varepsilon)\Bigg].
\end{aligned}\tag{9}$$

By Ito's rule with dynamics (8),

$$\begin{aligned}
\widetilde{V}^\varepsilon(T,y_T^\varepsilon)-\widetilde{V}^\varepsilon(s,x)=\int_s^T\Bigg[&\widetilde{V}_t^\varepsilon+(f(y_t^\varepsilon,u_t)+\sigma(y_t^\varepsilon)w_t)\cdot\nabla\widetilde{V}^\varepsilon+\frac{\varepsilon}{2\gamma^2}\sum_{i,j=1}^n a_{i,j}(y_t^\varepsilon)\widetilde{V}^\varepsilon_{x_ix_j}\Bigg]dt\\
&+\sqrt{\frac{\varepsilon}{\gamma^2}}\int_s^T(\nabla\widetilde{V}^\varepsilon)^T\sigma(y_t^\varepsilon)\,d\widehat{B}_t.
\end{aligned}$$

Taking $w_t\doteq\frac{1}{\gamma^2}\sigma^T(y_t^\varepsilon)\nabla\widetilde{V}^\varepsilon(y_t^\varepsilon)$ yields

$$\begin{aligned}
\psi(y_T^\varepsilon)-\widetilde{V}^\varepsilon(s,x)=\int_s^T\Bigg[&\widetilde{V}_t^\varepsilon+f(y_t^\varepsilon,u_t)\cdot\nabla\widetilde{V}^\varepsilon+\frac{1}{2\gamma^2}(\nabla\widetilde{V}^\varepsilon)^T a(y_t^\varepsilon)\nabla\widetilde{V}^\varepsilon\\
&+\frac{\varepsilon}{2\gamma^2}\sum_{i,j=1}^n a_{i,j}(y_t^\varepsilon)\widetilde{V}^\varepsilon_{x_ix_j}\Bigg]dt\\
&+\frac{1}{2\gamma^2}\int_s^T(\nabla\widetilde{V}^\varepsilon)^T a(y_t^\varepsilon)\nabla\widetilde{V}^\varepsilon\,dt+\sqrt{\frac{\varepsilon}{\gamma^2}}\int_s^T(\nabla\widetilde{V}^\varepsilon)^T\sigma(y_t^\varepsilon)\,d\widehat{B}.
\end{aligned}\tag{10}$$

Now, since this choice of u is not necessarily optimal in (4), one has at any time

$$0\le\widetilde{V}_t^\varepsilon+\frac{\varepsilon}{2\gamma^2}\sum_{i,j=1}^n a_{i,j}(y_t^\varepsilon)\widetilde{V}^\varepsilon_{x_ix_j}+\frac{1}{2\gamma^2}(\nabla\widetilde{V}^\varepsilon)^T a(y_t^\varepsilon)\nabla\widetilde{V}^\varepsilon+f(y_t^\varepsilon,u_t)\nabla\widetilde{V}^\varepsilon+\ell(y_t^\varepsilon,u_t).\tag{11}$$

Combining (10) and (11) yields

$$\begin{aligned}
\exp\frac{1}{\varepsilon}\widetilde{V}^\varepsilon(s,x)\le\exp\frac{1}{\varepsilon}\Bigg[&\int_s^T\ell(y_t^\varepsilon,u_t)\,dt+\psi(y_T^\varepsilon)-\sqrt{\frac{\varepsilon}{\gamma^2}}\int_s^T(\nabla\widetilde{V}^\varepsilon)^T\sigma(y_t^\varepsilon)\,d\widehat{B}\\
&-\frac{1}{2\gamma^2}\int_s^T(\nabla\widetilde{V}^\varepsilon)^T a(y_t^\varepsilon)\nabla\widetilde{V}^\varepsilon\,dt,\Bigg],
\end{aligned}$$

which, by (9), implies

$$\exp\frac{1}{\varepsilon}\widetilde{V}^\varepsilon(s,x)\le\mathrm{E}\exp\frac{1}{\varepsilon}\left[\int_s^T\ell(y_t^\varepsilon,u_t)\,dt+\psi(y_T^\varepsilon)\right].\qquad\square$$

Summarizing the results in this section, we have the following.

LEMMA 2.3. *Let $0 \le s \le T$ and $x \in \Re^n$. For any $u \in \mathcal{U}_s$ such that (7) holds, one has*

$$\varepsilon \log J^\varepsilon(s,x,u) \ge \widetilde{V}^\varepsilon(s,x),$$

while with optimal $\widetilde{u}$ given by (6) one has

$$\varepsilon \log J^\varepsilon(s,x,\widetilde{u}) \le \widetilde{V}^\varepsilon(s,x).$$

Lemma 2.3 is not quite what we will obtain for the risk-sensitive prelimit result. The lemma is deficient, due to the requirement that (7) hold in order to obtain the first assertion. Without that condition, the lemma would show that $\widetilde{V}^\varepsilon$ is the value of the risk-sensitive control problem. In the next section, while proving the existence of a classical solution to (4), we will obtain results that allow one to improve Lemma 2.3 in this way; specifically, that the value, V^ε, given by (3) is this classical solution. The improved result will be Theorem 3.11.

3. Existence and uniformity. The representation results of the previous section required the existence of a $C^{1,2}$ solution to (4). That will be obtained in this section. Also, in order to prove the risk averse limit result to follow, we will need continuity and local boundedness estimates on $\widetilde{V}^\varepsilon$ independent of $\varepsilon > 0$. These will be obtained at the same time as the existence result. Finally, we will obtain the promised improvement over Lemma 2.3. These results will all be obtained through the stochastic game representation.

Before proceeding with the technical details, we give some formal indications of the general concepts.

Using convex duality, one may rewrite DPE (4) as

$$0 = V_s + \frac{\varepsilon}{2\gamma^2}\sum_{i,j=1}^n a_{i,j}(x)V_{x_ix_j} + \min_{u\in U}\max_{w\in\Re^m}\left\{[f(x,u)+\sigma(x)w]\nabla V + \ell(x,u) - \frac{\gamma^2}{2}|w|^2\right\},$$

$$V(T,x) = \psi(x). \tag{12}$$

In this form, it is intuitive that this PDE might be associated with the stochastic differential game with dynamics

$$dy_t^\varepsilon = [f(y_t^\varepsilon,u_t) + \sigma(y_t^\varepsilon)w_t]\,dt + \sqrt{\frac{\varepsilon}{\gamma^2}}\sigma(y_t^\varepsilon)\,dB_t,$$

$$y_s^\varepsilon = x \tag{13}$$

and payoff

$$J_g^\varepsilon(s,x;u,w) = \mathrm{E}\left[\int_s^T \ell(y_t^\varepsilon,u_t) - \frac{\gamma^2}{2}|w_t|^2\,dt + \psi(y_T^\varepsilon)\right].$$

In this game, $w.$ is the control for the opposing (disturbance) player who is trying to maximize this payoff which we are trying to minimize. The definitions of value for such a game [24] are discussed further below. The connections between such games and their corresponding DPEs (now second-order Isaacs equations) are not known for the class of systems considered here, so it will be necessary to consider some more restrictive assumptions and then to relax these assumptions to obtain the above game and DPE.

One technique that is sometimes useful for games associated with H_∞ control is to obtain an L_2-bound on ε-optimal controls, w, for the disturbance player [41], [39], [40]. This is then used to obtain continuity estimates on the value function. The simplest analogue is to prove a bound of the form $\|w\|_{L_2([s,T]\times\Omega)} \le M < \infty$ for ε-optimal w. Such a bound is not difficult to obtain in this case, but it does not yield the equicontinuity estimates for the value needed for the proofs to follow. Instead, one would ideally obtain an almost sure bound of some form. This bound is rather technical due to the presence of the Brownian motion with its unbounded variation. Instead, a discrete game is analyzed here in order to obtain an analogous bound and consequent continuity estimates. A limit is then taken to obtain the continuity estimates for the original game. We now proceed with the analysis.

We first work with a system where the dynamics and cost functions are all bounded. Let

$$\widetilde{\ell}^R(x,u) = \begin{cases} \ell(x,u) & \text{if } \ell(x,u) \le R, \\ R & \text{otherwise,} \end{cases} \tag{14}$$

$$\widetilde{\psi}^R(x) = \begin{cases} \psi(x) & \text{if } \psi(x) \le R, \\ R & \text{otherwise,} \end{cases} \tag{15}$$

$$\widetilde{f}^b(x,u) = \begin{cases} f(x,u) & \text{if } |f(x,u)| \le b, \\ b\frac{f(x,u)}{|f(x,u)|} & \text{otherwise.} \end{cases} \tag{16}$$

The final assumption in the paper is that for all R sufficiently large one can mollify $\widetilde{\ell}^R, \widetilde{\psi}^R, \widetilde{f}^b$ yielding ℓ^R, ψ^R, f^b such that

$$\left\{ \begin{array}{ll} \ell^R \in C^2(\Re^n \times U) & \\ \ell^R(x,u) = \widetilde{\ell}^R(x,u) & \forall |x| \le D_R^\ell \\ \ell^R(x,u) \le R & \forall x \in \Re^n \\ |\nabla \ell^R(x,u)| \le 2C'|x| & \forall x \in \Re^n \\ |\ell^R_{xx}(x,u)| \le \overline{C}' & \forall x \in \Re^n \end{array} \right\}, \quad \left\{ \begin{array}{ll} \psi^R \in C^2(\Re^n) & \\ \psi^R(x) = \widetilde{\psi}^R(x) & \forall |x| \le D_R^\psi \\ \psi^R(x) \le R & \forall x \in \Re^n \\ |\nabla \psi^R(x)| \le 2C'|x| & \forall x \in \Re^n \\ |\psi^R_{xx}(x)| \le \overline{C}' & \forall x \in \Re^n \end{array} \right\},$$

$$\left\{ \begin{array}{ll} f^b \in C^2(\Re^n \times U) & \\ f^b(x,u) = \widetilde{f}^b(x,u) & \forall |x| \le D_b^f, u \in U \\ |f^b(x,u)| \le b & \forall x \in \Re^n, u \in U \\ |f^b_x(x,u)| \le K' & \forall x \in \Re^n, u \in U \\ |f^b_{xx}(x,u)| \le K'/(1+|x|) & \forall x \in \Re^n, u \in U \end{array} \right\}, \tag{A5}$$

where $D_R^\ell, D_R^\psi \to \infty$ as $R \to \infty$, $D_b^f \to \infty$ as $b \to \infty$. We claim that this can be achieved by a relatively standard mollification, but as we do not want to take up space with the technical details, we make it an assumption. In particular, for the ψ case we claim that one would use the following convolution for $R > C$: $\psi^R(x) = \int J_\varepsilon(x-y)\widetilde{\psi}^R(y)\,dy$, where $J_1(z) = k\exp\{-1/(1-|z|^2)\}$ on $|z| \le 1$ and 0 elsewhere, where k is such that $\int_{\Re^n} J_1 = 1$, $J_\varepsilon(z) = \varepsilon^{-n} J_1(z/\varepsilon)$, $D_R^\psi = \frac{1}{2}\sqrt{R/C-1}$, $\varepsilon = 0$ for $|x| \le D_R^\psi$, $\varepsilon = \frac{1}{2}D_R^\psi \exp\{1 - 1/[1 - |(|x| - 3D_R^\psi/2)/(D_R^\psi/2)|]\}$ for $D_R^\psi < |x| \le 3D_R^\psi/2$, and $\varepsilon = D_R^\psi/2$ for $3D_R^\psi/2 < |x|$. The bounds on the derivatives use the smoothness of ψ itself (as well as $\max_{x,R}|\varepsilon_x|$, $\max_{x,R}|\varepsilon_{xx}|$ bounded for sufficiently large R) on $|x| \le 3D_R^\psi/2$ and then rely on the fact that the derivatives can be translated onto J_ε itself in the convolution for $|x| > 3D_R^\psi/2$ (cf. [1],[53]). Again, by making (A5) an assumption, we avoid the details and leave them only to an interested reader.

Also, let

$$\mathcal{W}_s \doteq \{\mathcal{F}_t\text{-adapted, right-continuous } w. \text{ such that } \|w\|_{L_2([s,T]\times\Omega)} < \infty\}, \tag{17}$$

$$\mathcal{W}_s^m \doteq \{w \in \mathcal{W}_s : |w_t(\omega)| \le m \quad \forall t \in [s,T], \omega \in \Omega\}. \tag{18}$$

Note that condition (17) implies that any $w \in \mathcal{W}_s$ is $\mathcal{F}_t$-progressively measurable; see [22], [10]. Let the dynamics of the modified game be

$$\begin{aligned} dy_t^{\varepsilon,b} &= [f^b(y_t^{\varepsilon,b}, u_t) + \sigma(y_t^{\varepsilon,b})w_t]\,dt + \sqrt{\frac{\varepsilon}{\gamma^2}}\sigma(y_t^{\varepsilon,b})\,dB_t, \\ y_s^{\varepsilon,b} &= x, \end{aligned} \tag{19}$$

and define the payoff by

$$J_g^{\varepsilon,b,R}(s,x;u,w) = \mathrm{E}\left\{\int_s^T \left[\ell^R(y_t^{\varepsilon,b}, u_t) - \frac{\gamma^2}{2}|w_t|^2\right] dt + \psi^R(y_T^{\varepsilon,b})\right\}. \tag{20}$$

An admissible strategy for the player with control u is a mapping $\theta : \mathcal{W}_s \to \mathcal{U}_s$ (alternatively, $\theta : \mathcal{W}_s^m \to \mathcal{U}_s$ when the disturbance set is $\mathcal{W}_s^m$) such that if, almost surely, $w_r = \widetilde{w}_r$ for almost every $r \in [s,t]$, then, almost surely, $\theta[w]_r = \theta[\widetilde{w}]_r$ for almost every $r \in [s,t]$, for any $t \in [s,T]$; see [24]. The set of all admissible strategies will be denoted by Θ_s. Define the set of admissible strategies for the player with control w analogously, and denote them by Λ_s in the case in which the disturbance set is $\mathcal{W}_s$. (In the case in which the disturbance set is $\mathcal{W}_s^m$, denote it by Λ_s^m.) Then (see [24]), the upper and lower values are

$$\underline{V}^{\varepsilon,b,R,m}(s,x) = \inf_{\theta\in\Theta_s}\sup_{w\in\mathcal{W}_s^m} J_g^{\varepsilon,b,R}(s,x,\theta[w],w), \tag{21}$$

$$\overline{V}^{\varepsilon,b,Rm}(s,x) = \sup_{\lambda\in\Lambda_s^m}\inf_{u\in\mathcal{U}_s} J_g^{\varepsilon,b,R}(s,x,u,\lambda[u]), \tag{22}$$

and we note that, by (20), these are bounded. Then, by [24], $\underline{V}^{\varepsilon,b,R,m} = \overline{V}^{\varepsilon,b,R,m}$ is the unique, bounded, continuous viscosity solution to the Isaacs equation

$$\begin{aligned} 0 = V_s + \frac{\varepsilon}{2\gamma^2}\sum_{i,j=1}^n a_{i,j}(x)V_{x_ix_j} + \min_{u\in U}\max_{|w|\le m}\Big\{&[f^b(x,u)+\sigma(x)w]\cdot\nabla V + \ell^R(x,u) \\ &- \frac{\gamma^2}{2}|w|^2\Big\}, \\ V(T,x) = \psi^R(x). \end{aligned} \tag{23}$$

LEMMA 3.1. *$\underline{V}^{\varepsilon,b,R,m} = \overline{V}^{\varepsilon,b,R,m}$ is the unique, bounded, $C^{1,2}$ solution to* (23).

Proof. From the above, we know that the equality in Lemma 3.1 is the unique, bounded, continuous viscosity solution. Additionally, [24] tells us that it is Lipschitz in x and Holder continuous in s. (We will be interested in specific bounds independent of ε, b, R, m further below, but for simplicity we do not look for these here.) Under our assumptions, it is not difficult to show that it is also Lipschitz in s, and we do so now.

Fix $s, s+\delta \in [0,T]$, $\delta > 0$. We will bound $|\underline{V}^{\varepsilon,b,R,m}(s,x) - \underline{V}^{\varepsilon,b,R,m}(s+\delta,x)|$. Fix any $x \in \Re^n$, $u \in \mathcal{U}_s$, and $w \in \mathcal{W}_s^m$. Let the corresponding solution of (19) be denoted

by $y_\cdot^{0,\varepsilon,b}$. Recall that the probability space is $(\Omega, \{\mathcal{F}_t\}, P)$. Define a new Brownian motion $B_\cdot^\delta$ by $B_t^\delta = B_{t-\delta}$ on a new probability space with filtration $\mathcal{F}_t^0 = \mathcal{F}_{t-\delta}$. Let $u_t^\delta = u_{t-\delta}$ and $w_t^\delta = w_{t-\delta}$. Let the solution of (19) corresponding to this new Brownian motion, controls u^δ, w^δ, and initial condition x at time $s+\delta$ be denoted by $y_\cdot^{\delta,\varepsilon,b}$. Then

$$y_t^{\delta,\varepsilon,b} = y_{t-\delta}^{0,\varepsilon,b} \qquad \forall\, t \in [\delta, T+\delta].$$

Consequently, one has

$$J_g^{\varepsilon,b,R}(s+\delta, x; u^\delta, w^\delta) = \mathrm{E}\left[\int_s^{T-\delta} \ell^R(y_t^{0,\varepsilon,b}, u_t) - \frac{\gamma^2}{2}\|w_t\|^2\, dt + \psi^R(y_{T-\delta}^{0,\varepsilon,b})\right].$$

Therefore,

$$\begin{aligned} &J_g^{\varepsilon,b,R}(s+\delta, x; u^\delta, w^\delta) - J_g^{\varepsilon,b,R}(s, x; u, w) \\ &\quad = \mathrm{E}\left[\psi^R(y_{T-\delta}^{0,\varepsilon,b}) - \int_{T-\delta}^{T}\left(\ell^R(y_t^{0,\varepsilon,b}, u_t) - \frac{\gamma^2}{2}\|w_t\|^2\right) dt - \psi^R(y_T^{0,\varepsilon,b})\right] \\ &\quad \le \left(R + \frac{\gamma^2}{2}m^2\right)\delta + \mathrm{E}\left[\psi^R(y_{T-\delta}^{0,\varepsilon,b}) - \psi^R(y_T^{0,\varepsilon,b})\right] \end{aligned}$$

which by Ito's rule

$$\begin{aligned} &= \left(R + \frac{\gamma^2}{2}m^2\right)\delta + \mathrm{E}\Bigg\{\int_{T-\delta}^{T} \nabla\psi^R(y_t^{0,\varepsilon,b}) \cdot \left[f^b(y_t^{0,\varepsilon,b}, u_t) + \sigma(y_t^{0,\varepsilon,b}) w_t\right] dt \\ &\qquad\qquad + \frac{\varepsilon}{2\gamma^2}\int_{T-\delta}^{T} \sum_{i=1,j=1}^{n} \psi_{x_i x_j}^R(y_t^{0,\varepsilon,b}) a_{ij}(y_t^{0,\varepsilon,b})\, dt\Bigg\} \\ &\le \left[R + \frac{\gamma^2}{2}m^2 + 2M_R(b+Mm) + \frac{\varepsilon \overline{C} M^2}{2\gamma^2}\right]\delta. \end{aligned}$$

From this (and an analogous bound on the other side using approximately optimal w), one easily finds that $\underline{V}^{\varepsilon,b,R,m}$ is Lipschitz in s.

On the other hand, (23) has a $C^{1,2}$ solution which is globally Lipschitz and bounded ([45]; see also [35], [21]). Since classical solutions are viscosity solutions, the uniqueness of viscosity solutions for (23) implies that $\underline{V}^{\varepsilon,b,R,m}$ is, in fact, this $C^{1,2}$ solution. □

Lipschitz bounds dependent on ε, b, R, m were sufficient to obtain the above result. Below, however, we will need bounds independent of these parameters. These bounds require more technical arguments, which we now present. The first bound is actually quite straightforward.

LEMMA 3.2. *There exists $\widetilde{\gamma} < \infty$ such that for all $\gamma > \widetilde{\gamma}$,*

$$\underline{V}^{\varepsilon,b,R,m}(s,x) \le M_T(1+|x|^2) \qquad \forall\, s \in [0,T], \quad x \in \Re^n,$$

and further, for $\delta \le 1$, any δ-optimal $w_\delta \in \mathcal{W}_s$ with respect to any $\theta \in \Theta_s$ (i.e., $J_g^{\varepsilon,b,R}(s,x;\theta[w_\delta], w_\delta) \ge \sup_{w\in\mathcal{W}_s} J_g^{\varepsilon,b,R}(s,x;\theta[w],w) - \delta$) satisfies

$$\|w_\delta\|^2_{L_2(\Omega\times[0,T])} \le M_T'(1+|x|^2). \tag{24}$$

We note that M_T, M_T' depend on T but are independent of ε, b, R, m.

Remark 3.3. A value for $\widetilde{\gamma}$ is easily obtained from the proof below. Since this $\widetilde{\gamma}$ is general for all systems satisfying the assumptions, it is quite conservative. If one has additional structure such as contractivity of f, a much smaller value for $\widetilde{\gamma}$ can be obtained [39], [40].

Proof. Suppressing certain arguments for notational simplicity, one has

$$|y_t^{\varepsilon,b}|^2 \le |x|^2 + 2\int_s^t (y_r^{\varepsilon,b})^T[f^b + \sigma w_r]\,dr + 2\sqrt{\frac{\varepsilon}{\gamma^2}}\int_s^t (y_r^{\varepsilon,b})^T\sigma\,dB_r + \frac{\varepsilon}{\gamma^2}\int_s^t \sum a_{ii}\,dr,$$

where (see [36], for example) the stochastic integral is a square-integrable martingale. Then, using (A1.iii), (A2.ii),

$$\begin{aligned}\mathrm{E}|y_t^{\varepsilon,b}|^2 &\le |x|^2 + 2K\mathrm{E}\int_s^t (|y_r^{\varepsilon,b}| + |y_r^{\varepsilon,b}|^2)\,dr + 2M\mathrm{E}\int_s^t |y_r^{\varepsilon,b}||w_r|\,dr + \frac{\varepsilon}{\gamma^2}M^2(t-s)\\ &\le |x|^2 + (3K+M)\int_s^t \mathrm{E}|y_r^{\varepsilon,b}|^2\,dr + \Big(\frac{\varepsilon}{\gamma^2}M^2 + K\Big)(t-s) + M\int_s^t \mathrm{E}|w_r|^2\,dr.\end{aligned}$$

By Gronwall's inequality, there exists $C_t < \infty$ such that

$$\mathrm{E}|y_t^{\varepsilon,b}|^2 \le C_t(1 + |x|^2 + \|w\|^2_{L_2(\Omega\times[0,T])}).$$

Consequently,

$$J_g^{\varepsilon,b,R}(s,x;u,w) \le C_T'(1+|x|^2) + (\widetilde{\gamma}^2 - \gamma^2)\|w\|^2_{L_2(\Omega\times[0,T])}$$

for proper choice of $C_T', \widetilde{\gamma}$ which depend on T but not b, R, and $\varepsilon \le 1$. This yields the first assertion. The second then follows by comparison with $\overline{w} \equiv 0$. □

The proof of continuity estimates uniform over b, R, m, and $\varepsilon \le 1$ is considerably more technical than the above results. In the deterministic setting, an L_2-bound on δ-optimal w is sufficient to yield equicontinuity estimates. However, in the stochastic case, the type of L_2-bound obtained in the previous lemma is not sufficient. This is due to the fact that L_2-bounds over the sample space as well as the state space do not imply bounds on the expectation of the exponentiation of the state space L_2-norm of w, and this is exactly what appears when using Gronwall-type estimates. (Recall that we are only assuming Lipschitz conditions on the dynamics, not stability.) However, one expects that if a set of Brownian paths does not diverge much from the origin, then the near-optimal disturbance paths over that portion of the sample space should not have "too large" L_2-norms. Measurability conditions make it difficult to refine that idea into a proof in the continuous-time case, so we work first with discrete-time approximations and then use well-known limit results to obtain the continuity estimate.

LEMMA 3.4. *For γ sufficiently large (i.e., $\gamma \ge \overline{\gamma}$), there exists $\overline{C}_{T,D} < \infty$ (depending on $T, D \in [0,\infty)$ but independent of b, R, m, ε) such that*

$$|\underline{V}^{\varepsilon,b,R,m}(s,x) - \underline{V}^{\varepsilon,b,R,m}(\hat{s},\hat{x})| \le \overline{C}_{T,D}[|x-\hat{x}| + |s-\hat{s}|] \qquad \forall\, s,\hat{s} \in [0,T],\ x,\hat{x} \in B_D(0).$$

Actually, $C_{T,D}$ is linear in D for the space estimate.

Proof. We first obtain the Lipschitz estimate in the space variable. This requires the discrete game approximation. Once one has this, the estimate in the time variable follows without having to use discrete approximations. The proof is very technical, and a significant portion of it is moved to Appendix A.

Define the discrete approximation as follows. Let the number of time-steps be N, and the step size is then $\Delta \doteq T/N$. We assume $\Delta \leq 1$ freely throughout the proof. Let B_n be a sequence of independent identically distributed m-dimensional random variables with $\mathrm{E}[B_n] = 0$, $\mathrm{E}[B_n B_n^T] = I_m$, and finite range. In particular, we will actually take the B_n such that the components always consist entirely of the terms -1 and 1 in order to simplify the proof. Let the B_n be adapted to the sequence of σ-algebras $\mathcal{F}_n$, where $\mathcal{F}_0$ is trivial. Let the control and disturbance at step n be u_n and w_n, which are measurable with respect to $\mathcal{F}_{n-1}$. We suppress the b and R superscripts on f, ℓ, ψ and also use the original bounds (without the $'$ superscripts) of assumption (A5) to simplify matters—we note that these coefficients were independent of b, R. The dynamics for the discrete game are then

$$\overline{y}_n = \overline{y}_{n-1} + [f(\overline{y}_{n-1}, u_n) + \sigma(\overline{y}_{n-1})w_n]\Delta + \sqrt{\frac{\varepsilon\Delta}{\gamma^2}}\sigma(\overline{y}_{n-1})B_n. \tag{25}$$

Let $\mathcal{W}^N$ be the set of sequences of L_2 disturbances, $\{w_n\}_{n=1}^N$, of length N, where each w_n is $\mathcal{F}_{n-1}$-measurable. Let Θ^N be the set of admissible strategies for the player with control u, where admissibility is defined analogously to that in the continuous-time case above; we do not repeat it here. For any $\theta \in \Theta^N$, $w \in \mathcal{W}^N$, and initial state $y_0 = x$, define the payoff

$$J^N(x, \theta, w) = \mathrm{E}\left[\sum_{i=1}^{N}\left[\ell(\overline{y}_{n-1}, u_n) - \frac{\gamma^2}{2}|w_n|^2\right]\Delta + \psi(\overline{y}_N)\right]$$

and value

$$V^N(x) = \inf_{\theta\in\Theta^N} \overline{V}^N(x, \theta),$$
$$\overline{V}^N(x, \theta) = \sup_{w\in\mathcal{W}^N} J^N(x, \theta, w).$$

Note that if one obtains a Lipschitz bound in x on $\overline{V}^N(x, \theta)$ that is uniform over $\theta \in \Theta^N$, then one automatically obtains the desired Lipschitz bound in x on $V^N(x)$. Consequently, we will work only with $\overline{V}^N(x, \theta)$ and, again to simplify the notation, drop the θ argument, simply referring to $\overline{V}^N(x)$.

We will use dynamic programming to work back from our assumed Lipschitz continuity estimate on ψ to obtain the estimate on V^N. Let

$$\overline{V}_N^N(x_N) = \psi(x_N),$$
$$\overline{J}_{n-1}^N(x_{n-1}, v_n) = \mathrm{E}\left[\left(\ell(x_{n-1}) - \frac{\gamma^2}{2}|v_n|^2\right)\Delta + \overline{V}_n^N(y_n)\right],$$
$$\overline{V}_{n-1}^N(x_{n-1}) = \sup_{v_n\in\Re^m} \overline{J}_{n-1}^N(x_{n-1}, v_n) \qquad \forall\, n = 1, 2, \ldots, N,$$

where we again drop θ, and let y_n be given by

$$y_n = x_{n-1} + [f(x_{n-1}) + \sigma(x_{n-1})w_n]\Delta + \sqrt{\frac{\varepsilon\Delta}{\gamma^2}}\sigma(x_{n-1})B_n$$

and $w_n = v_n$, where x_{n-1}, v_n are deterministic. Then, of course,

$$\overline{V}^N(x) = \overline{V}_0^N(x). \tag{26}$$

We now proceed to obtain the Lipschitz bound in x on near-optimal $\overline{V}_0^N(x)$. First we obtain a bound on $|v_N|$ depending on $|x_{N-1}|$. Given v_N, let

$$y_N = x_{N-1} + [f(x_{N-1}) + \sigma(x_{N-1})v_N]\Delta + \sqrt{\frac{\varepsilon\Delta}{\gamma^2}}\sigma(x_{N-1})B_N, \tag{27}$$

$$y_N^0 = x_{N-1} + [f(x_{N-1})]\Delta + \sqrt{\frac{\varepsilon\Delta}{\gamma^2}}\sigma(x_{N-1})B_N. \tag{28}$$

Then by assumption (A4),

$$\begin{aligned}
&\psi(y_N) - \psi(y_N^0)\\
&\quad\le C(3/2 + 2|y_N| + 2|y_N^0|)|\sigma(x_{N-1})v_N\Delta|\\
&\quad\le C\left[(3/2 + 4K\Delta) + 2(1 + 2K\Delta)|x_{N-1}| + 2M|v_N|\Delta + 4\sqrt{\frac{\varepsilon\Delta}{\gamma^2}}M|B_N|\right]M|v_N|\Delta\\
&\quad\le C\Big[(3/2 + 4K\Delta) + 2(1 + 2K\Delta)|x_{N-1}|\Big]M|v_N|\Delta + \ 3CM^2|v_N|^2\Delta^2\\
&\qquad + 2CM^2\varepsilon\Delta|B_N|^2/\gamma^2.
\end{aligned}$$

(The weaker constants $3/2$ and 2 appearing in the first term above are being used here in place of 1 merely for reasons of consistency with later estimates; this will become clear upon a reading of Appendix A.) This implies that for $\delta\Delta$-optimal v_N,

$$\begin{aligned}
-\delta\Delta\le\ &\overline{J}_{N-1}^N(x_{N-1}, v_n) - \overline{J}_{N-1}^N(x_{N-1}, 0)\\
&\le -\left(\frac{\gamma^2}{2} - 3CM^2\Delta\right)|v_N|^2\Delta + CM\Big[(3/2 + 4K\Delta) + 2(1 + 2K\Delta)|x_{N-1}|\Big]|v_N|\Delta\\
&\quad + 2CM^2\varepsilon\Delta/\gamma^2
\end{aligned}$$

or

$$\begin{aligned}
0 \le\ &-\left(\frac{\gamma^2}{2} - 3CM^2\Delta\right)|v_N|^2 + CM\Big[(3/2 + 4K\Delta) + 2(1 + 2K\Delta)|x_{N-1}|\Big]|v_N|\\
&\quad + 2CM^2\varepsilon/\gamma^2 + \delta.
\end{aligned}$$

(One can use $a^2 + b^2 \ge 2ab$ to eliminate the cross-terms in $|v_N|, |x_{N-1}|$, leading to an additional term in the coefficient in the quadratic term, and thus leading to a stricter requirement below on γ, but that is avoidable in this proof.) Consequently, with Δ sufficiently small so that the coefficient on the quadratic term is negative, one obtains

$$|v_N| \le \frac{C}{\gamma}\left[C_1 + C_2|x_{N-1}|\right] \tag{29}$$

for proper choice of C_1, C_2, which may be chosen independently of Δ for sufficiently small Δ.

Next we propagate the Lipschitz property back one step. This is a rather technical step, and consequently we have moved the estimates to Appendix A. The resulting Lipschitz bound is

$$\begin{aligned}
&\left|\overline{V}_{N-1}^N(x_{N-1}) - \overline{V}_{N-1}^N(\hat{x}_{N-1})\right|\\
&\qquad\le b(|\hat{x}_{N-1}|)\left[1 + c_1\Delta + c_2\frac{K_4}{\gamma}\Delta + c_3\frac{K_4^2}{\gamma^2}\Delta^2\right]|x_{N-1} - \hat{x}_{N-1}|\\
&\qquad\quad + \overline{C}\sqrt{\frac{\varepsilon}{\gamma^2}}\left[K_3\Delta + M\frac{K_4}{\gamma}\Delta\right]|x_{N-1} - \hat{x}_{N-1}|,
\end{aligned} \tag{30}$$

provided that $\Delta \le \min\{1, d_1\}$, where the constants and restrictions are given in the appendix. Also, the function b is given in the appendix as

$$b(x) = \begin{cases} C(\frac{3}{2} + 2|x|^2) & \text{if } |x| \le 1, \\ C(\frac{3}{2} + 2|x|) & \text{if } |x| > 1. \end{cases} \tag{31}$$

This bound requires a one-sided bound on ψ_{xx}. In fact, the semiconvexity of ψ is sufficient to obtain this step. Consequently, the next step is to propagate the semiconvexity of ψ backward one step to obtain a semiconvexity bound on $\overline{V}^N_{N-1}$. This step also appears in Appendix A.

Next one works backwards with minor variants of (29), (30), and the semiconvexity bound. This appears in Appendix A and leads of course to a Lipschitz bound on $\overline{V}^N_0(x)$. One must then let $N \to \infty$ and show that this bound does not blow up. The corresponding sequences are given in the appendix, and the limit bound is denoted simply as $M_\infty b(|x|)$. Since there are well-known results demonstrating the convergence of such discrete-time games to their continuous-time counterparts (see, for instance, [14], [24]), one obtains a Lipschitz bound on the limit continuous-time game. In other words (reinstating the superscripts), one has obtained the Lipschitz bound

$$|\underline{V}^{\varepsilon,b,R,m}(s,x) - \underline{V}^{\varepsilon,b,R,m}(s,\hat{x})| \le M_\infty b(|x|)|x - \hat{x}| \qquad \forall\, s \in [0,T],\ |\hat{x}| \le |x| < \infty.$$

Once one obtains the above Lipschitz bound in the space variable, one may obtain the bound in the time variable without resorting to the discrete-time setting. The details are relatively long and technical, but rather standard. Consequently, they will only be sketched. Perhaps it should be noted that they rely on the bounds on the second derivatives of ℓ and ψ, of course. As a first step, let us note a result which is a slightly more specific bound on near-optimal disturbances than that of Lemma 3.2 and which is particularly oriented toward small time–horizons. It also allows for a random initial state. It is stated as a lemma below, and the proof is sketched. Following the sketched proof of the lemma, the remaining steps in the proof of the Lipschitz bound in the time variable are presented.

LEMMA 3.5. *Consider the above stochastic game, but now let the terminal time be denoted by τ, so as to differentiate this from the main problem of the section (with terminal time denoted by T), let $\tau \le 1$, and let $s \in [0,\tau]$. Also, now let the initial state, $y^\varepsilon_s = x$, be random and independent of B_t for $t \ge s$. Then there exists $\widetilde{\gamma} < \infty$ such that for all $\gamma > \widetilde{\gamma}$, for $\delta \le \tau - s$, any δ-optimal $w_\delta \in \mathcal{W}_s$ with respect to any $\theta \in \Theta_s$ satisfies*

$$\|w_\delta\|_{L_2(\Omega\times[s,\tau])} \le \overline{M}' \sqrt{1 + \mathrm{E}\big[|x|^2\big]}\sqrt{\tau - s} \tag{32}$$

for properly chosen $\overline{M}'$ (independent of x, s, τ).

Proof. The proof relies again on comparison of the near-optimal disturbance with the use of the disturbance process $w \equiv 0$. To simplify notation, we will work directly with f, ℓ, ψ, rather than with their cut-offs (f^b, ℓ^R, ψ^R); the structure of the proof is unchanged. Let $\theta \in \Theta_s$. Let y^ε satisfy (13) with disturbance w^δ, which is δ-optimal with respect to θ (for the problem over $[s,\tau]$). Using standard techniques, one shows that there exists $\widehat{C}_1 < \infty$ such that

$$\mathrm{E}\left[|y^\varepsilon_t - x|^2\right] \le \widehat{C}_1 \left[\Big(1 + \mathrm{E}\big[|x|^2\big]\Big)(t-s) + \int_s^t \mathrm{E}|w^\delta_r|^2\, dr\right] \qquad \forall\, t \in [s,\tau]. \tag{33}$$

Similarly,

$$\mathrm{E}|y_t^\varepsilon - x| \le \widehat{C}_2 \left[\Big(1 + \mathrm{E}|x|\Big)(t-s) + \int_s^t \mathrm{E}|w_r^\delta|\, dr \right] \qquad \forall\, t \in [s,\tau]. \tag{34}$$

Denote the payoff (with $u = \theta[w^\delta]$) as

$$\begin{aligned} J^\varepsilon_{g,\tau,\delta}(s,x;u,w^\delta) &\doteq \mathrm{E}\left[\int_s^\tau \ell(y_t^\varepsilon, u_t) - \frac{\gamma^2}{2}|w_t^\delta|^2\, dt + \psi(y_\tau^\varepsilon) \right] \\ &\le \mathrm{E}[\psi(x)] + \mathrm{E}[\ell(x)](\tau - s) \\ &\quad + \mathrm{E}\Bigg\{ \int_s^\tau \left[|\ell_x(x)||y_t^\varepsilon - x| + \frac{1}{2}|\ell_{xx}(\xi_t')||y_t^\varepsilon - x|^2 - \frac{\gamma^2}{2}|w_t^\delta|^2 \right] dt \\ &\qquad + |\psi_x(x)||y_\tau^\varepsilon - x| + \frac{1}{2}|\psi_{xx}(\xi_\tau')||y_\tau^\varepsilon - x|^2 \Bigg\}, \end{aligned}$$

where ξ_t' is on the line from x to y_t^ε for each t. Then, using (33), (34), (A3), and (A4) yields

$$\begin{aligned} J^\varepsilon_{g,\tau,\delta}(s,x;u,w^\delta) \le{}& \mathrm{E}[\psi(x)] + \mathrm{E}[\ell(x)](\tau - s) \\ &+ \widehat{C}_3 \left[\big(1 + \mathrm{E}\big[|x|^2\big]\big)\,(\tau - s) + \int_s^\tau \mathrm{E}\big[(1+|x|)|w_t^\delta|\big]\, dt \right] \\ &+ \widehat{C}_4 \left[\big(1 + \mathrm{E}\big[|x|^2\big]\big)\,(\tau - s) + \int_s^\tau \mathrm{E}|w_t^\delta|^2\, dt \right] - \frac{\gamma^2}{2}\int_s^\tau \mathrm{E}|w_t^\delta|^2\, dt \end{aligned} \tag{35}$$

for proper choice of $\widehat{C}_3, \widehat{C}_4$.

On the other hand, let y^0 be the solution of (13) with the same θ but with $w_0 \equiv 0$. Let the corresponding payoff be denoted by $J^\varepsilon_{g,\tau,0}(s,x;u_0,w_0)$. By similar techniques, one has

$$\mathrm{E}|y_t^0 - x|^2 \le \widehat{C}_5\big(1 + \mathrm{E}\big[|x|^2\big]\big)(t-s) \quad \text{and} \quad \mathrm{E}|y_t^0 - x| \le \widehat{C}_6\big(1 + \mathrm{E}|x|\big)(t-s)$$

for proper choice of $\widehat{C}_5, \widehat{C}_6$. This leads to

$$J^\varepsilon_{g,\tau,0}(s,x;u_0,w_0) \ge \mathrm{E}[\psi(x)] + \mathrm{E}[\ell(x)](\tau - s) - \widehat{C}_7\big(1 + \mathrm{E}\big[|x|^2\big]\big)(\tau - s) \tag{36}$$

for proper choice of $\widehat{C}_7$.

Now, using the δ-optimality of w^δ and combining (35) and (36) yields

$$\begin{aligned} 0 \le{}& \delta + \widehat{C}_8\big(1 + \mathrm{E}\big[|x|^2\big]\big)(\tau - s) + \widehat{C}_3 \int_s^\tau \mathrm{E}\big[(1+|x|)|w_t^\delta|\big]\, dt - \left(\frac{\gamma^2}{2} - \frac{\widetilde{\gamma}^2}{2}\right)\int_s^\tau \mathrm{E}|w_t^\delta|^2\, dt \\ \le{}& \delta + \widehat{C}_8\big(1 + \mathrm{E}\big[|x|^2\big]\big)(\tau - s) + \widehat{C}_3\sqrt{2(1+\mathrm{E}|x|^2)}\sqrt{\tau - s}\left[\int_s^\tau \mathrm{E}|w_t^\delta|^2\, dt\right]^{\frac{1}{2}} \\ &- \left(\frac{\gamma^2}{2} - \frac{\widetilde{\gamma}^2}{2}\right)\int_s^\tau \mathrm{E}|w_t^\delta|^2\, dt, \end{aligned}$$

where $\widehat{C}_8 \doteq \widehat{C}_3 + \widehat{C}_4 + \widehat{C}_7$ and $\frac{\widetilde{\gamma}^2}{2} \doteq \widehat{C}_4$. Letting $\delta \le \tau - s$, solving this last quadratic equation yields the desired result. ∎

Now we return to the sketch of the proof of the Lipschitz condition in the time variable. Again, to simplify matters, we work directly with f, ℓ, ψ, rather than with their cut-offs (f^b, ℓ^R, ψ^R). Consider $\underline{V}^\varepsilon(s,x)$ and $\overline{V}^\varepsilon(\hat{s},x)$ (where we are suppressing the cut-off superscripts), where $0 \le s \le \hat{s} \le T$ and $\tau \doteq \hat{s} - s$. Note that, under the given conditions, one has strong solutions of (13). Let the probability triple under which $\underline{V}^\varepsilon(s,x)$ is computed be $(\Omega, P, \mathcal{F}_\cdot)$. As in the proof of Lemma 3.1, for the computation of $\underline{V}^\varepsilon(\hat{s},x)$, one may use instead $(\Omega, P, \widehat{\mathcal{F}}_\cdot)$, where $\widehat{\mathcal{F}}_t \doteq \mathcal{F}_{t-\tau}$ for all $t \ge \hat{s} = s + \tau$; the corresponding Brownian motion is $\widehat{B}_t = B_{t-\tau}$ for all $t \ge \hat{s}$.

Let $\mathcal{U}_{[r_1,r_2]}$ ($\widehat{\mathcal{U}}_{[r_1,r_2]}$) be the set of $\mathcal{F}_t$- ($\widehat{\mathcal{F}}_t$-)progressively measurable controls over time interval $[r_1, r_2]$. Let $\widetilde{\theta}$ be $\frac{\varepsilon}{2}$-optimal for problem $\underline{V}^\varepsilon(\hat{s},x)$, and let $\widehat{\theta}$ be any extension of $\widetilde{\theta}$ to time $T+\tau$. Noting that $\widehat{\theta} : L_2[s+\tau, T+\tau] \to \widehat{\mathcal{U}}_{[s+\tau,T+\tau]}$, let $\overline{\theta} : L_2[s,T] \to \mathcal{U}_{[s,T]}$ be given by $\overline{\theta}_t[w] = \widehat{\theta}_{t+\tau}[\widetilde{w}]$, where $\widetilde{w}_{t+\tau} = w_t$. Then let w be $\frac{\varepsilon}{2}$-optimal for $\underline{V}^\varepsilon(s,x)$ corresponding to $\overline{\theta}$. Let $\widehat{w}$ be given by $\widehat{w}_t = w_{t-\tau}$ for all $t \ge \hat{s}$. Let y^ε be the solution to (13) with the above w and $\overline{u} \doteq \overline{\theta}[w]$. Let $\widehat{y}^\varepsilon$ be the solution to (13) with new initial condition $\widehat{y}^\varepsilon_{\hat{s}} = x$ and the above $\widehat{w}$ and $\widehat{u} \doteq \widehat{\theta}[\widehat{w}]$. Then

$$\widehat{y}^\varepsilon_t = y^\varepsilon_{t-\tau} \qquad \forall\, t \in [\hat{s}, T].$$

Note that

$$\underline{V}^\varepsilon(s,x) - \underline{V}^\varepsilon(\hat{s},x) \le J^\varepsilon_g(s,x,u,w) - J^\varepsilon_g(\hat{s},x,\widehat{u},\widehat{w}) + \varepsilon,$$

and, by the time shift, this is equal to

$$\mathrm{E}\left[\int_{T-\tau}^{T} \ell(y^\varepsilon_t, \overline{u}_t) - \frac{\gamma^2}{2}|w_t|^2\, dt + \psi(y^\varepsilon_T) - \psi(\widehat{y}^\varepsilon_{T-\tau})\right] + \varepsilon.$$

Letting $\varepsilon \downarrow 0$, one must show that this is bounded above by $\overline{C}_{T,D}\tau$ for any $|x| \le D < \infty$ (for proper choice of $\overline{C}_{T,D}$). The bound from below is similar, and thus is not discussed here. Also, since the bound on the integral term is somewhat simpler than the bound on the terminal cost difference, we consider only a bound on $\mathrm{E}[\psi(y^\varepsilon_T) - \psi(\widehat{y}^\varepsilon_{T-\tau})]$ of the desired form here. Note that by (A4),

$$\mathrm{E}\left[\psi(y^\varepsilon_T)\right] - \mathrm{E}\left[\psi(\widehat{y}^\varepsilon_{T-\tau})\right] \le \mathrm{E}\left[\psi_x(y^\varepsilon_{T-\tau})[y^\varepsilon_T - y^\varepsilon_{T-\tau}]\right] + \frac{1}{2}\overline{C}\mathrm{E}\left[|y^\varepsilon_T - y^\varepsilon_{T-\tau}|^2\right]. \tag{37}$$

Also, from Lemma 3.5 and the choice of w,

$$\|w\|_{L_2(\Omega\times[T-\tau,T])} \le \overline{M}'\sqrt{1 + \mathrm{E}\left[|y^\varepsilon_{T-\tau}|^2\right]}\sqrt{\tau} \tag{38}$$

for ε sufficiently small. Using (37), using techniques similar to those in the proof of Lemma 3.5, and employing (38) yields the desired bound. This completes the sketch of the proof of the Lipschitz bound in the time variable. □

LEMMA 3.6. *Given any sequences $\{b_n\}, \{m_n\}$ such that $b_n, m_n \to \infty$, there exists a subsequence which we also subscript by n (to reduce notational complexity) such that*

$$\underline{V}^{\varepsilon, b_n, R, m_n} \to \widetilde{V}^{\varepsilon, R}$$

uniformly on compact sets, where $\widetilde{V}^{\varepsilon,R} \in C^{1,2}$ (and the derivatives converge uniformly on compact sets as well). Further, $\widetilde{V}^{\varepsilon,R}$ is a solution to

$$0 = V_s + \frac{\varepsilon}{2\gamma^2}\sum_{i,j=1}^{n} a_{i,j}(x)V_{x_ix_j} + \min_{u\in U}\max_{w\in\Re^m}\left\{[f(x,u)+\sigma(x)w]\nabla V + \ell^R(x,u) - \frac{\gamma^2}{2}|w|^2\right\},$$

$$V(T,x) = \psi^R(x). \tag{39}$$

Proof. By Lemmas 3.1, 3.2, and 3.4, we have that $\underline{V}^{\varepsilon,b,R,m}, \underline{V}_s^{\varepsilon,b_n,R,m_n}, \nabla\underline{V}^{\varepsilon,b_n,R,m_n}$ are all uniformly bounded on compact sets. Combining this with (23) implies that $\sum_{i,j=1}^n a_{i,j}\cdot\underline{V}_{x_i,x_j}^{\varepsilon,b_n,R,m_n}$ are also bounded uniformly on compact sets. Then following the method in [21, Appendix E] (see also [43]), one obtains the result. □

Recall process (1) of section 2. Define the modified risk-sensitive cost and value,

$$J^{\varepsilon,R}(s,x,u.) = \mathrm{E}\exp\left\{\frac{1}{\varepsilon}\left[\int_s^T \ell^R(y_t^\varepsilon,u_t)\,dt + \psi^R(y_T^\varepsilon)\right]\right\} \tag{40}$$

and

$$V^{\varepsilon,R}(s,x) = \inf_{u\in\mathcal{U}_s}\varepsilon\log J^{\varepsilon,R}(s,x,u.). \tag{41}$$

THEOREM 3.7. *Let $0\le s\le T$ and $x\in\Re^n$. For any $u\in\mathcal{U}_s$, one has*

$$\varepsilon\log J^{\varepsilon,R}(s,x,u) \ge \widetilde{V}^{\varepsilon,R}(s,x),$$

while with optimal $\widetilde{u}^R$ given by (6), with $\widetilde{V}^{\varepsilon,R}, \ell^R$ replacing $\widetilde{V}^\varepsilon, \ell$, one has

$$\varepsilon\log J^{\varepsilon,R}(s,x,\widetilde{u}^R) = \widetilde{V}^{\varepsilon,R}(s,x).$$

Proof. By $\widetilde{V}^{\varepsilon,R}\in C^{1,2}$, assumptions (A3.iv), (A4.iv), and (40), one easily shows that there exists $C_R<\infty$ such that

$$|\nabla\widetilde{V}^{\varepsilon,R}(s,x)|\le C_R \qquad \forall\, s\in[0,T], x\in\Re^n.$$

In this case, (7) holds for any $u\in\mathcal{U}_s$. Consequently, the proof of Lemma 2.2 holds for $\widetilde{V}^{\varepsilon,R}$ for any control $u\in\mathcal{U}_s$, and so

$$\varepsilon\log J^{\varepsilon,R}(s,x,u) \ge \widetilde{V}^{\varepsilon,R}(s,x)$$

for any $u\in\mathcal{U}_s$. Lemma 2.1 also holds with $\widetilde{V}^{\varepsilon,R}, J^{\varepsilon,R}$ replacing $\widetilde{V}^\varepsilon, J^\varepsilon$. □

COROLLARY 3.8.

$$\underline{V}^{\varepsilon,b,R,m} \to \widetilde{V}^{\varepsilon,R}$$

uniformly on compact sets (not just sequentially).

Proof. Combine the uniqueness implied by Theorem 3.7 with the subsequential convergence of Lemma 3.6. □

COROLLARY 3.9. *$\widetilde{V}^{\varepsilon,R}$ is monotonically increasing as a function of R and is bounded above by*

$$\varepsilon\log J^\varepsilon(s,x,\widetilde{u}) \le \widetilde{V}^\varepsilon(s,x).$$

Proof. The first assertion follows by the definitions of ℓ^R, ψ^R. For the second assertion, note that by Theorem 3.7

$$\begin{aligned}\widetilde{V}^{\varepsilon,R}(s,x) &\le \varepsilon\log J^{\varepsilon,R}(s,x,\widetilde{u})\\ &\le \varepsilon\log J^\varepsilon(s,x,\widetilde{u}),\end{aligned}$$

which by Lemma 2.1

$$\leq \widetilde{V}^{\varepsilon}(s,x). \qquad \square$$

We now obtain the promised existence result for (4).

THEOREM 3.10. *$\widetilde{V}^{\varepsilon,R}(s,x) \uparrow \widetilde{V}^{\varepsilon}(s,x) \in C^{1,2}$ as $R \to \infty$, where $\widetilde{V}^{\varepsilon}$ is a solution to (4).*

Proof. The result follows exactly as in Lemma 3.6, with the added monotonicity from Corollary 3.9. $\square$

As an added bonus, we obtain the promised improvement over Lemma 2.3.

THEOREM 3.11. *There exists a solution $\widetilde{V}^{\varepsilon} \in C^{1,2}$ of (4). Further, for any $0 \leq s \leq T$, $x \in \Re^n$, and $u \in \mathcal{U}_s$, one has*

$$\varepsilon \log J^{\varepsilon}(s,x,u) \geq \widetilde{V}^{\varepsilon}(s,x),$$

while with optimal $\widetilde{u}$ given by (6) one has

$$\varepsilon \log J^{\varepsilon}(s,x,\widetilde{u}) = \widetilde{V}^{\varepsilon}(s,x).$$

Proof. Note that for any $u \in \mathcal{U}_s$,

$$\varepsilon \log J^{\varepsilon}(s,x,u) \geq \varepsilon \log J^{\varepsilon,R}(s,x,u),$$

which by Theorem 3.7

$$\geq \widetilde{V}^{\varepsilon,R}(s,x).$$

Combining this with Theorem 3.10 yields the first assertion. The second assertion follows by combining the first assertion with Lemma 2.1. $\square$

4. Uniqueness and risk-sensitive limit results. In this section we consider the following terminal value problem:

$$\begin{cases} -W_t(t,x) + H(x, D_x W(t,x)) = 0 & \text{in } (0,T) \times \Re^n, \\ W(T,x) = \psi(x) & \text{in } \Re^n, \end{cases} \tag{42}$$

where the Hamiltonian H is

$$\begin{aligned} H(x,p) &:= \max_{u \in U} \min_{\omega \in \Re^m} \left\{ -(f(x,u) + \sigma(x)\omega) \cdot p - \ell(x,u) + \frac{\gamma^2}{2}|\omega|^2 \right\} \\ &= \min_{\omega \in \Re^m} \left\{ \frac{\gamma^2}{2}|\omega|^2 - \sigma(x)\omega \cdot p \right\} + \max_{u \in U} \{ -f(x,u) \cdot p - \ell(x,u) \}. \end{aligned} \tag{43}$$

This is the DPE corresponding to a robust/H_∞ control problem. This problem can be considered a *differential game* with the following cost functional:

$$J(s,x,u,\omega) := \int_s^T \ell(y_x(t),u(t)) - \frac{\gamma^2}{2}|\omega(t)|^2 dt + \psi(y_x(T)), \tag{44}$$

where $u(\cdot) \in \mathcal{U} := \{\text{measurable functions } [0,T] \to U\}$ is the control of the minimizing player, $\omega(\cdot) \in \mathcal{B} := L^2([0,T], \Re^m)$ is the control of the maximizing player, and $y_x(\cdot)$ is the unique solution of the following dynamical system:

$$\begin{cases} y'(t) = f(y(t),u(t)) + \sigma(y(t))\omega(t), \\ y(s) = x. \end{cases} \tag{S}$$

Note that we switch notation for the time argument here from y_t to $y(t)$, to emphasize that the y paths are now (deterministic) solutions of ODEs rather than (stochastic) solutions of SDEs.

We define the set of nonanticipating strategies for the minimizing player to be

$$\Theta := \{\theta : \mathcal{B} \to \mathcal{U} : \text{ given any } \tau \in [0,T],\ \omega(t) = \overline{\omega}(t) \\ \text{for all } t \in [0,\tau] \text{ implies } \theta[\omega](t) = \theta[\overline{\omega}](t) \text{ for all } t \in [0,\tau]\},$$

and the set of nonanticipating strategies of the maximizing player to be

$$\Lambda := \{\lambda : \mathcal{U} \to \mathcal{B} : \text{ given any } \tau \in [0,T],\ u(t) = \overline{u}(t) \\ \text{for all } t \in [0,\tau] \text{ implies } \lambda[u](t) = \lambda[\overline{u}](t) \text{ for all } t \in [0,\tau]\}.$$

The *lower value* and the *upper value* of the game are given by (see, for instance, [11])

$$V(s,x) := \inf_{\theta \in \Theta} \sup_{\omega \in \mathcal{B}} J(s,x,\theta[\omega],\omega), \tag{45}$$

$$W(s,x) := \sup_{\lambda \in \Lambda} \inf_{u \in \mathcal{U}} J(s,x,u,\lambda[u]). \tag{46}$$

This section will be concerned with the *lower value*; analogous results hold for the *upper value* as well.

In the following theorem we prove a uniqueness result for (42) in the set of nonnegative continuous functions with locally bounded superdifferential in the x variable and growing at most quadratically. More precisely, given $U : [0,T] \times \Re^n \to \Re$, we define

$$D_x^+ U(t,x) = \left\{ p \in \Re^n : \ \limsup_{|x-y| \to 0} \frac{U(t,y) - U(t,x) - (y-x)\cdot p}{|x-y|} \le 0 \right\},$$

$$||U||_R := \sup\{|U(t,x)| + |p|,\ (t,x) \in [0,T] \times \overline{B}_R(0)\,,\ p \in D_x^+ U(t,x)\},$$

and

$$\mathcal{K} := \Bigg\{ U \in C([0,T] \times \Re^n) : \ U(t,x) \ge 0,\ ||U||_R < +\infty \qquad \forall\, R > 0 \\ \text{and} \quad \sup_{(t,x) \in [0,T] \times \Re^n} \frac{|U(t,x)|}{1+|x|^2} < +\infty \Bigg\}.$$

For the uniqueness result, we use, besides some hypotheses stated in section 2, the following less restrictive assumptions:

(H1) $f \in C(\Re^n \times U, \Re^n)$, $\sigma \in C(\Re^n, \mathcal{L}(\Re^m, \Re^n))$, $\ell \in C(\Re^n \times U, \Re)$, $\psi \in C(\Re^n, \Re)$;

(H2) there exists $M > 0$ such that

$$|\sigma(x)| \le M(1+|x|) \quad \forall x \in \Re; \tag{47}$$

(H3) for any $R > 0$ there exists $L_R > 0$ such that

$$|f(x,u) - f(y,u)| \le L_R |x-y| \quad \forall x,y \in \overline{B}_R(0),\ u \in U,$$
$$|\sigma(x) - \sigma(y)| \le L_R |x-y| \quad \forall x,y \in \overline{B}_R(0).$$

THEOREM 4.1. *Assume* (A0), (A1.iii), (A3.ii)–(A3.iv), (A4.ii)–(A4.iv), *and* (H1)–(H3). *If* $W_1, W_2 \in \mathcal{K}$ *are viscosity solutions of* (42), *then* $W_1 = W_2$ *in* $[0,T] \times \Re^n$.

Proof. We follow some arguments used in the proof of Theorem 3.1 in [3] for convex Hamiltonians. Let $W_1, W_2 \in \mathcal{K}$ be viscosity solutions of (42). We first note that if (x,p) belongs to a compact $K \subseteq \Re^{2n}$, then there exists $\widetilde{R} > 0$, depending on K, such that

$$H(x,p) = H_{\widetilde{R}}(x,p) := \min_{|\omega| \le \widetilde{R}} \left\{ \frac{\gamma^2}{2}|\omega|^2 - \sigma(x)\omega \cdot p \right\} + \max_{u \in U} \{-f(x,u) \cdot p - \ell(x,u)\}. \tag{48}$$

In fact, it can be easily verified that

$$\lim_{|\omega| \to \infty} \inf_{(x,p) \in K} \left\{ \frac{\gamma^2}{2}|\omega|^2 - \sigma(x)\omega \cdot p \right\} = +\infty.$$

Thus, since $\gamma^2|0|^2 - \sigma(x)0 \cdot p = 0$, the min in (43) can be computed in a compact subset of $\Re^m$. Now fix $r > 0$ and choose $R > r$. First we suppose that

$$T \le \delta(\gamma, r, R) := \min\left(1, \frac{R-r}{K(1+R)}, \frac{\gamma^2(R-r)^2}{2M^2(1+R)^2\overline{C}(1+R^2) + 2\gamma^2 K(R-r)(1+R)}\right), \tag{49}$$

where $\overline{C} = \max(C, \sup_{(t,x)\in[0,T]\times\Re^n} \frac{|W_1(t,x)|}{1+|x|^2}, \sup_{(t,x)\in[0,T]\times\Re^n} \frac{|W_2(t,x)|}{1+|x|^2})$ and C is the growth constant of ℓ and ψ.

Since W_1, $W_2 \in \mathcal{K}$ and (48) holds, there exists $\widetilde{R} > 0$ such that

$$H(x,p) = H_{\widetilde{R}}(x,p) \qquad \forall\, p \in D_x^+ W_1(t,x) \cup D_x^+ W_2(t,x),\ (t,x) \in [0,T] \times \overline{B}_R(0).$$

Thus W_i $(i = 1,2)$ is a viscosity solution also of

$$\begin{cases} -W_t(t,x) + H_{\widetilde{R}}(x, D_x W(t,x)) = 0 & \text{in } (0,T) \times B_R(0), \\ W(x,T) = \psi(x) & \text{in } B_R(0). \end{cases} \tag{50}$$

Under the current assumptions, the continuous viscosity solutions W of (50) satisfy the so-called *optimality principle* (see, e.g., Propositions 2.1 and 2.2 in [34]); namely, for all $s \in [0,T]$ and $0 \le \rho \le T - s$, the following estimate holds:

$$W(s,x) = \inf_{\theta \in \Theta} \sup_{\substack{\omega \in \mathcal{B} \\ |\omega| \le \widetilde{R}}} \Big\{ \mathcal{I}(s, (s+\rho) \wedge t_x^R(\theta[\omega], \omega)) + W((s+\rho) \wedge t_x^R(\theta[\omega],\omega), y_x((s+\rho) \wedge t_x^R(\theta[\omega],\omega))) \Big\}, \tag{51}$$

where for all $s, t \in [0,T]$, $\mathcal{I}(s,t) := \int_s^t \ell(y(\tau), \theta[\omega](\tau)) - \frac{\gamma^2}{2}|\omega(\tau)|^2 d\tau$, $y_x(t) = y_x(t, \theta[\omega], \omega)$ is the unique solution of (S) corresponding to $\theta[\omega]$, ω (we drop the dependence on $\theta[\omega]$, ω for simplicity of notations), and $t_x^R(\theta[\omega],\omega)$ is the first exit time from $B_R(0)$, i.e.,

$$t_x^R(\theta[\omega],\omega) := \inf\{t \ge s : |y_x(t)| \ge R\}.$$

In particular, (51) is satisfied by W_i $(i = 1,2)$. We observe that, since $W_i \ge 0$, for any strategy θ the sup in (51), corresponding to $\rho = T - s$, may be confined to the controls

ω such that $\mathcal{I}(s, T \wedge t_x^R(\theta[\omega], \omega)) + W_i((T \wedge t_x^R(\theta[\omega], \omega)), y(T \wedge t_x^R(\theta[\omega], \omega))) \geq 0$. For any strategy θ, we define the set

$$\mathcal{A}_\theta := \left\{\omega \in \mathcal{B} :\ \mathcal{I}(s, T \wedge t_x^R(\theta[\omega], \omega)) + W_i((T \wedge t_x^R(\theta[\omega], \omega)), y_x(T \wedge t_x^R(\theta[\omega], \omega))) \geq 0\right\}.$$

We observe that $\mathcal{A}_\theta \neq \emptyset$, since it contains the control $\omega \equiv 0$. We claim that if $t_x^R(\theta[\omega], \omega)) < T$, then $\omega \notin \mathcal{A}_\theta$. In fact, suppose that $y_x(\bar{t}) \in \partial B_R(0)$, for some $\bar{t} \in [s, T)$, and $y_x(t) \in B_R(0)$ for all $t \in [s, \bar{t})$. Then we have the following estimate:

$$\begin{aligned} R - |x| = |y_x(\bar{t})| - |x| &\leq \int_s^{\bar{t}} K(1 + |y_x(\tau)|)d\tau + \int_s^{\bar{t}} M(1 + |y_x(\tau)|)\omega(\tau)|d\tau \\ &\leq K(1 + R)(\bar{t} - s) + M(1 + R)||\omega||_{L^2(s,\bar{t})}(\bar{t} - s)^{1/2}. \end{aligned}$$

Thus $||\omega||_{L^2(s,\bar{t})} \geq \chi[r, R](\bar{t})$, where

$$\chi[r, R](\bar{t}) := \frac{R - r - K(1 + R)(\bar{t} - s)}{M(1 + R)(\bar{t} - s)^{1/2}}.$$

We observe that, by the assumption of (49), $\chi[r, R](t)$ is positive for all $t \in [s, T]$. Hence we have

$$\begin{aligned} \mathcal{I}(s, \bar{t}) + W_i(\bar{t}, y_x(\bar{t})) &\leq \int_s^{\bar{t}} C(1 + |y_x(t)|^2) - \frac{\gamma^2}{2}|\omega(t)|^2\ dt + \widetilde{C}(1 + |y_x(\bar{t})|^2) \\ &\leq C(1 + R^2)(\bar{t} - s) - \frac{\gamma^2}{2}|\chi[r, R](\bar{t})|^2 + \widetilde{C}(1 + R^2) \\ &\leq \overline{C}(1 + R^2) + \gamma^2 \frac{K(R - r)(1 + R)}{M^2(1 + R)^2} - \frac{\gamma^2}{2}\frac{(R - r)^2}{M^2(1 + R)^2(\bar{t} - s)} \\ &\quad - \frac{\gamma^2}{2}\frac{K^2(\bar{t} - s)}{M^2} \\ &\leq \overline{C}(1 + R^2) + \gamma^2 \frac{K(R - r)}{M^2(1 + R)} - \frac{\gamma^2}{2}\frac{(R - r)^2}{M^2(1 + R)^2(T - s)}, \end{aligned}$$

where $\widetilde{C} := (\sup_{(t,x)\in[0,T]\times\Re^n} \frac{|W_1(t,x)|}{1+|x|^2}, \sup_{(t,x)\in[0,T]\times\Re^n} \frac{|W_2(t,x)|}{1+|x|^2})$.

By the condition (49), we have $\mathcal{I}(s, \bar{t}) + W(\bar{t}, y_x(\bar{t})) < 0$, and this proves the claim. Therefore for all $(s, x) \in [0, T] \times \overline{B}_r(0)$, each W_i $(i = 1, 2)$ satisfies

$$W_i(s, x) = \inf_{\theta\in\Theta} \sup_{\substack{\omega\in\mathcal{A}_\theta \\ |\omega|\leq\widetilde{R}}} \left\{\mathcal{I}(s, T) + \psi(y_x(T))\right\}, \tag{52}$$

and we can conclude that $W_1 = W_2$ in $[0, T] \times \overline{B}_r(0)$. In the case of $T > \delta(\gamma, r, R)$, we can divide the interval $[0, T]$ into subintervals whose length is less than $\delta(\gamma, r, R)$. Let $0 = t_0, t_1, \ldots, t_n = T$ be the points of such a division, and for any $k = 1, \ldots, n$ let us consider the following Cauchy problem:

$$\begin{cases} -W_t(t, x) + H_{\widetilde{R}}(x, D_x W(t, x)) = 0 & \text{in } (t_{k-1}, t_k) \times B_R(0), \\ W(x, t_k) = \psi_k(x) & \text{in } B_R(0), \end{cases} \tag{53}$$

where the terminal value $\psi_k(x)$ can coincide either with $W_1(t_k, x)$ or with $W_2(t_k, x)$. We start with $k = n$. Since $|t_n - t_{n-1}| < \delta(\gamma, r, R)$, we can argue as above and obtain

that the W_i's coincide in $[t_{n-1}, t_n] \times \overline{B}_r(0)$. Then, by proceeding backward in the time variable, we obtain that for all $k \in \{0, \dots, n\}$, $W_1 = W_2$ in $[t_{k-1}, t_k] \times \overline{B}_r(0)$, and this completes the proof. □

In [41] it is proved that, under the hypotheses of Theorem 4.1 (with (A2.ii) instead of (H2)), for sufficiently large γ the *lower value* V is a locally Lipschitz continuous function growing at most quadratically in the state variable (thus $V \in \mathcal{K}$), and it is a viscosity solution of (42). Thus we have the following corollary.

COROLLARY 4.2. *Assume the hypotheses of Theorem* 4.1*, with* (H2) *replaced by* (A2.ii). *Then for γ large enough the value V is the unique viscosity solution of* (42) *in $\mathcal{K}$.*

As a consequence of the uniqueness Theorem 4.1 we get the convergence of the value function $\widetilde{V}^\varepsilon$ of the risk-sensitive problem to the value V of the robust/H_∞ control problem as $\varepsilon \to 0$. The proof of this result is standard (see, e.g., [2]), and it follows directly from the stability of the viscosity solutions with respect to the locally uniform convergence of the DPE for the risk-sensitive problem to the DPE for the corresponding robust control problem and the local uniform boundedness and equicontinuity estimates obtained in section 3.

THEOREM 4.3. *Assume* (A0)–(A5). *Then, for γ large enough, $\widetilde{V}^\varepsilon \to V$ uniformly on the compact subsets of $[0, T] \times \Re^n$.*

Proof. From Lemmas 3.2, 3.4, 3.6 and Theorem 3.10 it follows that, for γ large enough and for any $D > 0$, there exist $C_{T,D}$, $M_{T,D} > 0$ (independent of ε) such that

$$|\widetilde{V}^\varepsilon(s_1, x) - \widetilde{V}^\varepsilon(s_2, z)| \le C_{T,D}[|x - z| + |s_1 - s_2|] \quad \forall\, s_1, s_2 \in [0, T],\ x, z \in \overline{B}_D(0) \tag{54}$$

and

$$|\widetilde{V}^\varepsilon(t, x)| \le M_{T,D} \qquad \forall\, t \in [0, T],\ x \in \overline{B}_D(0). \tag{55}$$

By using the Ascoli–Arzelà theorem and a standard *diagonal procedure* (see, e.g., [2]), we get the existence of a subsequence of $\widetilde{V}^\varepsilon$ converging uniformly on the compact subsets of $[0, T] \times \Re^n$ to a function $U \in \mathcal{K}$. From the stability of viscosity solutions with respect to the locally uniform convergence of the DPE (4) for the risk-sensitive problem to the DPE (42) for the robust control problem as $\varepsilon \to 0$, it follows that U is a viscosity solution of (42), and the uniqueness Theorem 4.1 implies that $U = V$. Thus we can conclude that the entire sequence $\widetilde{V}^\varepsilon$ converges to V as $\varepsilon \to 0$, uniformly on compact subsets of $[0, T] \times \Re^n$. □

Appendix A. In this appendix, we first propagate the Lipschitz bound for $\overline{V}^N$ back one step from the terminal time; that is, we obtain the Lipschitz bound for $\overline{V}^N_{N-1}$, given the corresponding bound for $\overline{V}^N_N = \psi$. We will then continue this procedure back to the initial time.

Let v^δ_N be δ-optimal at x_{N-1}. Then

$$\overline{V}^N_{N-1}(x_{N-1}) - \overline{V}^N_{N-1}(\hat{x}_{N-1}) \le \mathrm{E}\big[\psi(y_N) - \psi(\hat{y}_N)\big] + C(1 + |x_{N-1}| + |\hat{x}_{N-1}|) \cdot |x_{N-1} - \hat{x}_{N-1}|\Delta + \delta, \tag{56}$$

where y_N is given by (27) with disturbance v^δ_N, and

$$\hat{y}_N = \hat{x}_{N-1} + [f(\hat{x}_{N-1}) + \sigma(\hat{x}_{N-1})v^\delta_N]\Delta + \sqrt{\frac{\varepsilon\Delta}{\gamma^2}}\sigma(\hat{x}_{N-1})B_N. \tag{57}$$

The following rather technical estimates will lead to a Lipschitz bound on the $\mathrm{E}[\psi(y_N) - \psi(\hat{y}_N)]$ term in (56). In particular, a one-sided Lipschitz bound is obtained using a one-sided bound on ψ_{xx}. Symmetry then leads to the usual (two-sided) Lipschitz bound. This approach is employed due to the semiconvexity-preserving nature of the maximizing control problem, which has previously been applied, for instance, in [20].

First, we will replace the bound on ψ_x so as to smooth it near the origin. Note that by Assumption (A4), $|\psi_x(x)| \leq C(1+2|x|)$ for all x. A quick computation shows that $C(1+2|x|) \leq C(\frac{3}{2} + 2|x|^2)$ for all x. Consequently, we have

$$\psi_x(x) \leq b(x) \qquad \forall\, x \in \Re^n,$$

where

$$b(x) = \begin{cases} C(\frac{3}{2} + 2|x|^2) & \text{if } |x| \leq 1, \\ C(\frac{3}{2} + 2|x|) & \text{if } |x| > 1. \end{cases} \tag{58}$$

Now note that, by the mean value theorem,

$$\psi(y_N) - \psi(\hat{y}_N) = \psi_x(\xi)(y_N - \hat{y}_N),$$

where

$$\xi \doteq \lambda y_N + (1-\lambda)\hat{y}_N$$

and λ is a random variable with range $[0,1]$. Use the notation $(\delta f) \doteq f(x_{N-1}) - f(\hat{x}_{N-1})$ and $(\delta\sigma) \doteq \sigma(x_{N-1}) - \sigma(\hat{x}_{N-1})$, so that

$$y_N - \hat{y}_N = (x_{N-1} - \hat{x}_{N-1}) + \big((\delta f) + (\delta\sigma)v_N^\delta\big)\,\Delta + \sqrt{\frac{\varepsilon\Delta}{\gamma^2}}(\delta\sigma)B_N.$$

(The δ in the notation (δf) is not intended to be confused with the δ superscript in v_N^δ; it is merely intended as a shorthand to reduce the size of the displayed equations.) Consequently, we have

$$\begin{aligned} \mathrm{E}[\psi(y_N) - \psi(\hat{y}_N)] = {} & \mathrm{E}[\psi_x(\xi)]\left[(x_{N-1} - \hat{x}_{N-1}) + ((\delta f) + (\delta\sigma)v_N^\delta)\Delta\right] \\ & + \sqrt{\frac{\varepsilon\Delta}{\gamma^2}}\mathrm{E}[\psi_x(\xi)(\delta\sigma)B_N]. \end{aligned} \tag{59}$$

We will work with each of the two terms on the right-hand side of (59) separately. The second is easier, and we begin with that.

As mentioned in section 3, we let each component of B_N, B_{Nj} be independent, with

$$P\Big(B_{Nj} = 1\Big) = \frac{1}{2}, \qquad P\Big(B_{Nj} = -1\Big) = \frac{1}{2} \qquad \text{for } j = 1, \ldots, m.$$

Then, by the definition of y_N, (A1), (A2), and (29), there exists $K_1 < \infty$ such that

$$|y_N - x_{N-1}| \leq K(1 + |x_{N-1}|)\Delta + K_1\sqrt{\Delta} + MD_v(x_{N-1})\Delta \qquad \forall\, \omega \in \Omega, \tag{60}$$

where

$$K_1 = \sqrt{\frac{\varepsilon m}{\gamma^2}}M$$

and, as a shorthand,

$$D_v(x_{N-1}) = \frac{C}{\gamma}\left[C_1 + C_2|x_{N-1}|\right].$$

Similarly,

$$|\hat{y}_N - \hat{x}_{N-1}| \le K(1+|\hat{x}_{N-1}|)\Delta + K_1\sqrt{\Delta} + MD_v(\hat{x}_{N-1})\Delta \qquad \forall\,\omega\in\Omega. \tag{61}$$

Let us take

$$|x_{N-1} - \hat{x}_{N-1}| \le \Delta \le 1, \tag{62}$$

in order to simplify matters below. (The bound on $|x_{N-1} - \hat{x}_{N-1}|$ is irrelevant to the Lipschitz bound; the bound on Δ is also not important, since we will be taking $\Delta = T/N$ to zero.) Also, without loss of generality, let us suppose $|\hat{x}_{N-1}| \le |x_{N-1}|$. By the definition of ξ,

$$|\xi - x_{N-1}| \le \max\left\{|y_N - x_{N-1}|,\, |\hat{y}_N - \hat{x}_{N-1}| + |\hat{x}_{N-1} - x_{N-1}|\right\},$$

which by (62)

$$\le \max\left\{|y_N - x_{N-1}|,\, |\hat{y}_N - \hat{x}_{N-1}| + \Delta\right\}. \tag{63}$$

Then by (60), (61), and (63),

$$|\xi - x_{N-1}| \le [K(2+|x_{N-1}|) + K_1]\sqrt{\Delta} + MD_v(x_{N-1})\Delta \qquad \forall\,\omega\in\Omega. \tag{64}$$

Now,

$$\begin{aligned}
&\mathrm{E}\left[\psi_x(\xi)(\delta\sigma)B_N\right]\\
&\quad= \mathrm{E}\left[\psi_x(x_{N-1})(\delta\sigma)B_N\right] + \mathrm{E}\left[(\psi_x(\xi) - \psi_x(x_{N-1}))(\delta\sigma)B_N\right]\\
&\quad= \mathrm{E}\left[(\psi_x(\xi) - \psi_x(x_{N-1}))(\delta\sigma)B_N\right],
\end{aligned}$$

which for proper choice of ζ on the line from x_{N-1} to ξ

$$= \mathrm{E}\left[\psi_{xx}(\zeta)(\xi - x_{N-1})(\delta\sigma)B_N\right],$$

which by (A2) and (A4)

$$\ge -\overline{C}L_\sigma \mathrm{E}\left\{|\xi - x_{N-1}|\frac{|x_{N-1} - \hat{x}_{N-1}|}{1+|\hat{x}_{N-1}|}|B_N|\right\},$$

which by (64) and (62)

$$\ge -\overline{C}\left[K_3 + M\frac{K_4}{\gamma}\right]|x_{N-1} - \hat{x}_{N-1}|\sqrt{\Delta} \qquad \forall\,\omega\in\Omega, \tag{65}$$

where

$$K_3 \doteq L_\sigma \max_{z\in\Re^n}\left\{\frac{K(3+|z|)+K_1}{1+|z|}\right\} \quad \text{and} \quad K_4 \doteq CL_\sigma \max_{z\in\Re^n}\left\{\frac{C_1 + C_2(|z|+1)}{1+|z|}\right\}.$$

Employing (65) in (59) yields

$$\begin{aligned}\mathrm{E}[\psi(y_N)-\psi(\hat{y}_N)] \geq\ & \mathrm{E}[\psi_x(\xi)]\left[(x_{N-1}-\hat{x}_{N-1})+((\delta f)+(\delta\sigma)v_N^\delta)\Delta\right] \\ & -\overline{C}\sqrt{\frac{\varepsilon}{\gamma^2}}\left[K_3+M\frac{K_4}{\gamma}\right]\Delta|x_{N-1}-\hat{x}_{N-1}|. \end{aligned} \tag{66}$$

We now turn to the first term. Employing (A1), (A2), and (29) in (66) yields

$$\begin{aligned}\mathrm{E}[\psi(y_N)-\psi(\hat{y}_N)] \geq\ & -\ |\mathrm{E}[\psi_x(\xi)]|\left[1+\left(K+L_\sigma M\frac{K_4}{\gamma}\right)\Delta\right]|x_{N-1}-\hat{x}_{N-1}| \\ & -\overline{C}\sqrt{\frac{\varepsilon}{\gamma^2}}\left[K_3+M\frac{K_4}{\gamma}\right]\Delta|x_{N-1}-\hat{x}_{N-1}|. \end{aligned} \tag{67}$$

A rather long argument will now be used to obtain a bound on

$$|\mathrm{E}[\psi_x(\xi)]|\,.$$

This will involve the use of the bound b from (58). Recalling the form of b, we see that this will involve bounds on $\mathrm{E}|y_N|$, $\mathrm{E}|\hat{y}_N|$, $\mathrm{E}|y_N|^2$, and $\mathrm{E}|\hat{y}_N|^2$.

We work first with $\mathrm{E}|y_N|$ (and $\mathrm{E}|\hat{y}_N|$). As a shorthand, let

$$z_{x_{N-1}} \doteq x_{N-1}+(f(x_{N-1})+\sigma(x_{N-1})v_N^\delta)\Delta.$$

Then

$$\begin{aligned}\mathrm{E}|y_N| &= \mathrm{E}\left|z_{x_{N-1}}+\sqrt{\frac{\varepsilon}{\gamma^2}}\sigma(x)\sqrt{\Delta}B_N\right| \\ &= \frac{1}{2^m}\sum_{j_1=1}^{2}\sum_{j_2=1}^{2}\cdots\sum_{j_m=1}^{2}\left|z_{x_{N-1}}+\sqrt{\frac{\varepsilon}{\gamma^2}}\sqrt{\Delta}\sigma(x)\left(-1^{j_1},-1^{j_2},\ldots,-1^{j_m}\right)^T\right|. \end{aligned} \tag{68}$$

Define G, $\overline{G}$ by

$$\begin{aligned} G(z_{x_{N-1}},\Delta) &\doteq \max_{u\in\Re^m,\,|u|=1}\frac{1}{2}\left[\left|z_{x_{N-1}}+\sqrt{\frac{\varepsilon}{\gamma^2}}\sqrt{\Delta}\sigma(x)u\right|+\left|z_{x_{N-1}}-\sqrt{\frac{\varepsilon}{\gamma^2}}\sqrt{\Delta}\sigma(x)u\right|\right] \\ &\leq \max_{|v|\leq M}\frac{1}{2}\left[\left|z_{x_{N-1}}+\sqrt{\frac{\varepsilon}{\gamma^2}}\sqrt{\Delta}v\right|+\left|z_{x_{N-1}}-\sqrt{\frac{\varepsilon}{\gamma^2}}\sqrt{\Delta}v\right|\right] \\ &\doteq \overline{G}(z_{x_{N-1}},\Delta), \end{aligned} \tag{69}$$

and note that $\mathrm{E}|y_N| \leq G(z_{x_{N-1}},\Delta)$. It is an exercise in Lagrange multipliers (which we do not include, since it is quite standard) to show that the maximum occurs when v is perpendicular to $z_{x_{N-1}}$. Consequently (using the Pythagorean theorem),

$$G(z_{x_{N-1}},\Delta) \leq \overline{G}(z_{x_{N-1}},\Delta) = \left[|z_{x_{N-1}}|^2+\frac{\varepsilon}{\gamma^2}M^2\Delta\right]^{\frac{1}{2}}. \tag{70}$$

Note that

$$\overline{G}(z_{x_{N-1}},0) = |z_{x_{N-1}}|,$$

$$\frac{d}{d\Delta}\overline{G}(z_{x_{N-1}},0) = \frac{\varepsilon}{2\gamma^2|z_{x_{N-1}}|}M^2,$$

and

$$\frac{d^2}{d\Delta^2}\overline{G}(z_{x_{N-1}},\Delta)=\left(\frac{-1}{4\left[|z_{x_{N-1}}|^2+\frac{\varepsilon}{\gamma^2}M^2\Delta\right]^{\frac{3}{2}}}\right)\frac{\varepsilon^2}{\gamma^4}M^4<0.$$

Consequently,

$$\overline{G}(z_{x_{N-1}},\Delta)\le|z_{x_{N-1}}|+\frac{\varepsilon}{2\gamma^2|z_{x_{N-1}}|}M^2\Delta. \tag{71}$$

Combining (68), (69), and (71), one finds

$$\mathrm{E}|y_N|\le|z_{x_{N-1}}|+\frac{\varepsilon}{4\gamma^2|z_{x_{N-1}}|}M^2\Delta. \tag{72}$$

Further, from the definition of $z_{x_{N-1}}$, (A1), (A2), and (29), $z_{x_{N-1}}\ge 1/4$ for $x_{N-1}\ge 1/2$ if

$$\Delta\le\frac{5}{4(2K+M(C_1+C_2))}\doteq d_1.$$

Consequently, if $\Delta\le\min\{1,d_1\}$, then

$$\mathrm{E}|y_N|\le|z_{x_{N-1}}|+\frac{\varepsilon}{\gamma^2}M^2\Delta \tag{73}$$

for any $x_{N-1}\ge 1/2$. Again using the definition of $z_{x_{N-1}}$, (A1), (A2), and (29), this implies

$$\mathrm{E}|y_N|\le|x_{N-1}|+\left[K(1+|x_{N-1}|)+MD_v(x_{N-1})+\frac{\varepsilon M^2}{\gamma^2}\right]\Delta\quad\forall|x_{N-1}|\ge\frac{1}{2}, \tag{74}$$

provided $\Delta\le\min\{1,d_1\}$. A similar statement holds for $\hat{y}_N$.

Noting that $\psi_x(x)\le b(x)\le C(\frac{3}{2}+2|x|)$ for all x, one finds using (74) and (A4) that

$$|\mathrm{E}\psi_x(\xi)|\le C\left\{\frac{3}{2}+2|x_{N-1}|+\left[2K_6(1+|x_{N-1}|)+2MD_v(x_{N-1})\right]\Delta\right\}$$
$$\forall\,|x_{N-1}|,|\hat{x}_{N-1}|\ge\frac{1}{2}, \tag{75}$$

where $K_6=\max\{K,\varepsilon M^2/(\gamma^2)\}$, provided $\Delta\le\min\{1,d_1\}$.

We now turn to the quadratic portion of $b(x)$, which occurs for $|x|\le 1$. Using the fact that B_N has zero mean,

$$\begin{aligned}\mathrm{E}|y_N|^2&=\mathrm{E}\left|z_{x_{N-1}}+\sqrt{\frac{\varepsilon\Delta}{\gamma^2}}\sigma(x_{N-1})B_N\right|^2\\&=|z_{x_{N-1}}|^2+\frac{\varepsilon}{\gamma^2}\Delta\mathrm{E}\left[B_N^T\sigma^T\sigma B_N\right]\\&\le|z_{x_{N-1}}|^2+\frac{\varepsilon}{\gamma^2}M^2\Delta,\end{aligned}$$

which by the definition of $z_{x_{N-1}}$, (A1), (A2), and (29) again,

$$\leq |x_{N-1}|^2 + K_7(1+|x_{N-1}|)^2\Delta + 2M^2 D_v^2(x_{N-1})\Delta^2,$$

where $K_7 = 2[K + K^2 + M(K_4/\gamma) + (\varepsilon/(2\gamma^2))M^2]$. In a similar manner to that above, this implies that

$$|\mathrm{E}\psi_x(\xi)| \leq C\left[\frac{3}{2} + 2|x_{N-1}|^2 + 2K_7(1+|x_{N-1}|)^2\Delta + \ 4M^2 D_v^2(x_{N-1})\Delta^2\right] \tag{76}$$

$$\forall\, x_{N-1}, \hat{x}_{N-1} \in \Re^n.$$

Combining (75) and (76), one finds

$$|\mathrm{E}\psi_x(\xi)| \leq b(|x_{N-1}|)\left[1 + K_8\Delta + 2M\frac{K_4}{\gamma}\Delta + 8M^2\frac{K_4^2}{\gamma^2}\Delta^2\right] \qquad \forall\, |x_{N-1} - \hat{x}_{N-1}| \leq \frac{1}{2}, \tag{77}$$

where $K_8 = \max\{2K_6, 4K_7\}$, provided $\Delta \leq \min\{1, d_1\}$, which is the inequality indicated earlier.

Combining (67) and (77), one finds

$$\begin{aligned}\mathrm{E}[\psi(y_N) - \psi(\hat{y}_N)] \geq\ & -b(|\hat{x}_{N-1}|)\left[1 + c_1\Delta + c_2\frac{K_4}{\gamma}\Delta + c_3\frac{K_4^2}{\gamma^2}\Delta^2\right]|x_{N-1} - \hat{x}_{N-1}| \\ & - \overline{C}\sqrt{\frac{\varepsilon}{\gamma^2}}\left[K_3\Delta + M\frac{K_4}{\gamma}\Delta\right]|x_{N-1} - \hat{x}_{N-1}|,\end{aligned} \tag{78}$$

where

$$\begin{aligned}\overline{c}_1 &= K + K_8 + KK_8, \\ c_2 &= 2MK + K_8 M L_\sigma, \\ c_3 &= M^2[2L_\sigma + 8(1+K)],\end{aligned}$$

provided $\Delta \leq \min\{1, d_1\}$ and $|x_{N-1} - \hat{x}_{N-1}| \leq \frac{1}{2}$. By symmetry, and by taking intermediary steps (to remove the $|x_{N-1} - \hat{x}_{N-1}| \leq \frac{1}{2}$ bound), this implies

$$\begin{aligned}\left|\mathrm{E}[\psi(y_N) - \psi(\hat{y}_N)]\right| \leq\ & b(|\hat{x}_{N-1}|)\left[1 + c_1\Delta + c_2\frac{K_4}{\gamma}\Delta + c_3\frac{K_4^2}{\gamma^2}\Delta^2\right]|x_{N-1} - \hat{x}_{N-1}| \\ & + \overline{C}\sqrt{\frac{\varepsilon}{\gamma^2}}\left[K_3\Delta + M\frac{K_4}{\gamma}\Delta\right]|x_{N-1} - \hat{x}_{N-1}|,\end{aligned} \tag{79}$$

provided $\Delta \leq \min\{1, d_1\}$.

Finally, combining (79) with (56) (and letting $\delta \downarrow 0$) yields

$$\begin{aligned}&\left|\overline{V}_{N-1}^N(x_{N-1}) - \overline{V}_{N-1}^N(\hat{x}_{N-1})\right| \\ &\quad \leq b(|\hat{x}_{N-1}|)\left[1 + c_1\Delta + c_2\frac{K_4}{\gamma}\Delta + c_3\frac{K_4^2}{\gamma^2}\Delta^2\right]\cdot|x_{N-1} - \hat{x}_{N-1}| \\ &\quad + \overline{C}\sqrt{\frac{\varepsilon}{\gamma^2}}\left[K_3\Delta + M\frac{K_4}{\gamma}\Delta\right]\cdot|x_{N-1} - \hat{x}_{N-1}|,\end{aligned} \tag{80}$$

provided $\Delta \leq \min\{1, d_1\}$, where $c_1 = \overline{c}_1 + 1$.

LEMMA A.1. *$\overline{V}^N_{N-1}$ is semiconvex with uniform constant $C[2K\Delta + 4L_\sigma \frac{K_4}{\gamma}\Delta] + \overline{C}\{1 + c_4\Delta + c_5\frac{K_4}{\gamma}\Delta + c_6\frac{K_4^2}{\gamma^2}\Delta^2\}$, where c_4, c_5, c_6 are given in the proof.*

Proof. Recall that

$$\overline{V}^N_{N-1}(x_{N-1}) = \sup_{|v_N| \le C_1 + C_2|x_{N-1}|} \overline{J}^N_{N-1}(x_{N-1}, v_N),$$

where

$$\overline{J}^N_{N-1}(x_{N-1}, v_N) = \mathrm{E}\left[\left(\ell(x_{N-1}) - \frac{\gamma^2}{2}|v_N|^2\right)\Delta + \psi(y_N)\right].$$

Fix a v_N satisfying $|v_N| \le C_1 + C_2|x_{N-1}|$. First, a lower bound on

$$\frac{d^2}{dh^2}\overline{J}^N_{N-1}(x_{N-1} + h\eta, v_N)\Big|_{h=0} \tag{81}$$

will be obtained for any $\eta \in \Re^n$ with $|\eta| = 1$.

One has

$$\frac{d^2}{dh^2}\overline{J}^N_{N-1}(x_{N-1}+h\eta, v_N)|_{h=0} = \eta^T l_{xx}(x_{N-1})\eta\Delta + \mathrm{E}\{{\zeta^1}^T\psi_{xx}(y_N)\zeta^1\} + \mathrm{E}\{\psi_x(y_N)\zeta^2\},$$

where

$$\zeta^1 \doteq \frac{d}{dh}y^h_N\big|_{h=0},$$

$$\zeta^2 \doteq \frac{d^2}{dh^2}y^h_N\big|_{h=0},$$

$$y^h_N = (x_{N-1} + h\eta) + [f(x_{N-1} + h\eta) + \sigma(x_{N-1} + h\eta)v_N]\Delta + \sqrt{\frac{\varepsilon\Delta}{\gamma^2}}\sigma(x_{N-1} + h\eta)B_N.$$

Consequently, by Assumption (A4),

$$\frac{d^2}{dh^2}\overline{J}^N_{N-1}(x_{N-1} + h\eta, v_N)|_{h=0} \ge -\overline{C}\Delta + \mathrm{E}\{{\zeta^1}^T\psi_{xx}(y_N)\zeta^1\} + \mathrm{E}\{\psi_x(y_N)\zeta^2\}. \tag{82}$$

Note that

$$\zeta^1 = \eta + [f_x(x_{N-1})\eta + \sigma_x(x_{N-1})\eta v_N]\,\Delta + \sqrt{\frac{\varepsilon\Delta}{\gamma^2}}\sigma_x(x_{N-1})\eta B_N.$$

Using the facts that $\mathrm{E}(B_N) = 0$ and $\mathrm{E}(B_N^2) = 1$, this yields

$$\begin{aligned}\mathrm{E}\left\{{\zeta^1}^T\psi_{xx}(y_N)\zeta^1\right\} \ge& -\overline{C}\mathrm{E}\left|\eta + [f_x(x_{N-1})\eta + \sigma_x(x_{N-1})\eta v_N]\,\Delta + \sqrt{\frac{\varepsilon\Delta}{\gamma^2}}\sigma_x(x_{N-1})\eta B_N\right|^2 \\ \ge& -\overline{C}\left[1 + \left(2|f_x(x_{N-1})| + 2|\sigma_x(x_{N-1})v_N| + \frac{\varepsilon}{\gamma^2}|\sigma_x(x_{N-1})|^2\right)\Delta \right. \\ & \left. + \left(|f_x(x_{N-1})| + |\sigma_x(x_{N-1})v_N|\right)^2\Delta^2\right],\end{aligned}$$

which by Assumptions (A1) and (A2), and by (29),

$$\geq -\overline{C}\left[1 + K_9\Delta + 2\frac{K_4}{\gamma}\Delta + 2\frac{K_4^2}{\gamma^2}\Delta^2\right], \tag{83}$$

where $K_9 = 2K + 2K^2 + \varepsilon L_\sigma^2/\gamma^2$.

Next, one finds

$$\zeta^2 = \left[\eta^T f_{xx}(x_{N-1})\eta + \eta^T \sigma_{xx}(x_{N-1})\eta v_N\right]\Delta + \sqrt{\frac{\varepsilon\Delta}{\gamma^2}}\eta^T\sigma_{xx}(x_{N-1})\eta B_N.$$

Then using (componentwise)

$$(\psi_x)_i(y_N) = (\psi_x)_i(x_{N-1}) + [(\psi_{xx})_i(\xi_i)](y_N - x_{N-1})$$

for the proper choice of ξ_i on the line segment from x_{N-1} to y_N (leaving out the x_{N-1} argument where it is clear), and letting $H(\{\xi_i\})$ be the matrix composed of the rows $(\psi_{xx})_i(\xi_i)$, one finds

$$\begin{aligned}
\mathrm{E}\{\psi_x(y_N)\zeta^2\} &= \psi_x(x_{N-1})\left[\eta^T f_{xx}(x_{N-1})\eta + \eta^T\sigma_{xx}(x_{N-1})\eta v_N\right]\Delta \\
&\quad + \mathrm{E}\Bigg\{H(\{\xi\}_i)\Bigg[\eta^T(f_{xx} + \sigma_{xx}v_N)\eta(f + \sigma v_N)\Delta^2 \\
&\qquad + \left[(f + \sigma v_N)\eta^T\sigma_{xx}\eta\sqrt{\frac{\varepsilon}{\gamma^2}} + \eta^T(f_{xx} + \sigma_{xx}v_N)\eta\sigma\sqrt{\frac{\varepsilon}{\gamma^2}}\right]B_N\Delta^{3/2} \\
&\qquad + \eta^T\sigma_{xx}\eta\sigma\frac{\varepsilon}{\gamma^2}B_N^2\Delta\Bigg]\Bigg\} \\
&\geq -|\psi_x(x_{N-1})|\left[|f_{xx}(x_{N-1})| + |\sigma_{xx}(x_{N-1})v_N|\right]\Delta \\
&\quad - \overline{C}\Bigg[\big(|f_{xx}| + |\sigma_{xx}v_N|\big)\big(|f| + |\sigma v_N|\big)\Delta^2 \\
&\quad + \Big(\big(|f| + |\sigma v_N|\big)|\sigma_{xx}| + \big(|f_{xx}| + |\sigma_{xx}v_N|\big)|\sigma|\Big)\sqrt{\frac{\varepsilon}{\gamma^2}}\Delta^{3/2} + |\sigma_{xx}|\,|\sigma|\frac{\varepsilon}{\gamma^2}\Delta\Bigg],
\end{aligned}$$

from which with a few bounds on rational functions of $|x|$ (such as $(3/2 + 2|x|)/(1 + |x|) \leq 2$) one has

$$\begin{aligned}
&\geq -C\left[2K + 4L_\sigma\frac{K_4}{\gamma}\right]\Delta \\
&\quad - \overline{C}\Bigg\{\left[K^2 + ML_\sigma\frac{K_4^2}{\gamma^2} + KM\frac{K_4}{\gamma} + 2KL_\sigma\frac{K_4}{\gamma}\right]\Delta^2 \\
&\qquad + \left[K(L_\sigma + M) + 2ML_\sigma\frac{K_4}{\gamma}\right]\sqrt{\frac{\varepsilon}{\gamma^2}}\Delta^{3/2} + ML_\sigma\frac{\varepsilon}{\gamma^2}\Delta\Bigg\} \\
&\geq -C\left[2K\Delta + 4L_\sigma\frac{K_4}{\gamma}\Delta\right] - \overline{C}\left\{K_{11}\Delta + K_{12}\frac{K_4}{\gamma}\Delta + K_{13}\frac{K_4^2}{\gamma^2}\Delta^2\right\},
\end{aligned} \tag{84}$$

where

$$K_{11} = K^2\Delta + K(L_\sigma + M)\sqrt{\frac{\varepsilon}{\gamma^2}}\sqrt{\Delta} + ML_\sigma\frac{\varepsilon}{\gamma^2},$$

$$K_{12} = (KM + 2KL_\sigma)\Delta + 2ML_\sigma\sqrt{\frac{\varepsilon}{\gamma^2}}\sqrt{\Delta},$$
$$K_{13} = ML_\sigma.$$

Combining (82), (83), and (84) yields

$$\frac{d^2}{dh^2}\overline{J}^N_{N-1}(x_{N-1}+h\eta, v_N)|_{h=0} \geq -C\left[2K\Delta + 4L_\sigma\frac{K_4}{\gamma}\Delta\right] - \overline{C}\left\{1 + c_4\Delta + c_5\frac{K_4}{\gamma}\Delta + c_6\frac{K_4^2}{\gamma^2}\Delta^2\right\}, \tag{85}$$

where $c_4 = 1 + K_9 + K_{11}$, $c_5 = 2 + K_{12}$, and $c_6 = 2 + K_{13}$.

This implies that $\overline{J}^N_{N-1}(x, v_N)$ is semiconvex, with uniform constant given by the right-hand side of (85) over all of $\Re^n$ if $|v_N| \leq (C/\gamma)[C_1 + C_2|x|]$. By the definition of $\overline{V}^N_{N-1}$ as a supremum of these, $\overline{V}^N_{N-1}$ is semiconvex with this same constant. This completes the proof of the lemma. □

We now proceed to step backwards. Note that we used the derivatives of ψ in the above argument. One can first smooth $\overline{V}^N$ via convolution with approximations to the identity, before proceeding backwards; since this is an obvious, but time-consuming, technicality, we do not include it. Instead, simply assuming differentiability, we now have from (29), (80), and the above lemma,

$$|v_N(x)| \leq \frac{C}{\gamma}\left[C_1 + C_2|x|\right], \tag{86}$$

$$\left|\overline{V}^N_{N-1_x}(x)\right| \leq b(|x|)\left[1 + c_1\Delta + c_2\frac{K_4}{\gamma}\Delta + c_3\frac{K_4^2}{\gamma^2}\Delta^2\right] + \overline{C}\sqrt{\frac{\varepsilon}{\gamma^2}}\left[K_3\Delta + M\frac{K_4}{\gamma}\Delta\right], \tag{87}$$

and that $\overline{V}^N_{N-1}$ is semiconvex with constant

$$\overline{C}_1 = C\left[2K\Delta + 4L_\sigma\frac{K_4}{\gamma}\Delta\right] + \overline{C}_0\left\{1 + c_4\Delta + c_5\frac{K_4}{\gamma}\Delta + c_6\frac{K_4^2}{\gamma^2}\Delta^2\right\}, \tag{88}$$

where $\overline{C}_0 \doteq \overline{C}$.

Define

$$K_4^{(0)} = K_4,$$
$$M_0 = 1,$$
$$M_1 = \left\{\left[1 + c_1\Delta + c_2\frac{K_4^{(0)}}{\gamma}\Delta + c_3\frac{{K_4^{(0)}}^2}{\gamma^2}\Delta^2\right] + \frac{\overline{C}_0}{b(0)}\sqrt{\frac{\varepsilon}{\gamma^2}}\left[K_3\Delta + M\frac{K_4^{(0)}}{\gamma}\Delta\right]\right\}M_0.$$

Note then that the above Lipschitz and semiconvexity bounds become

$$\left|\overline{V}^N_{N-1_x}(x)\right| \leq b(|x|)M_1,$$
$$\overline{C}_1 = \overline{C}_0\left\{1 + c_4\Delta + c_5\frac{K_4^{(0)}}{\gamma}\Delta + c_6\frac{{K_4^{(0)}}^2}{\gamma^2}\Delta^2\right\} + CM_0\left[2K\Delta + 4L_\sigma\frac{K_4^{(0)}}{\gamma}\Delta\right].$$

Stepping backward, one finds

$$
\begin{aligned}
&|v_{N-1}(x)| \le \frac{C}{\gamma}\left[C_1 + C_2|x|\right] M_1,\\
&\left|\overline{V}^N_{N-2_x}(x)\right| \le b(|x|)M_2,\\
&\overline{C}_2 \doteq \overline{C}_1\left\{1 + c_4\Delta + c_5\frac{K_4^{(1)}}{\gamma}\Delta + c_6\frac{K_4^{(1)^2}}{\gamma^2}\Delta^2\right\} + CM_1\left[2K\Delta + 4L_\sigma\frac{K_4^{(1)}}{\gamma}\Delta\right],
\end{aligned}
$$

where

$$
\begin{aligned}
&K_4^1 = M_1K_4^{(0)},\\
&M_2 = \left\{\left[1 + c_1\Delta + c_2\frac{K_4^{(1)}}{\gamma}\Delta + c_3\frac{K_4^{(1)^2}}{\gamma^2}\Delta^2\right] + \frac{\overline{C}_1}{b(0)}\sqrt{\frac{\varepsilon}{\gamma^2}}\left[K_3\Delta + M\frac{K_4^{(1)}}{\gamma}\Delta\right]\right\}M_1.
\end{aligned}
$$

Continuing this process, one has

$$
\begin{aligned}
&|v_{N-n}(x)| \le \frac{C}{\gamma}\left[C_1 + C_2|x|\right] M_n,\\
&\left|\overline{V}^N_{N-(n+1)_x}(x)\right| \le b(|x|)M_{n+1},\\
&\overline{C}_{n+1} \doteq \overline{C}_n\left\{1 + c_4\Delta + c_5\frac{K_4^{(n)}}{\gamma}\Delta + c_6\frac{K_4^{(n)^2}}{\gamma^2}\Delta^2\right\} + CM_n\left[2K\Delta + 4L_\sigma\frac{K_4^{(n)}}{\gamma}\Delta\right],
\end{aligned}
$$

where

$$
\begin{aligned}
&K_4^{(n)} = M_nK_4^{(0)},\\
&M_{n+1} = \left\{\left[1 + c_1\Delta + c_2\frac{K_4^{(n)}}{\gamma}\Delta + c_3\frac{K_4^{(n)^2}}{\gamma^2}\Delta^2\right] + \frac{\overline{C}_n}{b(0)}\sqrt{\frac{\varepsilon}{\gamma^2}}\left[K_3\Delta + M\frac{K_4^{(n)}}{\gamma}\Delta\right]\right\}M_n.
\end{aligned}
$$

Now, we note that one is interested here in the Lipschitz constant for $\overline{V}^N_0$, and in particular, in whether this stays bounded as $N \to \infty$. Noting in the above that $\Delta = T/N$, we see that this question reduces to consideration of the sequence

$$
\begin{aligned}
M_{N,0} &= 1,\\
\overline{C}_{N,0} &= \overline{C},\\
M_{N,n+1} &= \left\{\left[1 + c_1\frac{T}{N} + c_2\frac{K_4^{(N,n)}}{\gamma}\frac{T}{N} + c_3\frac{K_4^{(N,n)^2}}{\gamma^2}\frac{T^2}{N^2}\right]\right.\\
&\qquad \left. + \frac{\overline{C}_{N,n}}{b(0)}\sqrt{\frac{\varepsilon}{\gamma^2}}\left[K_3\frac{T}{N} + M\frac{K_4^{(N,n)}}{\gamma}\frac{T}{N}\right]\right\}M_{N,n},\\
\overline{C}_{N,n+1} &\doteq \overline{C}_{N,n}\left\{1 + c_4\frac{T}{N} + c_5\frac{K_4^{(N,n)}}{\gamma}\frac{T}{N} + c_6\frac{K_4^{(N,n)^2}}{\gamma^2}\frac{T^2}{N^2}\right\}\\
&\qquad + CM_{N,n}\left[2K\frac{T}{N} + 4L_\sigma\frac{K_4^{(N,n)}}{\gamma}\frac{T}{N}\right],
\end{aligned}
$$

where

$$K_4^{(N,n)} = M_{N,n}K_4.$$

Note that $|\overline{V}^N_{0\ x}(x)| \le b(|x|)M_{N,N}$.

Showing that $M_{N,N}$ stays bounded for sufficiently large γ can be done with some rather crude estimates. Again, we are not interested here in the minimal possible γ, but mainly in developing the machinery to handle risk-sensitive control limit problems under quadratic growth assumptions. In that spirit, one may define

$$R_{N,n} \doteq \max\{M_{N,n}, \overline{C}_{N,n}\}.$$

Then, for $N > K_4T$ one has

$$M_{N,n+1} \le R_{N,n}\left\{1 + c_1\frac{T}{N} + \frac{\overline{c}_2}{\gamma}R_{N,n}\frac{T}{N} + \frac{\overline{c}_3}{\gamma^2}R^2_{N,n}\frac{T}{N}\right\},$$

$$\overline{C}_{N,n+1} \le R_{N,n}\left\{1 + c_4\frac{T}{N} + \frac{\overline{c}_5}{\gamma}R_{N,n}\frac{T}{N} + \frac{\overline{c}_6}{\gamma^2}R^2_{N,n}\frac{T}{N}\right\},$$

where $\overline{c}_2 = c_2 + \sqrt{\varepsilon}K_3/b(0)$, $\overline{c}_3 = c_3 + \sqrt{\varepsilon}MK_4/b(0)$, $\overline{c}_4 = c_4 + 2CK$, $\overline{c}_5 = c_5K_4 + 4L_\sigma CK_4$, and $\overline{c}_6 = c_6K_4$. This implies

$$R_{N,n+1} \le R_{N,n}\left\{1 + c_7\frac{T}{N} + \frac{c_8}{\gamma}R_{N,n}\frac{T}{N} + \frac{c_9}{\gamma^2}R^2_{N,n}\frac{T}{N}\right\},$$

where $c_7 = c_1 + \overline{c}_4$, $c_8 = c_2 + \overline{c}_5$, and $c_7 = \overline{c}_3 + \overline{c}_6$. Letting $\overline{\rho}_{N,n} \doteq \ln(R_{N,n})$, one has

$$\overline{\rho}_{N,n+1} \le \overline{\rho}_{N,n} + \left[c_7 + \frac{c_8}{\gamma}e^{\overline{\rho}_{N,n}} + \frac{c_9}{\gamma^2}e^{2\overline{\rho}_{N,n}}\right]\frac{T}{N},$$

and perhaps one should note $\overline{\rho}_{N,0} = \max\{0, \ln(\overline{C})\}$. Then $\overline{\rho}_{N,n} \le \rho_{N,n}$ for all N, n, where $\rho_{N,n}$ satisfies

$$\rho_{N,n+1} = \rho_{N,n} + \left[c_7 + \frac{c_{10}}{\gamma}e^{2\rho_{N,n}}\right]\frac{T}{N}, \tag{89}$$

$$\rho_{N,0} = \max\{0, \ln(\overline{C})\}, \tag{90}$$

and $c_{10} = c_8 + (c_9/\gamma)$. Note that then

$$M_{N,N} \le e^{\rho_{N,N}} \qquad \forall\, N.$$

Also, note that (89), (90) is simply the Euler numerical method (with N time-steps) for the ODE problem over $[0, T]$ given by

$$\dot{\widehat{\rho}} = c_7 + \frac{c_{10}}{\gamma}e^{2\rho},$$

$$\widehat{\rho}(0) = \max\{0, \ln(\overline{C})\}.$$

By elementary calculus, one can show that, given $T < \infty$, there is $\widetilde{\gamma} < \infty$ such that this ODE has a solution over $[0, T]$. Letting the maximum of this solution be ρ_M, one sees that $M_{N,N} \le e^{\rho_M + 1}$ for sufficiently large N. This behavior (increasingly large γ required for increasingly large T) is expected, since the Riccati equation for the linear-quadratic case displays the finite-time blow-up property. (A reader curious about this point might note that letting $\widehat{\mathcal{M}} \doteq \exp(\widehat{\rho})$, one finds that $\widehat{\mathcal{M}}$ satisfies the ODE $\dot{\widehat{\mathcal{M}}} = c_7\widehat{\mathcal{M}} + (c_{10}/\gamma)\widehat{\mathcal{M}}^3$, which is somewhat worse than the Riccati equation which has a squared term only on the right-hand side; thus we have lost a bit in all the bounding estimates.)

Acknowledgments. The authors wish to thank M. Bardi, W. H. Fleming, S.-J. Sheu, the referee, and the associate editor for helpful discussions and/or comments.

REFERENCES

[1] R.A. Adams, *Sobolev Spaces*, Academic Press, New York, London, 1975.
[2] M. Bardi and I. Capuzzo Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser Boston, Cambridge, MA, 1997.
[3] M. Bardi and F. Da Lio, *On the Bellman equation for some unbounded control problems*, Nonlinear Differential Equations Appl., 4 (1997), pp. 491–510.
[4] A. Bensoussan, J. Frehse, and H. Nagai, *Some results on risk–sensitive control with full observation*, Appl. Math. Optim., 37 (1998), pp. 1–42.
[5] A. Bensoussan and H. Nagai, *Conditions for no breakdown and Bellman equations of risk–sensitive control*, Appl. Math. Optim., to appear.
[6] A. Bensoussan and J.H. van Schuppen, *Optimal control of partially observable stochastic systems with an exponential–of–integral performance index*, SIAM J. Control Optim., 23 (1985), pp. 599–613.
[7] C.D. Charalambous and R.J. Elliott, *Remarks on the explicit solutions for nonlinear partially observable stochastic control problems and relations to H_∞ robust control*, in Proceedings of the 34th IEEE Conference on Decision and Control, 1995, pp. 2858–2863.
[8] F.H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.
[9] P. Dupuis and W.M. McEneaney, *Risk-sensitive and robust escape criteria*, SIAM J. Control Optim., 35 (1997), pp. 2021–2049.
[10] R.J. Elliott, *Stochastic Calculus and Applications*, Springer-Verlag, New York, 1982.
[11] R.J. Elliott and N.J. Kalton, *The existence of value in differential games*, Mem. Amer. Math. Soc., 126 (1972).
[12] C.-H. Fan, J.L. Speyer, and C.R. Jaensch, *Centralized and decentralized solutions of the linear-exponential-Gaussian problem*, IEEE Trans. Automat. Control, 39 (1994), pp. 1986–2003.
[13] E. Fernandez-Gaucherand and S.I. Marcus, *Risk–sensitive optimal control of hidden Markov models: A case study*, in Proceedings of the 33rd IEEE Conference on Decision and Control, Orlando, FL, 1994, pp. 1657–1662.
[14] W.H. Fleming, *The Cauchy problem for degenerate parabolic equations*, J. Mathematics and Mechanics, 13 (1964), pp. 987–1008.
[15] W.H. Fleming and D. Hernandez-Hernandez, *Risk-sensitive control of finite state machines on an infinite horizon* I, SIAM J. Control Optim., 35 (1997), pp. 1790–1810.
[16] W.H. Fleming and M.R. James, *The risk–sensitive index and the H_2 and H_∞ norms for nonlinear systems*, Math. Control Signals Systems, 8 (1995), pp. 199–221.
[17] W.H. Fleming and W.M. McEneaney, *Risk–sensitive optimal control and differential games*, in Proceedings of the Stochastic Theory and Adaptive Controls Workshop, Lecture Notes in Control and Inform. Sci. 184, Springer-Verlag, New York, 1992.
[18] W.H. Fleming and W.M. McEneaney, *Risk–sensitive control with ergodic cost criteria*, in Proceedings of the 31st IEEE Conference on Decision and Control, 1992.
[19] W.H. Fleming and W.M. McEneaney, *Risk-sensitive control on an infinite time horizon*, SIAM J. Control Optim., 33 (1995), pp. 1881–1915.
[20] W.H. Fleming and W.M. McEneaney, *A max-plus-based algorithm for a Hamilton–Jacobi–Bellman equation of nonlinear filtering*, SIAM J. Control Optim., 38 (2000), pp. 683–710.
[21] W.H. Fleming and R.W. Rishel, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, Berlin, 1975.
[22] W.H. Fleming and H.M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1992.
[23] W.H. Fleming and S.–J. Sheu, *Asymptotics for the first eigenvalue and eigenfunction of a nearly first order operator with large potential*, Ann. Probab., 25 (1997), pp. 1953–1994.
[24] W.H. Fleming and P.E. Souganidis, *On the existence of value functions of two–player, zero–sum stochastic differential games*, Indiana Univ. Math. J., 38 (1989), pp. 293–314.
[25] I.V. Girsanov, *On transforming a certain class of stochastic processes by absolutely continuous substitution of measures*, Theory Probab. Appl., 5 (1960), pp. 285–301.
[26] H. Ishii, *Comparison results for Hamilton–Jacobi equations without growth condition on solutions from above*, Appl. Anal., 67 (1997), pp. 357–372.
[27] H. Ishii, H. Nagai, and F. Teramoto, *A singular limit on risk sensitive control and semi–classical analysis*, in Proceedings of the 7th Japan-Russia Symposium on Probability

Theory and Math. Stats., S. Watanabe et al., eds., World Scientific, River Edge, NJ, 1996, pp. 164–173.

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

[29] M.R. James, *Asymptotic analysis of non–linear stochastic risk–sensitive control and differential games*, Math. Control Signals Systems, 5 (1992), pp. 401–417.

[30] M.R. James, J.S. Baras, and R.J. Elliott, *Output feedback risk–sensitive control and differential games for continuous–time nonlinear systems*, in Proceedings of the 32nd IEEE Conference on Decision and Control, San Antonio, TX, 1993.

[31] H. Kaise and H. Nagai, *Bellman–Isaacs equations of ergodic type related to risk–sensitive control and their singular limits*, Asympt. Anal., 16 (1998), pp. 347–362.

[32] I. Karatzas and S.E. Shreve, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York, Berlin, 1988.

[33] G. Lebourg, *Valeur moyenne pour gradient généralisé*, Co. R. Acad. Sci. Paris, 281 (1975), pp. 795–797.

[34] P.-L. Lions and P.E. Souganidis, *Differential games, optimal control and directional derivates of viscosity solutions of Bellman's and Isaacs' equations*, SIAM J. Control Optim., 23 (1985), pp. 566–583.

[35] O.A. Ladyzhenskaja, V.A. Solonnikov, and N.N. Ural'ceva, *Linear and Quasi-linear Equations of Parabolic Type*, Amer. Math. Soc. Transl. Ser., 1968.

[36] R.S. Lipster and A.N. Shiryayev, *Statistics of Random Processes,* I, Springer-Verlag, New York, Berlin, 1977.

[37] W.M. McEneaney and P. Dupuis, *A risk–sensitive escape criterion and robust limit*, Proceedings of the 33rd IEEE Conference on Decision and Control, 1994, pp. 4195–4197.

[38] W.M. McEneaney and K. Ito, *Infinite time–horizon risk sensitive systems with quadratic growth*, in Proceedings of the 36th IEEE Conference on Decision and Control, 1997.

[39] W.M. McEneaney, *Elimination of troublesome disturbances with application to representation results for H_∞ control DPEs*, in Proceedings of the Seventh International Symposium on Dynamic Games and Applications, Shonan Village, Japan, International Society of Dynamic Games, 1996, pp. 662–671.

[40] W.M. McEneaney, *A uniqueness result for the Isaacs equation corresponding to nonlinear H_∞ control*, Math. Control Signals Systems, 11 (1998), pp. 303–334.

[41] W.M. McEneaney, *Robust control and differential games on a finite time horizon*, Math. Control Signals Systems, 8 (1995), pp. 138–166.

[42] W.M. McEneaney, *Uniqueness for viscosity solutions of nonstationary Hamilton–Jacobi–Bellman equations under some a priori conditions (with applications)*, SIAM J. Control Optim., 33 (1995), pp. 1560–1576.

[43] W.M. McEneaney, *Connections between Risk–Sensitive Stochastic Control, Differential Games and H_∞ Control: The Nonlinear Case*, Doctoral Thesis, Brown University, Providence, RI, 1993.

[44] H. Nagai, *Bellman equations of risk-sensitive control*, SIAM J. Control Optim., 34 (1996), pp. 74–101.

[45] O.A. Oleinik and S.N. Kruzhkov, *Quasi-linear parabolic equations of the second order with many independent variables*, Uspekhi Mat. Nauk., 16 (1961), pp. 115–155.

[46] Z. Pan and T. Başar, *Backstepping controller design for nonlinear stochastic systems under a risk-sensitive cost criterion*, SIAM J. Control Optim., 37 (1999), pp. 957–995.

[47] T. Runolfsson, *The Equivalence Between Infinite–Horizon Optimal Control of Stochastic Systems with Exponential-of-Integral Performance Index and Stochastic Differential Games*, Technical report JHU–ECE, John Hopkins University, Baltimore, MD, pp. 91–07.

[48] T. Runolfsson, *Risk–sensitive control of Markov chains and differential games*, in Proceedings of the 32nd IEEE Conference on Decision and Control, 1993.

[49] A. Swiech, *Another approach to the existence of value functions of stochastic differential games*, J. Math. Anal. Appl., 204 (1996), pp. 884–897.

[50] A.J. Veretennikov, *On strong solutions and explicit formulas for solutions of stochastic integral equations*, Math. USSR Sbornik, 39 (1981), pp. 387–403.

[51] P. Whittle, *Risk–sensitive linear/quadratic/Gaussian control*, Adv. Appl. Prob., 13 (1981), pp. 764–777.

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

[53] W.P. Ziemer, *Weakly Differentiable Functions*, Springer-Verlag, New York, Berlin, 1989.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 6, pp. 1663–1696

PERPETUAL AMERICAN OPTIONS UNDER LÉVY PROCESSES*

S. I. BOYARCHENKO† AND S. Z. LEVENDORSKIĬ‡

Abstract. We consider perpetual American options, assuming that under a chosen equivalent martingale measure the stock returns follow a Lévy process. For put and call options, their analogues for more general payoffs, and a wide class of Lévy processes that contains Brownian motion, normal inverse Gaussian processes, hyperbolic processes, truncated Lévy processes, and their mixtures, we obtain formulas for the optimal exercise price and the fair price of the option in terms of the factors in the Wiener–Hopf factorization formula, i.e., in terms of the resolvents of the supremum and infimum processes, and derive explicit formulas for these factors. For calls, puts, and some other options, the results are valid for any Lévy process.

We use Dynkin's formula and the Wiener–Hopf factorization to find the explicit formula for the price of the option for any candidate for the exercise boundary, and by using this explicit representation, we select the optimal solution.

We show that in some cases the principle of the smooth fit fails and suggest a generalization of this principle.

Key words. Lévy processes, perpetual American options, Wiener–Hopf factorization

AMS subject classifications. 60G40, 90A09, 93E20

PII. S0363012900373987

1. Introduction.

Consider the market of a riskless bond and a stock whose returns follow a Lévy process. If the Lévy process is neither a Brownian motion nor a Poisson process, then the market is incomplete. According to the modern martingale approach to option pricing [16], arbitrage-free prices can be obtained as expectations under any equivalent martingale measure (EMM), which is absolutely continuous w.r.t. the historic measure.

Let the riskless rate $r > 0$ and the dividend rate $\lambda \geq 0$ be fixed, let $S = \{S_t\}_{t\geq 0}$, $S_t = \exp X_t$, be the price process of the stock, and let $\mathbf{Q}$ be an EMM chosen by the market. Let $\{X_t\}$ be a Lévy process under $\mathbf{Q}$, and $(\Omega, \mathcal{F}, \mathbf{Q})$ the corresponding probability space. (For general definitions of the theory of Lévy processes, see, e.g., [32], [5], and [33].)

Let $g(X_t)$ be the payoff function for a perpetual American option on the stock (e.g., for a put, $g(x) = K - e^x$, and for a call, $g(x) = e^x - K$, where K is the strike price; for the formulation of our results, it is more convenient to use $g(X_t)$ rather than $\max\{g(X_t), 0\}$). Set $q = r + \lambda$, and denote by $V_*(x)$, where $x = \ln S$, the rational price of the perpetual American option. It is given by

$$V_*(x) = \sup E^x[e^{-q\tau} g(X_\tau)], \tag{1.1}$$

where E^x denotes the expectation under $\mathbf{Q}$, and the supremum is taken over a set $\mathcal{M}$ of all stopping times $\tau = \tau(\omega)$ satisfying $0 \leq \tau(\omega) < \infty$, $\omega \in \Omega$ (see, e.g., [34, Chapter XVIII, section 2]).

*Received by the editors June 19, 2000; accepted for publication (in revised form) August 18, 2001; published electronically February 14, 2002.

http://www.siam.org/journals/sicon/40-6/37398.html

†Department of Economics, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104 (sboyarch@ssc.upenn.edu).

‡Rostov State University of Economics, 69, B. Sadovaya, Rostov-on-Don, 344007, Russia (leven@ns.rnd.runnet.ru).

Suppose that the optimal stopping time is the hitting time of the exterior of an open set $\mathcal{C} \subset \mathbf{R}$:

$$\tau_* = \inf\{t \geq 0 \mid X_t \notin \mathcal{C}\}. \tag{1.2}$$

In the pure diffusion case, one finds a candidate for the optimal stopping time (1.2) or, equivalently, a boundary of $\mathcal{C}$, by using the smooth fit principle as in [21] and in [27]; see also [34]. When jumps are present, this principle may fail. This effect was demonstrated in [30] for sequential testing problems for the Poisson process, and in [7], [8] and [10], [11], [12] for a discrete-time model of the investment under uncertainty, the perpetual American put in discrete time, and the perpetual American put in continuous time, respectively. (In [7], [8] and [10], [11], [12], only the free boundary value problem was considered.)

We use a direct reduction of the problems for puts and calls to the free boundary problem based on Dynkin's formula, and we solve this problem directly by using the Wiener–Hopf factorization method in the form which is standard in the theory of pseudodifferential operators (PDO) (see, e.g., [20]).[1] A similar but less direct approach was used in [28] for pure jumps and jump-diffusion mixtures of special forms; only puts and calls were considered. In [15], the perpetual call for random walks was considered, and the answer in terms of the supremum process was obtained. In [29], the paper [15] was used to derive results in the same nonexplicit form for calls and puts and any Lévy process.

We do not use the smooth pasting principle but make the direct comparison of expected payoffs for different choices of candidates for the exercise price. We formulate the optimality conditions for a relatively general payoff, and verify them for puts, calls, and other options with payoffs of the form

$$g(x) = \sum_{j=1}^{m} c_j \exp(\gamma_j x); \tag{1.3}$$

the list of examples can be extended. An example of (1.3) is an option which gives its owner the right to sell the stock for $K + a\sqrt{S_t}$.

We obtain the optimal solution in the class $\mathcal{M}_0$ of hitting times of semi-infinite intervals; the verification in the class $\mathcal{M}$ is made for Brownian motions (BM), normal inverse Gaussian processes (NIG) and their generalizations, hyperbolic processes (HP), truncated Lévy processes (TLP), and any finite mixture of independent BM, NIG, HP, and TLP.

The results are formulated in terms of the infinitesimal generator and the factors in the Wiener–Hopf factorization formula (equivalently, in terms of the resolvents of the supremum and infimum processes); in this form, they make sense for any Lévy process. We prove the results by using explicit analytic expressions for the factors, obtained in the paper for a wide class of Lévy processes. This class can be loosely characterized as a class of Lévy processes with the Lévy measures exponentially decaying at infinity and having polynomial singularity at the origin; we call these processes regular Lévy processes of exponential type (RLPE).[2] Notice that BM, NIG, HP, and TLP and any finite mixture of independent BM, NIG, HP, and TLP are RLPE.

[1] Notice that the stochastic version of the Wiener–Hopf method was used to solve boundary value problems for stochastic processes in queuing theory and insurance (see [13] and [31]).

[2] In [9], [10], [11], [12], a misleading name generalized truncated Lévy processes was used. Here we use the name suggested in [4].

We exclude variance gamma processes (VGP), since they need special treatment at many places; in particular, the explicit formulas for the factors in the Wiener–Hopf factorization formula, which we use, need regularization in the case of VGP.

Not only BM, but also the other mentioned processes have been widely used to describe the behavior of stock prices in real financial markets: VGPs have been used by Madan and coauthors in a series of papers during the 90s (see [23] and the bibliography there); HP were constructed and used by Eberlein and coauthors [17], [18], [19]; hyperbolic distributions were constructed in [1]; NIG were constructed in [2] and used to model German stocks in [3]; TLP constructed in [22] were used for modeling in real financial markets in [6], [14], [26]; a generalization of this family was constructed in [9], [11], [12]. As A. N. Shiryaev and O. E. Barndorff-Nielsen remarked, the name TLP was misleading, and thus from now on we will call this family of processes the KoBoL family.

Earlier, noninfinitely divisible truncations of stable Lévy distributions were constructed and used to model the behavior of the Standard & Poor 500 Index by Mantegna and Stanley [24], [25].

Notice that the Lévy measure of any Lévy process can be approximated by a sequence of Lévy measures of RLPE so that the factors in the Wiener–Hopf factorization formula also converge, and in the case of payoffs of the form (1.3), the answers and conditions are formulated in terms of these factors. Hence, for these payoffs, our results are valid for any Lévy process. Whether they are valid for any Lévy process when payoffs are more general than (1.3) remains an open question.

As is well known, simple formulas for the factors in the Wiener–Hopf factorization formula can be obtained in few cases only. Here we obtain them (in two versions) by using only one integration, and for model classes HP, NIG, and the KoBoL family we derive really simple approximate formulas, with small errors if the rate of decay of the tails is large. As empirical studies in [3] and [26] suggest, this is usually the case, and so these approximate formulas may be of some interest.

The plan of the paper is as follows. In section 2, we introduce the class RLPE, give examples, and prove several properties of the characteristic exponents of RLPE. In section 3, we derive two sets of explicit formulas for factors in the Wiener–Hopf factorization formula and necessary bounds for these factors; for model classes of RLPE, we also obtain approximate effective formulas for the factors.

In section 4 (resp., section 5), we solve the problem for the perpetual American put (resp., call) and similar more general payoffs, in the class $\mathcal{M}_0$. In section 6, we formulate the free boundary value problem, prove that its solution solves the optimal stopping problem in the class $\mathcal{M}$, and for model classes of RLPE and mixtures of independent processes of the model classes, verify that the explicit solutions found in sections 4–5 for puts, calls, and some other options with the payoffs of the form (1.3) solve the free boundary value problem and, hence, solve problem (1.1).

In section 7 we show that in some cases the smooth pasting condition fails, and we offer its generalization, which is valid for RLPE; in the appendix, we prove the most technical statements of sections 2–6.

2. RLPEs.

2.1. Some basic facts about Lévy processes (See, e.g., [32, section I.4], [5, pp. 3, 13], and [33, p. 3]). We assume as given a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{0\le t<+\infty}, \mathbf{P})$ satisfying the usual hypotheses.

DEFINITION 2.1. *An adapted process $X = (X_t)_{0\le t<+\infty}$ with $X_0 = 0$ a.s. is a Lévy process if and only if*

(i) *X has increments independent of the past, i.e., $X_t - X_s$ is independent of $\mathcal{F}_s, 0 \le s < t < +\infty$;*

(ii) *X has stationary increments, i.e., $X_t - X_s$ has the same distribution as $X_{t-s}, 0 \le s < t < +\infty$;*

(iii) *X is continuous in probability.*

There exists a nice formula (the Lévy–Khintchine formula) which explicitly describes a Lévy process in terms of its characteristic exponent, ψ, defined by $E[e^{i\xi X_t}] = e^{-t\psi(\xi)}$. Since we consider only one-dimensional Lévy processes here, we formulate the corresponding theorem in the one-dimensional case only.

THEOREM 2.2. (a) *Let X be a Lévy process on $\mathbf{R}$. Then its characteristic exponent admits the representation*

$$\psi(\xi) = \frac{\sigma^2}{2}\xi^2 - i\gamma\xi - \int_{\mathbf{R}^n} \Big(e^{ix\xi} - 1 - ix\xi\mathbf{1}_{[-1,1]}(x)\Big)\Pi(dx), \tag{2.1}$$

where $\sigma \ge 0$, $\gamma \in \mathbf{R}$, and Π is a measure supported on $\mathbf{R}\setminus\{0\}$ that satisfies

$$\Pi(\{0\}) = 0, \qquad \int_{-\infty}^{+\infty} (|x|^2 \wedge 1)\Pi(dx) < \infty. \tag{2.2}$$

(b) *The representation* (2.1) *is unique.*

(c) *Conversely, if $\sigma \ge 0$, $\gamma \in \mathbf{R}$, and Π is a measure supported on $\mathbf{R}\setminus\{0\}$ that satisfies* (2.2)*, then there exists a Lévy process X with the characteristic exponent defined by* (2.1)*; X is uniquely defined in law.*

The triple (σ^2, Π, γ) is called the *generating triplet* of X. The σ^2 and Π are called the *Gaussian coefficient* and *Lévy measure* of X. When $\Pi = 0$, X is Gaussian, and if $\sigma = 0$, then X is called purely non-Gaussian.

The infinitesimal generator, L, of a Lévy process X acts as follows:

$$Lf(x) = \frac{\sigma^2}{2}f''(x) + \gamma f'(x) + \int_{-\infty}^{+\infty} (f(x+y) - f(x) - yf'(x)\mathbf{1}_{[-1,1]}(y))\Pi(dy). \tag{2.3}$$

Apply $-L$ to an oscillating exponent $f(x) = e^{ix\xi}$ and use (2.3):

$$(-L)e^{ix\xi} = \left[\frac{\sigma^2}{2}\xi^2 - i\gamma\xi - \int_{-\infty}^{+\infty} \Big(e^{iy\xi} - 1 - iy\xi\mathbf{1}_{(-1,1)}(y)\Big)\Pi(dy)\right] e^{ix\xi} = \psi(\xi)e^{ix\xi}. \tag{2.4}$$

By decomposing a sufficiently regular function u into the Fourier integral

$$u(x) = (2\pi)^{-1}\int_{-\infty}^{+\infty} e^{ix\xi}\hat{u}(\xi)d\xi,$$

where

$$\hat{u}(\xi) = \int_{-\infty}^{+\infty} e^{-ix\xi}u(x)dx \tag{2.5}$$

is the Fourier transform of a function u (this is the standard definition in the literature on PDOs), and using (2.4), we conclude that $-L$ is a pseudodifferential operator with the symbol $\psi(\xi)$:

$$-L = \psi(D).$$

Recall that a pseudodifferential operator with the (constant) symbol a is defined by

$$a(D)u(x) = (2\pi)^{-1} \int_{-\infty}^{+\infty} e^{ix\xi} a(\xi)\hat{u}(\xi)d\xi.$$

For an introduction to the general theory of PDOs, see [20].

2.2. Lévy processes of exponential type.

DEFINITION 2.3. *Let $\lambda_- < 0 < \lambda_+$. We call X a Lévy process of exponential type $[\lambda_-, \lambda_+]$ if its Lévy measure satisfies*

$$\int_{-\infty}^{-1} e^{-\lambda_+ x}\Pi(dx) + \int_{1}^{+\infty} e^{-\lambda_- x}\Pi(dx) < \infty. \tag{2.6}$$

LEMMA 2.4. *Let X be a Lévy process of exponential type $[\lambda_-, \lambda_+]$. Then*

(a) *the characteristic exponent ψ is holomorphic in the strip $\Im\xi \in (\lambda_-, \lambda_+)$ and continuous up to the boundary of the strip;*

(b) *there exist C and $\nu > 0$ such that for all ξ in the strip $\Im\xi \in [\lambda_-, \lambda_+]$*

$$|\psi(\xi)| \le C(1 + |\xi|)^\nu; \tag{2.7}$$

(c) *for any $q > 0$ there exist $\delta > 0$ and $\sigma_- < 0 < \sigma_+$ such that for any $[\omega_-, \omega_+] \subset (\sigma_-, \sigma_+)$ and all ξ in the strip $\Im\xi \in [\omega_-, \omega_+]$*

$$q + \Re\psi(\xi) \ge \delta, \tag{2.8}$$

where $\delta = \delta(\omega_-, \omega_+) > 0$;

(d) *if*

$$q + \psi(i(\sigma_- + 0)) \ge 0, \quad \text{and} \quad q + \psi(i(\sigma_+ - 0)) \ge 0, \tag{2.9}$$

then (2.8) holds;

(e) *for any $q > 0$, the equation*

$$q + \psi(\xi) = 0 \tag{2.10}$$

has at most one purely imaginary root in the lower half-plane (call it $-i\beta_+$) and at most one, $-i\beta_-$, in the upper half-plane;

(f) *the root $-i\beta_\mp$ exists if and only if*

$$q + \psi(i\lambda_\pm \mp 0) < 0, \tag{2.11}$$

and if it exists, it is a simple root.

Proof. (a) is immediate from (2.6), and (b) can be easily deduced from the Lévy–Khintchine formula, by considering separately the integral over $|x| \le |\xi|^{-1}$ and $|x| \ge |\xi|^{-1}$.

(c)–(d) Set $M_1(\sigma) = \int_{-\infty}^{+\infty} e^{-\sigma x}\mu^1(dx)$, where $\mu^1(dx)$ is the probability distribution of X_1. By differentiating twice, we conclude that M_1 is convex and, clearly,

$M_1(0) = 1 < e^q$. Hence, there exist $\omega_- < 0 < \omega_+$ and $\delta > 0$ such that for all $\sigma \in [\omega_-, \omega_+]$, $M_1(\sigma) \le e^{q-\delta}$.

Now, for any $\xi \in \mathbf{R}$ and these σ,

$$\begin{aligned}\exp(-\Re\psi(\xi + i\sigma)) &= |\exp(-\psi(\xi + i\sigma))| \\ &= \left|\int_{-\infty}^{+\infty} e^{i\xi x - \sigma x}\mu^1(dx)\right| \le \int_{-\infty}^{+\infty} e^{-\sigma x}\mu^1(dx);\end{aligned}$$

therefore (2.8) holds with $\sigma_- = \inf\omega_-, \sigma_+ = \sup\omega_+$, and (2.9) implies (2.8).

(e)–(f) Notice that by the proof of (c), $\sigma \mapsto q + \psi(i\sigma)$ is concave and equal to $q > 0$ at 0. □

2.3. Two definitions of RLPEs. For the sake of brevity, we consider processes with Lévy measures (almost) symmetric in a neighborhood of the origin.

DEFINITION 2.5. *Let $\lambda_- < 0 < \lambda_+$ and $\nu \in [0, 2)$. A purely non-Gaussian Lévy process is called an RLPE of type $[\lambda_-, \lambda_+]$ and order ν if its Lévy measure satisfies* (2.6) *and, in a neighborhood of zero, admits a representation $\Pi(dx) = f(x)dx$, where f satisfies the following condition: There exist $\nu' < \nu$, $c > 0$, and $C > 0$ such that*

$$\left|f(x) - c|x|^{-\nu-1}\right| \le C|x|^{-\nu'-1} \qquad \forall\ |x| \le 1. \tag{2.12}$$

If the sample paths of a Lévy process have bounded variation on every compact time interval a.s., one says that the Lévy process has bounded variation. A regular Lévy process of exponential type has bounded variation if and only if $\nu < 1$, since this is equivalent to $\int_{-\infty}^{+\infty}(|x| \wedge 1)\Pi(dx) < +\infty$ (see, e.g., [5], p. 15).

Straightforward calculation (see [9]) shows that an RLPE of order $\nu > 0$ in the sense of Definition 2.5 is an RLPE in the sense of the following definition.

DEFINITION 2.6. *Let $\lambda_- < 0 < \lambda_+$ and $\nu \in (0, 2]$. A Lévy process is called an RLPE of type $[\lambda_-, \lambda_+]$ and order $\nu > 0$ if the following two conditions are satisfied:*

(i) *the characteristic exponent admits a representation*

$$\psi(\xi) = -i\mu\xi + \phi(\xi), \tag{2.13}$$

where ϕ is holomorphic in the strip $\Im\xi \in (\lambda_-, \lambda_+)$, is continuous up to the boundary of the strip, and admits a representation

$$\phi(\xi) = c|\xi|^\nu + O(|\xi|^{\nu_1}) \tag{2.14}$$

as $\xi \to \infty$ in the strip $\Im\xi \in [\lambda_-, \lambda_+]$, where $\nu_1 < \nu$;

(ii) *there exist $\nu_2 < \nu$ and C such that the derivative of ϕ in* (2.13) *admits a bound*

$$|\phi'(\xi)| \le C(1 + |\xi|)^{\nu_2}, \qquad \Im\xi \in [\lambda_-, \lambda_+]. \tag{2.15}$$

One can easily generalize both definitions by using $c_\pm \ge 0$ in (2.12) on the half-axis $\pm x > 0$, and in (2.14) as $\Re\xi \to \pm\infty$.

2.4. Model classes of RLPEs. All model classes listed in the introduction except for VGP are RLPE:

- BM are RLPE of order 2 and any exponential type;

- a KoBoL process of order $\nu \in (0,2)$, with steepness parameters $\lambda_- < 0$ and $\lambda_+ > 0$, is an RLPE of order ν and exponential type $[\lambda_-, \lambda_+]$. An (asymmetric) version can be defined as a purely non-Gaussian Lévy process with the Lévy measure

$$\Pi(dx) = c_+ x_+^{-\nu-1} e^{\lambda_- x} dx + c_- x_-^{-\nu-1} e^{\lambda_+ x} dx, \tag{2.16}$$

where $x_+ = \max\{x, 0\}$ and $c_\pm > 0$. An RLPE in the sense of Definitions 2.5–2.6 holds with $c_+ = c_-$. Direct calculation shows that if $\nu \in (0,2)$, $\nu \neq 1$, and $c_+ = c_- = c$, then the characteristic exponent of a KoBoL process is of the form

$$\psi(\xi) = -i\mu\xi + c\Gamma(-\nu)[\lambda_+^\nu - (\lambda_+ + i\xi)^\nu + (-\lambda_-)^\nu - (-\lambda_- - i\xi)^\nu]. \tag{2.17}$$

In the case $\nu = 1$, the formula differs from (2.17) (see [9], [12]).

- a normal tempered stable Lévy process of order $\nu \in (0,2)$, with parameters $\delta > 0$, $\alpha > \beta > -\alpha$, is an RLPE of order ν and exponential type $[-\alpha+\beta, \alpha+\beta]$; in particular, NIG processes are RLPE of order 1. The characteristic exponent is of the form

$$\psi(\xi) = -i\mu\xi + \delta[(\alpha^2 - (\beta + i\xi)^2)^{\nu/2} - (\alpha^2 - \beta^2)^{\nu/2}]; \tag{2.18}$$

- an HP with parameters $\delta > 0$, $\alpha > \beta > -\alpha$, is an RLPE of order 1 and exponential type $[\lambda_-, \lambda_+]$ for any $[\lambda_-, \lambda_+] \subset (-\alpha+\beta, \alpha+\beta)$. Its characteristic exponent is equal to

$$\psi(\xi) = -i\mu\xi - \ln\left[\left(\frac{\alpha\delta}{K_1(\alpha\delta)}\right)\frac{K_1(\delta\sqrt{\alpha^2 - (\beta+i\xi)^2})}{\delta\sqrt{\alpha^2-\beta^2}}\right]. \tag{2.19}$$

2.5. Properties of the characteristic exponent of an RLPE.

2.5.1. General properties. Clearly, an RLPE is a Lévy process of exponential type; therefore the properties listed in Lemma 2.4 hold for any RLPE.

2.5.2. Additional properties of the characteristic exponents from model classes. In order to derive simple approximate formulas for the factors in the Wiener–Hopf factorization formula, we need the following lemma, which we managed to prove only for model classes on a case-by-case basis. We conjecture that this lemma holds for a much wider variety of RLPE, if not for all RLPE.

LEMMA 2.7. *Let X be one of the model processes, of order $\nu > 0$ and exponential type $[\lambda_-, \lambda_+]$. Then*

(a) *the ϕ in* (2.13) *admits the analytic continuation into the complex plane with two cuts, $(-i\infty, i\lambda_-]$ and $[i\lambda_+, +i\infty)$, and outside any neighborhood of $i\lambda_-$ and $i\lambda_+$ it satisfies the following estimate:*

$$|\phi(\xi)| \le C(1 + |\xi|)^\nu; \tag{2.20}$$

(b) *all the roots in the plane with the cuts are purely imaginary.*

For the proof, see the appendix. □

3. The Wiener–Hopf factorization.

3.1. General Lévy processes. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space on which a one-dimensional Lévy process X is defined, and let Ω_0 be a subset of Ω such that for each $\omega \in \Omega_0$ the trajectory $X_\cdot(\omega)$ is right-continuous with left limits. Define, on Ω_0, $M_t = \sup_{0\le s\le t} X_s$ and $N_t = \inf_{0\le s\le t} X_s$. On $\Omega \setminus \Omega_0$, both M_t and N_t are set to be 0. $M = \{M_t\}$ and $N = \{N_t\}$ are called the supremum process and the infimum process, respectively. The Laplace transform (in t) of the distribution of X_t, or more precisely,

$$qE^x\left[\int_0^\infty e^{-qt}e^{i\xi X_t}dt\right] = q(q+\psi(\xi))^{-1},$$

can be factorized by using the Laplace transforms (in t) of the distributions of the supremum and infimum processes. Among many factorization identities, we will use only the simplest one ([33, Theorems 45.2 and 45.5]; for more detailed exposition, see [33, section 45]).

Let μ^t be the law of X.

THEOREM 3.1. (a) *Let $q > 0$. There exists a unique pair of infinitely divisible distributions μ_q^+ and μ_q^- supported on $(-\infty, 0]$ and $[0, +\infty)$, respectively, such that their Fourier transforms ϕ_q^+ and ϕ_q^- satisfy*

$$q(q+\psi(\xi))^{-1} = \phi_q^+(\xi)\phi_q^-(\xi), \quad \xi \in \mathbf{R}. \tag{3.1}$$

(b) *The functions ϕ_q^+ and ϕ_q^- admit the following representations:*

$$\phi_q^+(\xi) = q\int_0^{+\infty} e^{-qt}E[e^{i\xi M_t}]dt = q\int_0^{+\infty} e^{-qt}E[e^{i\xi(X_t-N_t)}]dt, \tag{3.2}$$

$$\phi_q^-(\xi) = q\int_0^{+\infty} e^{-qt}E[e^{i\xi N_t}]dt = q\int_0^{+\infty} e^{-qt}E[e^{i\xi(X_t-M_t)}]dt \tag{3.3}$$

and

$$\phi_q^+(\xi) = \exp\left[\int_0^{+\infty} t^{-1}e^{-qt}dt\int_0^{+\infty}(e^{ix\xi}-1)\mu^t(dx)\right], \tag{3.4}$$

$$\phi_q^-(\xi) = \exp\left[\int_0^{+\infty} t^{-1}e^{-qt}dt\int_{-\infty}^{0}(e^{ix\xi}-1)\mu^t(dx)\right]. \tag{3.5}$$

Notice that $\phi_q^+(\xi)$ (resp., $\phi_q^-(\xi)$) admits the analytic continuation into the upper half-plane $\Im\xi > 0$ (resp., lower half-plane $\Im\xi < 0$) and does not vanish there. Thus, (3.1) is a special case of the *Wiener–Hopf factorization* introduced in solving integral equations by Wiener and Hopf in 1931 [35], and widely used in the theory of boundary value problems for PDE and PDO.

The formulas (3.2)–(3.3) are by no means explicit, though very convenient for theoretical considerations, and (3.4)–(3.5) are rather involved. Simple explicit analytical formulas can be obtained for special cases only.

Example 3.1. Let X be a BM with the drift γ and variance σ^2. Then the characteristic exponent is $\psi(\xi) = \frac{\sigma^2}{2}\xi^2 - i\gamma\xi$. It is clear that for $q > 0$ the equation $q+\psi(\xi) = 0$ has two roots $-i\beta_-$ and $-i\beta_+$ in the upper and lower half-planes, respectively, and therefore, $q(q+\psi(\xi))^{-1}$ admits the factorization (3.1) with

$$\phi_q^+(\xi) = \frac{\beta_+}{\beta_+ - i\xi}, \qquad \phi_q^-(\xi) = \frac{-\beta_-}{-\beta_- + i\xi}. \tag{3.6}$$

Clearly, ϕ_q^- is the Fourier transform of the exponential distribution with parameter $-\beta_-$, and ϕ_q^+ is the Fourier transform of the dual to the exponential distribution with parameter β_+.

3.2. Lévy processes of exponential type. We fix a branch of ln by the requirement $\ln a \in \mathbf{R}$ for $a > 0$. We also fix $\omega_- < 0 < \omega_+$, for which (2.8) holds.

THEOREM 3.2. *Let X be a Lévy process of exponential type, let there exist $C, c, \nu > 0$ such that*

$$\Re\psi(\xi) \geq c(1+|\xi|)^\nu, \qquad \Im\xi \in [\omega_-, \omega_+], \tag{3.7}$$

and let there exist $q > 0$ such that for $\omega = \omega_\pm$

$$\int_{-\infty+i\omega}^{+\infty+i\omega} \frac{|\psi'(\eta)|}{(1+|\eta|)(q+\Re\psi(\eta))} d\eta < +\infty. \tag{3.8}$$

Then (a) *$\phi_q^+(\xi)$ admits the analytic continuation into a half-plane $\Im\xi > \omega_-$ and can be calculated as follows:*

$$\phi_q^+(\xi) = \exp\left[(2\pi i)^{-1} \int_{-\infty+i\omega_-}^{+\infty+i\omega_-} \frac{\psi'(\eta)}{q+\psi(\eta)} \ln\frac{\eta-\xi}{\eta} d\eta\right] \tag{3.9}$$

$$= \exp\left[(2\pi i)^{-1} \int_{-\infty+i\omega_-}^{+\infty+i\omega_-} \frac{\xi \ln(q+\psi(\eta))}{\eta(\xi-\eta)} d\eta\right]; \tag{3.10}$$

(b) *$\phi_q^-(\xi)$ admits the analytic continuation into a half-plane $\Im\xi < \omega_+$ and can be calculated as follows:*

$$\phi_q^-(\xi) = \exp\left[-(2\pi i)^{-1} \int_{-\infty+i\omega_+}^{+\infty+i\omega_+} \frac{\psi'(\eta)}{q+\psi(\eta)} \ln\frac{\eta-\xi}{\eta} d\eta\right] \tag{3.11}$$

$$= \exp\left[-(2\pi i)^{-1} \int_{-\infty+i\omega_+}^{+\infty+i\omega_+} \frac{\xi \ln(q+\psi(\eta))}{\eta(\xi-\eta)} d\eta\right]; \tag{3.12}$$

(c) *$\phi_q^+(\xi)^{-1}$ (resp., $\phi_q^-(\xi)^{-1}$) admits the analytic continuation into a wider half-plane $\Im\xi > \lambda_-$ (resp., $\Im\xi < \lambda_+$) by*

$$\phi_q^+(\xi)^{-1} = q^{-1}(q+\psi(\xi))\phi_q^-(\xi), \quad \Im\xi \in (\lambda_-, \omega_-] \tag{3.13}$$

$$\phi_q^-(\xi)^{-1} = q^{-1}(q+\psi(\xi))\phi_q^+(\xi), \quad \Im\xi \in [\omega_+, \lambda_+). \tag{3.14}$$

Proof. (a) Consider the expression under the exponent sign in (3.4):

$$\begin{aligned} f(\xi) :&= \int_0^{+\infty} \frac{e^{-qt}}{t} \int_0^{+\infty} (e^{ix\xi} - 1)\mu^t(dx)dt \\ &= \int_0^{+\infty} \frac{e^{-qt}}{t} \int_0^{+\infty} (e^{ix\xi} - 1)(2\pi)^{-1} \int_{-\infty}^{+\infty} e^{-ix\eta - t\psi(\eta)} d\eta dx dt. \end{aligned}$$

On the strength of (3.7), we may apply the Cauchy theorem and shift the line of integration:

$$f(\xi) = \int_0^{+\infty} \frac{e^{-qt}}{t} \int_0^{+\infty} (e^{ix\xi} - 1)(2\pi)^{-1} \int_{-\infty+i\omega_-}^{+\infty+i\omega_-} e^{-ix\eta - t\psi(\eta)} d\eta dx dt.$$

Now the inner double integral converges absolutely; hence we can apply the Fubini theorem and integrate w.r.t. x first:

$$f(\xi)=\int_0^{+\infty}\frac{e^{-qt}}{t}(2\pi i)^{-1}\int_{-\infty+i\omega_-}^{+\infty+i\omega_-}e^{-t\psi(\eta)}((\eta-\xi)^{-1}-\eta^{-1})d\eta dt.$$

Integrate by parts:

$$\begin{aligned}f(\xi)&=\int_0^{+\infty}\frac{e^{-qt}}{t}(2\pi i)^{-1}\int_{-\infty+i\omega_-}^{+\infty+i\omega_-}\ln\frac{\eta-\xi}{\eta}t\psi'(\eta)e^{-t\psi(\eta)}d\eta dt\\&=(2\pi i)^{-1}\int_0^{+\infty}\int_{-\infty+i\omega_-}^{+\infty+i\omega_-}\ln\frac{\eta-\xi}{\eta}\psi'(\eta)e^{-t(q+\psi(\eta))}d\eta dt.\end{aligned}$$

From (3.8), the integral above, calculated in the reverse order $dtd\eta$, converges absolutely. Hence we can apply the Fubini theorem once again and obtain (3.9); integrating in (3.9) by parts, we arrive at (3.10).

(b) The dual process $\tilde X$ is of exponential type $[-\lambda_+,-\lambda_-]$, its characteristic exponent is $\bar\psi$, and $[-\omega_+,-\omega_-]$ plays the part of $[\omega_-,\omega_+]$ in Lemma 2.4. Write down the Wiener–Hopf factorization for $\tilde X$ and apply the complex conjugation; then the "+"-factor for $\tilde X$ becomes the "−"-factor for X, and (3.9) for $\tilde X$ becomes (3.11) for X.

(c) This part follows from (3.1) and Lemma 2.4(a). □

Remark 3.1. If X is an RLPE in the sense of Definition 2.6, then (3.8) and (3.7) hold; hence, Theorem 3.2 holds as well.

LEMMA 3.3. *Let ω_- and ω_+ be as in Theorem* 3.2. *Then there exists $C>0$ such that in the half-plane $\pm\Im\xi\ge\pm\omega_\mp$, $\phi_q^\pm$ admits estimates*

$$(1+|\xi|)^{-C}\le|\phi_q^\pm(\xi)|\le(1+|\xi|)^C. \tag{3.15}$$

Proof. In (3.9) and (3.11), make the change of variables $\eta\mapsto|\xi|\eta$ and use (2.7) to notice that the expressions under the exponential sign admit an estimate via $C\ln(2+|\xi|)$. □

Equation (3.15) is insufficient for the proofs in sections 4–6. More information about properties of the factors is obtained below.

3.3. RLPE. Let $\sigma_-<0<\sigma_+$ be from (2.8). Fix $\lambda>\max\{-\sigma_-,\sigma_+\}$, and set $\Lambda_\pm(\xi)^s=(\lambda\mp i\xi)^s=\exp[s\ln(\lambda\mp i\xi)]$. Next, choose $d>0$ and $\kappa_-,\kappa_+\in\mathbf{R}$ so that

$$B(\xi):=d^{-1}\Lambda_+(\xi)^{-\kappa_+}\Lambda_-(\xi)^{-\kappa_-}(q+\psi(\xi)) \tag{3.16}$$

satisfies

$$\lim_{\xi\to\pm\infty}B(\xi)=1. \tag{3.17}$$

Choices of d, κ_+, and κ_- depend on the properties of ψ, and hence on ν,μ, and c in (2.13)–(2.14). We have to consider four cases.

1. If $\nu\in(1,2)$ or $\nu\in(0,1]$ and $\mu=0$, we set $d=c$, $\kappa_+=\kappa_-=\nu/2$.
2. If $\nu\in(0,1)$ and $\mu>0$, we set $d=\mu$, $\kappa_+=1$, $\kappa_-=0$.
3. If $\nu\in(0,1)$ and $\mu<0$, we set $d=|\mu|$, $\kappa_+=0$, $\kappa_-=1$.
4. If $\nu=1$, we set $d=(c^2+\mu^2)^{1/2}$, $\kappa_\pm=1/2\pm\pi^{-1}\arctan(\mu/c)$.

In all cases, (3.17) follows from (2.13)–(2.14). In the first three cases, (3.17) is immediate, and in the last case, the simplest way is to check that $\ln B(\xi) \to 0$ as $\xi \to \pm\infty$:

$$\lim_{\xi\to\pm\infty} \ln B(\xi) = \pm\frac{\pi i}{2}\kappa_+ \mp \frac{\pi i}{2}\kappa_- + \ln\frac{c \mp i\mu}{(c^2+\mu^2)^{1/2}} + (-\kappa_+ - \kappa_- + 1)\ln|k|$$
$$= \pm(\kappa_+ - \kappa_-)\frac{\pi i}{2} \mp i \arctan\frac{\mu}{c} = 0$$

by our choice of κ_+ and κ_-.

The last factor in (3.16) assumes values in a half-plane $\Re z > 0$ by (2.8), and the same is true of the product of the first three factors, since the first one is positive, $\Lambda_-(\xi)$ and $\Lambda_+(\xi)$ assume values in the half-plane but in different quadrants, and $0 \le \kappa_\pm \le 1$. Hence, for all $\xi \in \mathbf{R}$, $-\pi < \arg B(\xi) < \pi$, and therefore $b = \ln B$ is well-defined on $\mathbf{R}$. Fix $\omega_- < 0 < \omega_+$ such that $\sigma_- < \omega_-, \omega_+ < \sigma_+$, where $\sigma_\pm$ are from (2.8), and notice that all the arguments above are valid on any line $\Im\xi = \sigma \in [\omega_-, \omega_+]$.

Next, for $\tau > \omega_-$, $\tau_1 \in [\omega_-, \tau)$ and real ξ, set

$$b_+(\xi + i\tau) = \frac{i}{2\pi}\int_{-\infty+i\tau_1}^{+\infty+i\tau_1} \frac{b(\eta)}{\xi + i\tau - \eta}d\eta \tag{3.18}$$

(by the Cauchy theorem, $b_+(\eta + i\tau)$ is independent of a choice of τ_1), and for $\tau < \omega_+$, $\tau_2 \in (\tau, \omega_+]$, and real ξ, set

$$b_-(\xi + i\tau) = -\frac{i}{2\pi}\int_{-\infty+i\tau_2}^{+\infty+i\tau_2} \frac{b(\eta)}{\xi + i\tau - \eta}d\eta. \tag{3.19}$$

It follows from (2.13), (2.14), (3.16), and (3.17) that there exist $C_1, \rho > 0$ such that for any η in a strip $\Im\eta \in [\omega_-, \omega_+]$

$$|b(\eta)| \le C_1(1 + |\eta|)^{-\rho}. \tag{3.20}$$

Hence, the integrals in (3.18)–(3.19) converge, and $b_+(\xi)$ (resp., $b_-(\xi)$) is well-defined in a half-plane $\Im\xi\ \omega_-$ (resp., $\Im\xi < \omega_+$). We set $a_\pm(\xi) = \Lambda_\pm(\xi)^{\kappa_\pm}\exp b_\pm(\xi)$.

THEOREM 3.4. (a) *a_+ (resp., a_-) is holomorphic in a half-plane $\Im\xi > \omega_-$ (resp., $\Im\xi < \omega_+$). It admits the analytic continuation into a wider half-plane $\Im\xi > \lambda_-$ (resp., $\Im\xi < \lambda_+$) and the continuous extension up to the boundary, by*

$$a_+(\xi) = a(\xi)/a_-(\xi), \quad \Im\xi \in [\lambda_-, \omega_-],$$

$$a_-(\xi) = a(\xi)/a_+(\xi), \quad \Im\xi \in [\omega_+, \lambda_+],$$

where $a(\xi) = d^{-1}(q + \psi(\xi))$;

(b) *on a strip $\Im\xi \in [\lambda_-, \lambda_+]$,*

$$q + \psi(\xi) = da_+(\xi)a_-(\xi); \tag{3.21}$$

(c) *there exist C, $c > 0$, and $\rho_1 > 0$ such that in a half-plane $\Im\xi \ge \omega_-$,*

$$c(1 + |\xi|)^{\kappa_+} \le |a_+(\xi)| \le C(1 + |\xi|)^{\kappa_+}, \tag{3.22}$$

$$|a_+(\xi)^{\pm 1} - \Lambda_+(\xi)^{\pm\kappa_+}| \le C(1 + |\xi|)^{\pm\kappa_+ - \rho_1}, \tag{3.23}$$

and in a half-plane $\Im\xi \le \omega_+$,

$$c(1+|\xi|)^{\kappa_-} \le |a_-(\xi)| \le C(1+|\xi|)^{\kappa_-}, \tag{3.24}$$

$$|a_-(\xi)^{\pm 1} - \Lambda_-(\xi)^{\pm\kappa_-}| \le C(1+|\xi|)^{\pm\kappa_- - \rho_1}; \tag{3.25}$$

(d) *factors in* (3.1) *and* (3.21) *are related by*

$$\phi_q^{\pm}(\xi)^{-1} = a_\pm(\xi)/a_\pm(0). \tag{3.26}$$

Proof. (a) The first statement is straightforward from (3.20), and once (c) is proven, the second one follows, since $a(\xi)$ is holomorphic on a strip $\Im\xi \in (\lambda_-, \lambda_+)$ and admits the continuous extension up to the boundary of the strip.

(b) By the residue theorem, we have for $\tau_1 \in (\omega_-, \Im\xi)$ and $\tau_2 \in (\Im\xi, \omega_+)$

$$b_+(\xi) = \frac{i}{2\pi}\left(\int_{-\infty+i\tau_1}^{+\infty+i\tau_1} - \int_{-\infty+i\tau_2}^{+\infty+i\tau_2}\right)\frac{b(\eta)}{\xi-\eta}d\eta + \frac{i}{2\pi}\int_{-\infty+i\tau_2}^{+\infty+i\tau_2}\frac{b(\eta)}{\xi-\eta}d\eta = b(\xi) - b_-(\xi).$$

Hence, $\exp b_+(\xi)\exp b_-(\xi) = B(\xi)$, and (3.21) is immediate on a narrow strip $\omega_- < \Im\xi < \omega_+$; on a wider strip $\Im\xi \in [\lambda_-, \lambda_+]$, it holds by construction.

(c) By using (3.20), we obtain

$$|(\xi + i\tau - \eta)^{-1} b(\eta)| \le C(1+|\xi-\eta|)^{-1}(1+|\eta|)^{-\rho}.$$

By separately considering a region, where $|\xi-\eta| \ge |\xi|/2$, and its complement, it is easy to show that the right-hand side (RHS) admits an upper bound via

$$C_1(1+|\xi|)^{-\rho_1}(1+|\eta|)^{-1-\rho_1} + C_1(1+|\xi|)^{-\rho_1}(1+|\xi-\eta|)^{-1-\rho_1},$$

where $\rho_1 = \min\{1,\rho\}/2 > 0$. By integrating, we obtain for ξ in a half-plane $\pm\Im\xi \ge \omega_\mp$ (see (3.18)–(3.19))

$$|b_\pm(\xi)| \le C_3(1+|\xi|)^{-\rho_1}, \tag{3.27}$$

and (3.22)–(3.25) follow from (3.27) and the definition of $a_\pm$.

(d) Notice that $a_\pm$, $1/a_\pm$, $\phi_q^\pm$, and $1/\phi_q^\pm$ are bounded by a polynomial in the half-plane $\pm\Im\xi \ge \pm\omega_\mp$; therefore, by comparing (3.1) and (3.21), we conclude that $a_\pm\phi_q^\pm$ is holomorphic, polynomially bounded, and nonvanishing on the complex plane. By the Liouville theorem, this is constant, and taking into account that $\phi_q^\pm(0) = 1$, we obtain (3.26). □

3.4. Approximate formulas for the factors in the case of NIG, HP, and KoBoL. We can write these formulas down for both representations (in (3.1) and (3.21)). In the case of the former, the argument and formulas are shorter. We use (3.9) and Lemma 2.7 to transform the line of integration into the contour along the banks of the cut $(-i\infty, i\lambda_-]$. In empirical studies (see, e.g., [3] and [26]), the λ_+ and $-\lambda_-$ are usually large, of order 40–50, and then for typical values of other parameters both roots $-i\beta_\pm$ in Lemma 2.7 exist. Therefore, in the process of transformation, the contour crosses the simple pole at $\eta = -i\beta_+$. By the residue theorem we obtain, for ξ in the upper half-plane,

$$\phi_q^+(\xi) = \exp\left[\ln\frac{-i\beta_+}{-i\beta_+ - \xi} + \Phi_q^+(\xi)\right],$$

where

$$\Phi_q^+(\xi) = (2\pi)^{-1} \int_{-\infty}^{\lambda_-} \left[\frac{\psi'(iz-0)}{q+\psi(iz-0)} - \frac{\psi'(iz+0)}{q+\psi(iz+0)} \right] \ln \frac{-z-i\xi}{-z} dz. \tag{3.28}$$

Thus,

$$\phi_q^+(\xi) = \frac{\beta_+}{\beta_+ - i\xi} \exp \Phi_q^+(\xi). \tag{3.29}$$

Similarly, from (3.11) we deduce, for ξ in the lower half-plane,

$$\phi_q^-(\xi) = \frac{-\beta_-}{-\beta_- + i\xi} \exp \Phi_q^-(\xi), \tag{3.30}$$

where

$$\Phi_q^-(\xi) = (2\pi)^{-1} \int_{\lambda_+}^{+\infty} \left[\frac{\psi'(iz-0)}{q+\psi(iz-0)} - \frac{\psi'(iz+0)}{q+\psi(iz+0)} \right] \ln \frac{z+i\xi}{z} dz. \tag{3.31}$$

If $-\lambda_-$ (resp., λ_+) is large, then $|\Phi_q^+(\xi)|$ (resp., $|\Phi_q^+(\xi)|$) is small uniformly in ξ in the upper (resp., lower) half-plane, which can be easily seen from the explicit formulas for the characteristic exponents and (3.28) (resp., (3.31)). Hence, we may calculate the integrals in (3.28) and (3.31) with large relative error and still obtain $\phi_q^+(\xi)$ from (3.29) and $\phi_q^-(\xi)$ from (3.30) with good accuracy. This observation can be used to develop effective numerical procedures. In fact, even the simple approximations

$$\phi_q^+(\xi) \sim \frac{\beta_+}{\beta_+ - i\xi}, \qquad \phi_q^-(\xi) \sim \frac{-\beta_-}{-\beta_- + i\xi} \tag{3.32}$$

produce errors of only several percent for many typical parameters values.

The comparison of (3.6) and (3.32) provides an analytical explanation of why a simple adjustment of parameters of the Gaussian model can give fairly good fit even in a very non-Gaussian situation.

4. Pricing of the perpetual American put and similar perpetual options.

4.1. Sufficient conditions for the solution for the perpetual put-like options, in the class $\mathcal{M}_0$ of hitting times $\tau(a)$ of segments $(-\infty, a]$. Let $\mathbf{Q}$ be an EMM chosen by the market, and assume that X under $\mathbf{Q}$ is an RLPE with the characteristic exponent ψ and the infinitesimal generator L. For $g(X_t)$ the payoff, set

$$V(h,x) := E^x[e^{-q\tau(h)} g(X_{\tau(h)})], \tag{4.1}$$

where E^x is the expectation operator of the process X started at x, under $\mathbf{Q}$.

LEMMA 4.1. *Let there exist h_* with the following properties:*

(a) *if $h < h_*$, then there exists x such that*

$$V(h,x) < g(x); \tag{4.2}$$

(b) *for any $x \geq h_*$,*

$$V(h^*,x) \geq g(x); \tag{4.3}$$

(c) *if $h > h^*$, then for any $x \geq h$,*

$$V(h^*, x) \geq V(h, x). \tag{4.4}$$

Then $\tau(h_*)$ is an optimal stopping time of the class $\mathcal{M}_0$.

Proof. Clearly, the rational price of the option must satisfy (4.3); hence (4.2) excludes $h < h_*$. Due to (4.3), h_* is an admissible choice, and (4.3)–(4.4) ensure that a choice $h > h_*$ is no better than h_*. □

To apply Lemma 4.1, we need an explicit formula for $V(h, x)$. We derive it by using Dynkin's formula and the solution to the Wiener–Hopf equation. Let $U^q = U^q_X$ be the potential operator (the resolvent) of the process X:

$$U^q W(x) = E^x \left[\int_0^{+\infty} e^{-qt} W(X_t) dt \right].$$

If $V \in C_0$ is sufficiently regular, for instance, $(q - L)V \in C_0$, then

$$U^q (q - L)V = V \tag{4.5}$$

(see, e.g., [33, Theorem 31.3]). We will need (4.5) for not-so-regular V.

LEMMA 4.2. *Let $W := (q - L)V := (q + \psi(D))V$ belong to L_1 and*

$$(q + \Re\psi)^{-1} \hat{W} \in L_1. \tag{4.6}$$

Then (4.5) *holds.*

Proof. Since $W \in L_1$, we have

$$\begin{aligned}(U^q W)(x) &= \int_0^{+\infty} e^{-qt} (P_t W)(x) dt \\ &= \int_0^{+\infty} e^{-qt} (2\pi)^{-n} \int_{\mathbf{R}^n} e^{-i\langle x, \xi\rangle - t\psi(\xi)} \hat{f}(\xi) d\xi dt.\end{aligned}$$

Due to (4.6), the last integral, computed in the reverse order $dtd\xi$, converges absolutely, and hence we can apply the Fubini theorem and obtain $U^q W = (q+\psi(D))^{-1} W$; (4.5) follows. □

If W is universally measurable, then for any stopping time τ, Dynkin's formula is valid (see, e.g., [33, equation (41.3)]):

$$U^q W(x) = E^x \left[\int_0^{\tau} e^{-qt} W(X_t) dt \right] + E^x \left[e^{-q\tau} U^q W(X_\tau) \right]. \tag{4.7}$$

It follows that (4.7) holds for $g \in L_1 := L_1(\mathbf{R}^n)$, which admits a representation $g = g_1 + g_2$, where $g_1 \in C_0$ and g_2 is a nonnegative (or nonpositive) function of the class L_1. Denote the class of such sums by $UL := UL(\mathbf{R}^n)$. This class is sufficiently wide for all the applications that we will need in the paper.

LEMMA 4.3. *Let $W := (q - L)V := (q + \psi(D))V \in UL$ satisfy* (4.6). *Then*

$$W(x) = E^x \left[\int_0^{\tau} e^{-qt} (q - L) W(X_t) dt \right] + E^x \left[e^{-q\tau} W(X_\tau) \right]. \tag{4.8}$$

Proof. Apply (4.7) to W. Due to (4.6), (4.5) holds, and hence, (4.7) becomes (4.8). □

Let $\sigma_+ > 0$ be from Lemma 2.4. Let $g^{(s)} = D^s g$, $s = 0, \dots, m$, be measurable, and let

$$\sum_{0 \le s \le m} |g^{(s)}(x)| \le C e^{-\omega'_+ x}, \quad x \le 0, \tag{4.9}$$

$$\sum_{0 \le s \le m} |g^{(s)}(x)| \le C e^{-\omega'_- x}, \quad x \ge 0, \tag{4.10}$$

In subsection 4.3, we will prove the following theorem.

THEOREM 4.4. *Let g satisfy (4.9)–(4.10) with $\omega'_- < \omega'_+ < \sigma_+$ and $m = 2$. Then*

(a) *for any $h \in \mathbf{R}$, a solution of the problem*

$$(q - L)V(x) = 0, \qquad x > h, \tag{4.11}$$

$$V(x) = g(x), \qquad x \le h, \tag{4.12}$$

in the class of measurable functions, bounded on $[h, +\infty)$, exists.

(b) (1) *If $\kappa_- = 1$, then a continuous bounded solution is unique. It is given by*

$$V = \phi_q^-(D)\mathbf{1}_{(-\infty,h)}\phi_q^-(D)^{-1}g. \tag{4.13}$$

(2) *If $\kappa_- \in (0,1)$, then a bounded solution is unique. It is given by (4.13), and it is continuous.*

(3) *If $\kappa_- = 0$, then a bounded solution is unique. It is given by (4.13), and it is continuous if and only if $(\phi_q^-(D)^{-1}g)(h) = 0$.*

(4) *If $\kappa_- \in (0,1]$, then $V'(h-0) = V'(h+0)$ if and only if $(\phi_q^-(D)^{-1}g)(h) = 0$.*

Remark 4.1. (a) The condition $\kappa_- = 0$ is equivalent to $\nu \in (0,1)$ and $\mu > 0$. This is the case of the process of bounded variation, with positive drift.

(b) The regularity condition on g can be relaxed: For some $s > \kappa_- + 1/2$ and $\omega'_- < \omega'_+ < \sigma_+$,

$$(e^{-\omega'_- x} + e^{-\omega'_+ x})^{-1} g \in H^s(\mathbf{R}).$$

(c) Equation (4.13) can be written as

$$V(h, \cdot) := V(\cdot) = q U_N^q \mathbf{1}_{(-\infty,h)} w(\cdot), \tag{4.14}$$

where

$$w := \phi_q^-(D)^{-1} g = U_M^q (q - L) g \tag{4.15}$$

and U_M^q and U_N^q are the resolvents of the supremum and infimum processes, respectively.

We continue to study the optimal stopping problem. If $\kappa_- > 0$ or $\kappa_- = 0$ and $w(h) = 0$, then the next lemma provides the representation of $W := (q - L)V$, which implies that $W \in UL$ and (4.6) holds. The lemma is formulated and proven under simplifying assumptions, which hold for model classes. We also require more regularity of g.

Thus, in these cases, we may use (4.8) due to Lemma 4.2. If $\kappa_- = 0$, the condition $w(h) = 0$ can be used formally to find the optimal exercise boundary as the only boundary for which the solution is continuous, and in section 6 we will show that this is really the optimal boundary (for model classes, at least). Otherwise, we cannot

justify the usage of Dynkin's formula for discontinuous V. This is the reason why we exclude the case $\nu \in (0,1)$ and $\mu > 0$ below.

LEMMA 4.5. *Assume that* $\nu_1 = \nu - 1$ *in* (2.15)*, and that* (4.9)–(4.10) *hold with* $m = 3$*. Then the following hold:*

(a) *If* $\kappa_+ < 1$*, then* W *is continuous on* $(h, +\infty)$*, exponentially decays as* $x \to +\infty$*, and admits the following representation in the right neighborhood of* h*:*

$$W(x) = a_+(0)^{-1} w(h) \Gamma(1-\kappa_+)^{-1} (x-h)^{-\kappa_+} (1 + O((x-h)^{\gamma_1}) + O((x-h)^{\gamma_2}) \tag{4.16}$$

for some $\gamma_1, \gamma_2 > 0$.

(b) *If* $\kappa_+ = 1$ *and* $w(h) = 0$*, then* W *is continuous on* $(h, +\infty)$ *and exponentially decays as* $x \to +\infty$*; in addition,* $W(h+0)$ *exists.*

(c) *If* $\kappa_+ = 1$ *and* (2.14) *holds with* $\nu = 2$ *and* $\nu_1 < 1$ *(that is, the process* X *is a mixture of a BM with independent RLPE of order less than* 1*), then the statement in* (b) *holds.*

(d) *In all cases,* W *satisfies* (4.6).

Proof of this lemma will be given in subsection 4.4.

Thus, our further considerations in this section do not apply in the case of the mixture of a BM with RLPE of order ≥ 1. Notice that we will not use the additional conditions when we verify the sufficient optimality conditions in section 6 for all mixtures of processes from model classes and put options.

Under the conditions of Lemma 4.5, we can use (4.8); from (4.11)–(4.12), we conclude that V is nothing else but $V(h,x)$ given by (4.1). Hence, (4.13) is the formula for $V(h,x)$ that we need, and we can formulate a simple sufficient optimality condition in the class $\mathcal{M}_0$.

THEOREM 4.6. *Let* p_q^-*, the (generalized) density of the distribution* μ_q^- *in* Theorem 3.1*, be continuous; let* (4.9)–(4.10) *hold with* $\omega'_- < \omega'_+ < \sigma_+$ *and* $m = 3$*; and let there exist* $\tilde{h}_1 \leq \tilde{h}_2$ *such that the following conditions are satisfied:*

$$w(x) > 0 \qquad \forall\ x < \tilde{h}_1, \tag{4.17}$$

$$w(x) = 0 \qquad \forall\ \tilde{h}_1 \leq x \leq \tilde{h}_2, \tag{4.18}$$

$$w(x) < 0 \qquad \forall\ x > \tilde{h}_2. \tag{4.19}$$

Then for any $\tilde{h} \in [\tilde{h}_1, \tilde{h}_2]$*,* $\tau(\tilde{h})$ *is an optimal stopping time in the class* $\mathcal{M}_0$.

Proof. Write (4.13), for $x > h$, as

$$V(h,x) = \int_{-\infty}^{h} p_q^-(x-y) w(y) dy \tag{4.20}$$

and as

$$V(h,x) = g(x) - \int_{h}^{+\infty} p_q^-(x-y) w(y) dy. \tag{4.21}$$

If $h < \tilde{h}_1$, we notice that $\operatorname{supp} p_q^- \subset [0, +\infty)$, and therefore from (4.17) and (4.21) we conclude that there exist x such that $V(h,x) < g(x)$, which violates the necessary optimality conditions. Now consider h on the half-axis $(\tilde{h}_2, +\infty)$, and $x > h$. By differentiating (4.20) w.r.t. h and using (4.19), we find

$$V'_h(h,x) = p_q^-(x-h) w(h) < 0;$$

hence for these h, x, we have $V(h,x) < V(\tilde{h}_2, x)$. Finally, for $\tilde{h}_1 \le h \le \tilde{h}_2$ and $x > h$, we have from (4.18)

$$V'_h(h,x) = p_q^-(x-h)w(h) = 0;$$

hence

$$V(h,x) = V(\tilde{h}_1, x) \qquad \forall\ h \in [\tilde{h}_1, \tilde{h}_2] \text{ and } x > h.$$

We conclude that any $h \in [\tilde{h}_1, \tilde{h}_2]$ is an optimal exercise boundary. □

Let us show that if (2.15) holds with any $\nu_1 \in (\nu - 1, \nu)$ (this condition is satisfied for model processes), then p_q^- is continuous on $(0, +\infty)$. For $x > 0$,

$$p_q^-(x) = (2\pi)^{-1} \int_{-\infty}^{+\infty} e^{ix\xi} \phi_q^-(\xi) d\xi. \tag{4.22}$$

By choosing ν_1 sufficiently closely to $\nu - 1$, it is possible to refine the proof of (3.25) and obtain (3.25) with any $\rho_1 \in (0,1)$ (and C depending on ρ_1). Then from (3.25) and (3.26), we deduce

$$\phi_q^-(\xi) = a_-(0)(\lambda + i\xi)^{-\kappa_-} + \hat{f}(\xi), \tag{4.23}$$

where $\hat{f}(\xi) = O((1+|\xi|)^{-s})$ as $\xi \to \infty$, with some $s > 1$. Hence, f, the inverse Fourier transform of $\hat{f}$, is a continuous function, and since for $\nu < 1$

$$\int_0^{+\infty} e^{-ix\xi} x^{-\nu-1} e^{-\lambda x} dx = \Gamma(-\nu)(\lambda + i\xi)^{\nu}, \tag{4.24}$$

we deduce from (4.22)–(4.23) that

$$p_q^-(x) = a_-(0)\Gamma(\kappa_-)^{-1} \mathbf{1}_{(0,+\infty)}(x) x^{\kappa_- - 1} e^{-\lambda x} + f(x) \tag{4.25}$$

is continuous on $(0, +\infty)$.

Example 4.1. Let g be given by (1.3). Then

$$w(x) = \sum_{j=1}^{l} c_j \phi_q^-(-i\gamma_j)^{-1} e^{\gamma_j x}, \tag{4.26}$$

and it is easy to verify the sufficient conditions of Theorem 4.6 in concrete cases. In particular, if they are satisfied, then $\tilde{h}_1 = \tilde{h}_2$; call it $\tilde{h}$.

For instance, if the option owner has the right to sell a share of the stock for $K + a\sqrt{S}$, where S is the spot price, then $g(x) = K + ae^{x/2} - e^x$, $w(x) = K + a\phi_q^-(-i/2)^{-1} e^{x/2} - \phi_q^-(-i)^{-1} e^x$, and the optimal exercise price is $\sqrt{Y}$, where Y is the only positive root of

$$K + a\phi_q^-(-i/2)^{-1} Y - \phi_q^-(-i)^{-1} Y^2 = 0.$$

When an optimal $\tilde{h}$ is found, we can calculate the rational price by using the explicit formulas for ϕ_q^-:

$$V(\tilde{h}, x) = (2\pi)^{-1} \int_{-\infty+i\sigma}^{+\infty+i\sigma} \phi_q^-(\xi) \hat{u}(\tilde{h}, \xi) d\xi, \tag{4.27}$$

where $\sigma \in (\omega'_+, \lambda_+)$ is arbitrary and $\hat{u}$ is the Fourier transform of

$$u(\tilde{h}, x) := \mathbf{1}_{(-\infty,\tilde{h})}(x)w(x)$$

w.r.t. x. If $\hat{u}(\xi)$ and ψ are holomorphic in the upper half-plane with pole(s) and/or cut(s), then we can reduce the calculation of the integral in (4.27) to the sum of terms corresponding to poles, and integrals over these cuts. This procedure allows one to derive more effective formulas. We illustrate this procedure for puts.

4.2. Perpetual American put. For puts, $g(x) = K - e^x$; (4.9) and (4.10) hold with $\omega'_+ = 0$ and $\omega'_- = -1$, respectively, and any m; and $\tilde{h}$ is defined as the solution to the equation $K - \phi_q^-(-i)^{-1}e^x = 0$, that is,

$$e^{\tilde{h}} = K\phi_q^-(-i) = KqE\left[\int_0^\infty e^{-qt+N_t}dt \mid N_0 = 0\right]. \tag{4.28}$$

Take $\sigma \in (0, \lambda_+)$ and calculate for $\Im\xi = \sigma$

$$\begin{aligned}\hat{u}(\tilde{h}, \xi) &= \int_{-\infty}^{\tilde{h}} e^{-ix\xi}(K - \phi_q^-(-i)^{-1}e^x)dx \\ &= \frac{Ke^{-i\tilde{h}\xi}}{(-i\xi)(1-i\xi)} = \frac{-Ke^{-i\tilde{h}\xi}}{\xi(\xi+i)}.\end{aligned}$$

By substituting into (4.27), we obtain the formula for the rational perpetual put price

$$V(\tilde{h}, x) = -\frac{K}{2\pi}\int_{-\infty+i\sigma}^{+\infty+i\sigma} \frac{\exp[i(x-\tilde{h})\xi]\phi_q^-(\xi)}{\xi(\xi+i)}d\xi, \tag{4.29}$$

where $\sigma \in (0, \lambda_+)$ is arbitrary.

Assume that ϕ in (2.13) admits the analytic continuation into the upper half-plane $\Im\xi > 0$ with the cut $[i\lambda_+, +i\infty)$ and satisfies (2.20) there. (If $\phi(\xi) = a\xi^2 + \phi_1(\xi)$, we assume that ϕ_1 satisfies (2.20) with $\nu_1 < 2$, in the upper half-plane with the cut.) Assume also that $q+\psi$ has the only zero $-i\beta_-$ in the upper half-plane, $0 < -\beta_- < \lambda_+$; by Lemma 2.7, these conditions are satisfied for model processes. Then ϕ_q^- admits the analytic continuation into the upper half-plane with one simple pole at $-i\beta_-$, and the cut $[i\lambda_+, +i\infty)$, by

$$\phi_q^-(\xi) = q(q + \psi(\xi))^{-1}\phi_q^+(\xi)^{-1}. \tag{4.30}$$

For $z \in (\lambda_+, +\infty)$, set

$$\Phi_q^-(z) = iq[(q + \psi(iz + 0))^{-1} - (q + \psi(iz - 0))^{-1}]\phi_q^+(iz)^{-1}. \tag{4.31}$$

By transforming the contour in (4.29) into the integral over the banks of the cut $[i\lambda_+, +i\infty)$, we meet the simple pole, which gives the first term in (4.32) below; in the integral over the banks of the contour, we make the change of variables $\xi = iz$, and, finally, obtain for $x > \tilde{h}$

$$V(\tilde{h}, x) = \frac{iqK\exp[\beta_-(x-\tilde{h})]}{\psi'(-i\beta_-)\phi_q^+(-i\beta_-)(-\beta_-)(1-\beta_-)} + (2\pi)^{-1}\int_{\lambda_+}^{+\infty} \frac{K\Phi_q^-(z)\exp[-(x-\tilde{h})z]}{z(1+z)}dz. \tag{4.32}$$

As the empirical studies of financial markets reveal, λ_+ is usually large; hence, the second term in (4.32) is small. Therefore, one may calculate it with a large relative error. This observation facilitates the numerical implementation of (4.32). The leading term is a decaying exponential function, as in the Gaussian case, when there is no cut at all, and the second term in (4.32) is zero.

In particular, in the Gaussian case,

$$(q+\psi(\xi))/q = ((-\beta_- + i\xi)/(-\beta_-))\,((\beta_+ - i\xi)/\beta_+) = \phi_q^-(\xi)^{-1}\phi_q^+(\xi)^{-1}, \tag{4.33}$$

and hence $q^{-1}\psi'(-i\beta_-) = i(-\beta_-)^{-1}\phi_q^+(-i\beta_-)^{-1}$. By substituting into (4.28) and (4.32), we obtain the optimal exercise price

$$e^{\tilde{h}} = \frac{K\beta_-}{\beta_- - 1} \tag{4.34}$$

and the rational put price, for $x > \tilde{h}$,

$$V(\tilde{h}, x) = \frac{K\exp[\beta_-(x-\tilde{h})]}{1-\beta_-} = \left(\frac{K}{1-\beta_-}\right)^{1-\beta_-}(-\beta_-)^{-\beta_-}e^{\beta_- x}. \tag{4.35}$$

This is Merton's result. One can easily calculate $\hat{u}$ for payoffs of the form (1.3), and obtain the analogues of (4.29) and (4.32) and, in the Gaussian case, of (4.34) and (4.35) as well.

4.3. Proof of Theorem 4.4. We need several basic definitions and facts of the theory of PDO. For the sake of completeness, and in order to demonstrate the role of the conditions that we impose, we give the proof of two crucial facts (for more details, see [20]).

$H^s(\mathbf{R}^n)$ is the space of generalized functions on $\mathbf{R}^n$ with the finite norm

$$||u||_s = \left(\int_{\mathbf{R}^n}(1+|\xi|^2)^s|\hat{u}(\xi)|^2 d\xi\right)^{1/2}. \tag{4.36}$$

Denote by $\overset{o}{H}{}^s(\mathbf{R}_+)$ (resp., by $\overset{o}{H}{}^s(\mathbf{R}_-)$) the subspace of $H^s(\mathbf{R})$ consisting of generalized functions supported on $[0,+\infty)$ (resp., on $(-\infty,0]$).

THEOREM 4.7. *Let $s, m \in \mathbf{R}$, and let ϕ be a measurable function which admits the following estimate, for $\xi \in \mathbf{R}$:*

$$|\phi(\xi)| \le C(1+|\xi|)^m. \tag{4.37}$$

Then

$$\phi(D) : H^s(\mathbf{R}) \to H^{s-m}(\mathbf{R}) \quad \textit{is bounded.} \tag{4.38}$$

Proof. Apply the Fourier transform and the definition of the norm (4.36) to obtain the conclusion. □

We call ϕ a symbol of order m, and $\phi(D)$ is called a PDO of order m.

THEOREM 4.8. *Let $s, m \in \mathbf{R}$. Let $\phi_\pm$ be holomorphic in the half-plane $\pm\Im\xi > 0$, continuous up to the boundary, and admitting of the estimate* (4.37) *in the closed half-plane. Then*

(a) *for any $v \in C_0^\infty((-\infty,0))$ (resp., $v \in C_0^\infty((0,+\infty))$), the function $\phi_+(D)v$ (resp., $\phi_-(D)v$) is supported on $(-\infty,0)$ (resp., on $(0,+\infty)$);*

(b) *for any* $s \in \mathbf{R}$, $\phi_{\mp}(D) : \overset{o}{H}{}^{s}(\mathbf{R}_{\pm}) \to \overset{o}{H}{}^{s-m}(\mathbf{R}_{\pm})$ *is bounded.*

We call ϕ_+ (resp., ϕ_-) a positive (resp., negative) symbol of order m.

Proof. Consider $\phi^-(D)$ and $v \in C_0^\infty((0,+\infty))$. (a) Let $x_0 := \inf \operatorname{supp} v (> 0)$. We will prove that $\phi_-(D)v(x) = 0$ for all $x \le x_0$. By changing the variable, we may assume $x_0 = 0$. Take $x \le 0$ and calculate

$$\phi_-(D)v(x) = (2\pi)^{-1} \int_{-\infty}^{+\infty} e^{ix\xi} \phi_-(\xi) \hat{u}(\xi) d\xi. \tag{4.39}$$

Change the line of integration in (4.39):

$$\phi_-(D)v(x) = (2\pi)^{-1} \int_{-\infty+i\sigma}^{+\infty+i\sigma} e^{ix\xi} \phi_-(\xi) \hat{u}(\xi) d\xi, \tag{4.40}$$

where $\sigma < 0$. Since $u \in C_0^\infty((0,+\infty))$, its Fourier transform admits the following estimate in the half-plane $\Im\xi \le 0$:

$$|\hat{u}(\xi)| \le C_N (1 + |\xi|)^{-N} \tag{4.41}$$

for any N. From (4.37) and (4.41), we conclude that the integrand admits the bound via

$$C_N e^{-\sigma x} (1 + |\xi|)^{m-N}$$

for any N. By choosing $N > m + 1$ and passing to the limit $\sigma \to -\infty$ in (4.40), we obtain 0.

(b) Since $C_0^\infty((0,+\infty))$ is dense in $\overset{o}{H}{}^{s}(\mathbf{R}_+)$, we deduce (b) from Theorem 4.7 and from (a). □

By the change of the variable $x \mapsto h + x$, we reduce the proof of Theorem 4.4 to the case $h = 0$. Next, for $\omega'_+ < \sigma_+$ in (4.9), take any $\gamma \in (\omega'_+, \sigma_+)$ and set $V_\gamma(x) = e^{\gamma x} V(x)$. Denote

$$a(D) := q + \psi(D) = q - L,$$

insert $V(x) = e^{-\gamma x} V_\gamma(x)$ into (4.11), after that multiply (4.11) and (4.12) by $e^{\gamma x}$ and use the equality

$$e^{\gamma x} a(D) e^{-\gamma x} = a(D + i\gamma). \tag{4.42}$$

We obtain

$$a(D + i\gamma) V_\gamma(x) = 0, \qquad x > 0, \tag{4.43}$$

$$V_\gamma(x) = g_\gamma(x), \qquad x \le 0. \tag{4.44}$$

Notice that g_γ decays exponentially as $x \to -\infty$: From (4.9), on $(-\infty, 0]$,

$$\sum_{0 \le s \le 2} |g_\gamma^{(s)}(x)| \le C e^{-\epsilon|x|}, \tag{4.45}$$

where $\epsilon = \gamma - \omega'_+ > 0$. Construct G_γ, which coincides with g_γ on $\mathbf{R}_-$ and admits a bound (4.45) on $\mathbf{R}$, and set $u_\gamma = V_\gamma - G_\gamma$, $F_\gamma = -a(D + i\gamma) G_\gamma$. Then u_γ solves the problem

$$a(D + i\gamma) u_\gamma(x) = F_\gamma(x), \qquad x > 0, \tag{4.46}$$

$$u_\gamma(x) = 0 \qquad x \le 0. \tag{4.47}$$

Now (4.45) implies that $G_\gamma \in H^2(\mathbf{R})$, and from (2.13)–(2.14) we conclude that $F_\gamma \in H^{2-\bar\nu}(\mathbf{R})$, where $\bar\nu = \nu$ if $\nu \geq 1$ or $\mu = 0$, and $\bar\nu = \max\{\nu, 1\}$, otherwise. Recall that we are looking for V, which is measurable and bounded on $(0, +\infty)$. Hence, u_γ is measurable and admits a bound via $Ce^{\gamma x}$. We want to reduce to the case of an unknown function of the class $L_2(\mathbf{R}_+)$. Since $\sigma_- < 0$, we can choose $\gamma' \in (\sigma_- - \gamma, -\gamma)$. Set $u_{\gamma,\gamma'}(x) = e^{\gamma' x} u_\gamma(x)$, insert $u_\gamma(x) = e^{-\gamma' x} u_{\gamma,\gamma'}(x)$ into (4.46) and (4.47), and after that multiply (4.46) by $e^{\gamma' x}$. By using (4.42), we obtain

$$a(D + i(\gamma + \gamma'))u_{\gamma,\gamma'}(x) = F_{\gamma,\gamma'}(x), \qquad x > 0, \tag{4.48}$$

$$u_{\gamma,\gamma'}(x) = 0, \qquad x \leq 0. \tag{4.49}$$

Now $u_{\gamma,\gamma'} \in L_2(\mathbf{R}_+)$, and on the strength of (2.13)–(2.14), Theorem 4.7 gives $a(D + i(\gamma + \gamma'))u_{\gamma,\gamma'} \in H^{-\bar\nu}(\mathbf{R})$. Hence, we can write the Wiener–Hopf equation (4.48) in the form

$$a(D + i(\gamma + \gamma'))u_{\gamma,\gamma'} = F_{\gamma,\gamma'} + F_-, \tag{4.50}$$

where $F_- \in \overset{o}{H}{}^{-\bar\nu}(\mathbf{R}_-)$. Multiply by q^{-1}, and then apply $\phi_q^+(D + i(\gamma + \gamma'))$. Since in the strip $\Im\xi \in (\lambda_-, \lambda_+)(\supset (\sigma_-, \sigma_+) \supset (\sigma_-, 0))$

$$q^{-1}a(\xi) = \phi_q^+(\xi)^{-1}\phi_q^-(\xi)^{-1},$$

and by our choice, $\gamma + \gamma' \in (\sigma_-, 0)$, we obtain

$$\phi_q^-(D + i(\gamma + \gamma'))^{-1}u_{\gamma,\gamma'} = K + K_-, \tag{4.51}$$

where

$$\begin{aligned} K :&= q^{-1}\phi_q^+(D + i(\gamma + \gamma'))F_{\gamma,\gamma'} \\ &= q^{-1}\phi_q^+(D + i(\gamma + \gamma'))e^{\gamma' x}(-a(D + i\gamma))G_\gamma \\ &= -\phi_q^-(D + i(\gamma + \gamma'))^{-1}e^{\gamma' x}G_\gamma \end{aligned}$$

and

$$K_- := q^{-1}\phi_q^+(D + i(\gamma + \gamma'))F_-.$$

By construction, $G_\gamma \in H^2(\mathbf{R})$, and $u_{\gamma,\gamma'} \in L_2(\mathbf{R}_+) = \overset{o}{H}{}^0(\mathbf{R}_+)$, $F_- \in \overset{o}{H}{}^{-\bar\nu}(\mathbf{R}_-)$. From Theorem 3.4, we know that for any $\sigma \in (\sigma_-, \sigma_+)$,

$$c(1 + |\xi|)^{\kappa_\pm} \leq |\phi_q^\pm(\xi)| \leq C(1 + |\xi|)^{\kappa_\pm}, \qquad \pm\Im\xi \geq \sigma;$$

therefore, by applying Theorem 4.7 and Theorem 4.8, we conclude that

$$\phi_q^-(D + i(\gamma + \gamma'))^{-1}u_{\gamma,\gamma'} \in \overset{o}{H}{}^{-\kappa_-}(\mathbf{R}_+), \qquad K_- \in \overset{o}{H}{}^{-\kappa_-}(\mathbf{R}_-), \quad K \in H^{-\kappa_-}(\mathbf{R}_+).$$

Notice that $\kappa_- \in [0, 1]$, and consider two cases: (i) $\kappa_- \in [0, 0.5)$ and (ii) $\kappa_- \in [0.5, 1]$.

In case (i), $H^{-\kappa_-}(\mathbf{R})$ is the direct sum of the subspaces $\overset{o}{H}{}^{-\kappa_-}(\mathbf{R}_\pm)$, the projections being $\theta_\pm$, the closures of the multiplication-by-$\mathbf{1}_{\mathbf{R}_\pm}$ operators defined on a dense subset $L_2(\mathbf{R}) \subset H^{-\kappa_-}(\mathbf{R})$ (see [20, Theorem 5.1 and Lemma 5.4]). Hence, from (4.51), we deduce

$$\phi_q^-(D + i(\gamma + \gamma'))^{-1}u_{\gamma,\gamma'} = -\theta_+\phi_q^-(D + i(\gamma + \gamma'))^{-1}e^{\gamma' x}G_\gamma. \tag{4.52}$$

Next, we multiply (4.52) by $\phi_q^-(D+i(\gamma+\gamma'))$, which establishes an isomorphism between $\overset{o}{H}{}^{-\kappa_-}(\mathbf{R}_+)$ and $L_2(\mathbf{R}_+)$:

$$u_{\gamma,\gamma'} = -\phi_q^-(D+i(\gamma+\gamma'))\theta_+\phi_q^-(D+i(\gamma+\gamma'))^{-1}e^{\gamma' x}G_\gamma. \tag{4.53}$$

Then we multiply (4.53) by $e^{-\gamma' x}$ and use (4.42):

$$u_\gamma = -\phi_q^-(D+i\gamma)\theta_+\phi_q^-(D+i\gamma)^{-1}G_\gamma.$$

After that, we return to

$$\begin{aligned} V_\gamma &= G_\gamma + u_\gamma \\ &= G_\gamma - \phi_q^-(D+i\gamma)\theta_+\phi_q^-(D+i\gamma)^{-1}G_\gamma \\ &= \phi_q^-(D+i\gamma)\theta_-\phi_q^-(D+i\gamma)^{-1}G_\gamma \end{aligned}$$

and notice that, since G_γ coincides with g_γ on $\mathbf{R}_-$, Theorem 4.8 ensures that $\operatorname{supp}\phi_q^-(D+i\gamma)^{-1}(G_\gamma - g_\gamma) \subset [0,+\infty)$. Thus,

$$\theta_-\phi_q^-(D+i\gamma)^{-1}G_\gamma = \theta_-\phi_q^-(D+i\gamma)^{-1}g_\gamma,$$

and in the formula for V_γ, we may replace G_γ with g_γ. By using (4.42), we finally arrive at

$$V = \phi_q^-(D)\theta_-\phi_q^-(D)^{-1}g. \tag{4.54}$$

Due to (4.9)–(4.10) and Theorem 4.7, $w := \phi_q^-(D)^{-1}g \in H^{2-\kappa_-}(\mathbf{R})$. Since $2-\kappa_- > 1/2$, we can apply Lemma 5.5 of [20] and obtain

$$\theta_-\phi_q^-(D)^{-1}g = w(0)(1-iD)^{-1}\delta + (1-iD)^{-1}\theta_-(1-iD)\phi_q^-(D)^{-1}g, \tag{4.55}$$

where δ is the Dirac delta-function. Notice that for any $\epsilon > 0$, $\delta \in H^{-1/2-\epsilon}(\mathbf{R})$ and $\theta_-(1-iD)\phi_q^-(D)^{-1}g \in H^0(\mathbf{R})$. Hence, if $\kappa_- > 0$, we obtain $V \in H^{1/2+\rho}(\mathbf{R})$ for any $\rho \in (0,\kappa_-)$. But for $s > 1/2$, $H^s(\mathbf{R}) \subset C(\mathbf{R})$, and therefore, V is continuous. By using (4.42) and (4.9), it is easy to show that the RHS in (4.54) decays exponentially.

If $\kappa_- = 0$, we have from (3.25) and (3.26)

$$\phi_q^-(D) = a_-(0) + T(D), \tag{4.56}$$

where $T(\xi)$ admits an estimate (4.37) with $m < 0$; therefore from (4.54) and (4.55) we conclude that

$$V = a_-(0)w(0)(1-iD)^{-1}\delta + V_1, \tag{4.57}$$

where $V_1 \in C(\mathbf{R})$. It is straightforward to check that the Fourier transform of $\mathbf{1}_{(-\infty,0]}(x)e^x$ is $(1-i\xi)^{-1}$; therefore $(1-iD)^{-1}\delta = \mathbf{1}_{(-\infty,0]}e^x$, and we conclude from (4.57) that V is continuous if and only if $w(0) = 0$.

In case (ii), we notice that for $s \in (-3/2,-1/2)$, the decomposition of $H^s(\mathbf{R})$ into the sum of the subspaces $\overset{o}{H}{}^s(\mathbf{R}_\pm)$ is not direct, the intersection of the latter couple being $\mathbf{C}\cdot\delta$, where δ is the Dirac delta-function. It follows that in (4.52), an additional term $C\delta$ may appear, and in (4.54), the term $C\phi_q^-(D)\delta$, where C is a constant.

If $\kappa_- = 0$, we use (3.25) and (3.26) and conclude from (4.55) that $V = C\delta + V_1$, where $V_1 \in H^s(\mathbf{R})$ for some $s > -1/2$. Since $\delta \notin H^s(\mathbf{R})$, for $s > -1/2$, and V is bounded, we conclude that C must be 0.

If $\kappa_- \in (0,1)$, we can show with the help of (3.25) and (3.26) that

$$\phi_q^-(D)\delta(x) = (2\pi)^{-1}\int_{-\infty}^{+\infty} e^{ix\xi}\phi_q^-(\xi)d\xi$$

is unbounded as $x \to +0$. Further, for $\kappa_- > 0$, V in (4.54) belongs to $H^s(\mathbf{R})$ for some $s > 1/2$ and hence is continuous. We conclude that $C\phi_q^-(D)\delta + V$ is bounded only in the case $C = 0$, and so we are left with the same (4.54). Finally, if $\kappa_- = 1$, then the argument in part (i) shows that V in (4.54) is continuous, and the same argument shows that $\phi_q^-(D)\delta(x)$ is discontinuous at 0. Hence, in order to get a continuous solution, we need $C = 0$. This finishes the proof of part (a) and part (b) (1)–(3). The last part (b) (4) can be proven by the same argument, after differentiating in (4.54).

Theorem 4.4 has thus been proven. □

4.4. Proof of Lemma 4.5. As in the proof of Theorem 4.4, we change the variable so that $h = 0$, and the usage of (4.42) establishes the exponential decay at infinity. The main difficulty is the regularity at 0. We have

$$(q - L)V = a(D)V = \phi_q^+(D)^{-1}\theta_-\phi_q^-(D)^{-1}g.$$

Under our regularity assumption on g, we can use Lemma 5.5 in [20] with one more term than in (4.55) and obtain

$$(q - L)V = W_1 + W_2 + W_3, \tag{4.58}$$

where

$$\begin{aligned} W_1 &= \phi_q^+(D)^{-1}(1 - iD)^{-1}(\phi_q^-(D)^{-1}g)(0)\delta, \\ W_2 &= \phi_q^+(D)^{-1}(1 - iD)^{-2}(\phi_q^-(D)^{-1}(1 - iD)g)(0)\delta, \\ W_3 &= \phi_q^+(D)^{-1}(1 - iD)^{-2}\theta_+(1 - iD)^2\phi_q^-(D)^{-1}g. \end{aligned}$$

Equality (4.58) holds, provided that for some $s > 5/2$, $\phi_q^-(D)^{-1}g \in H^s(\mathbf{R})$ (locally). By using (4.9)–(4.10), we conclude that, locally, $g \in H^3(\mathbf{R})$, and by Theorem 4.7, $\phi_q^-(D)^{-1}g \in H^{3-\kappa_-}(\mathbf{R})$. Since $\kappa_- \in [0,1]$, (4.58) holds.

For $|s| < 1/2$, $\theta_- : H^s(\mathbf{R}) \to \overset{o}{H}{}^s(\mathbf{R}_-)$ is bounded, and since $\kappa_+ \le 1$, we conclude that $W_3 \in \overset{o}{H}{}^s(\mathbf{R})$ for some $s > 1/2$. For such s, $H^s(\mathbf{R}) \subset C^0(\mathbf{R})$; hence W_3 is continuous and vanishes on $[0, +\infty)$. Since $\delta \in \overset{o}{H}{}^{-s}(\mathbf{R})$ for any $s > 1/2$ and since $\kappa_+ \le 1$, we obtain $W_2 \in \overset{o}{H}{}^{2-\kappa_+-s}(\mathbf{R}_-)$. If $\kappa_+ < 1$, we conclude that $W_2 \in \overset{o}{H}{}^s(\mathbf{R}_-)$. If $\kappa_+ = 1$, we use (3.23) and represent $\phi_q^+(D)^{-1}(1 - iD)^{-2}$ in the form

$$\phi_q^+(D)^{-1}(1 - iD)^{-2} = a_+(0)^{-1}(1 - iD)^{-1} + T(D), \tag{4.59}$$

where $T(\xi)$ is a positive symbol of order less than one. Hence, $f := T(D)\delta$ is continuous and vanishes on $[0, +\infty)$, and

$$(\phi_q^+(D)^{-1}(1 - iD)^{-2}\delta)(x) = a_+(0)^{-1}\mathbf{1}_{(-\infty,0]}(x)e^x + f(x). \tag{4.60}$$

It remains to consider W_1. Since in part (b), $w(h) = 0$ and hence $W_1 = 0$, the application of (4.60) finishes the proof of part (a). Let $\kappa_+ < 1$. By using (2.14) with $\nu_1 = \nu - 1$ and the same sort of argument as in the proof of (4.25), we obtain

$$\phi_q^+(D)^{-1}(1 - iD)^{-1}\delta = \Gamma(\kappa_+ - 1)^{-1}a_+(0)^{-1}\mathbf{1}_{(-\infty,0]}(x)(-x)^{-\kappa_+}e^x + f_1(x), \tag{4.61}$$

where f_1 is continuous and vanishes on $[0, +\infty)$, and (4.16) follows from (4.61).

The proof of part (c) differs from the one above. We represent $q - L$ in the form $q - L = a_2(D) + a_{\nu'}(D)$, where $a_2(D)$ is a differential operator of order 2 and $a_{\nu'}(D)$ is a PDO of order $\nu' < 1$, and consider separately $a_2(D)V$ and $a_{\nu'}(D)V$ on $(-\infty, 0)$. Since $a_2(D)$ is local, we obtain, for $x < 0$,

$$a_2(D)V(x) = \phi_q^-(D)\theta_+\phi_q^-(D)^{-1}a_2(D)g(x).$$

By Theorem 4.7, $\phi_q^-(D)^{-1}a_2(D)g \in H^{3-\kappa_- -2}(\mathbf{R}) = L_2(\mathbf{R})$, since $\kappa_- = 2 - \kappa_+ = 1$, and hence, by the same theorem, $\phi_q^-(D)\theta_+\phi_q^-(D)^{-1}a_2(D)g \in H^1(\mathbf{R})$. Hence, this is a continuous function. Since $\nu' < 1$ and the order of $\phi_q^-(D)$ is $-\kappa_- = -1$, the continuity of $a_{\nu'}(D)V = a_{\nu'}(D)\phi_q^-(D)\theta_-\phi_q^-(D)^{-1}g$ is established as in the proof of part (a) above.

It remains to verify (d). In the course of the proof of (a)–(c), we have shown that

$$|\hat{V}(\xi)| \le C(1 + |\xi|)^{-1-\kappa_-};$$

therefore, from (2.13)–(2.14),

$$(q + \Re\psi(\xi))^{-1}|\hat{W}(\xi)| = (q + \Re\psi(\xi))^{-1}|(q + \psi(\xi))\hat{V}(\xi)| \le C(1 + |\xi|)^{-1+\kappa_+-\nu}.$$

If $\kappa_- > 0$, then $\nu - \kappa_+ > 0$ and (4.6) holds.

Lemma 4.5 has thus been proven. □

5. Pricing of the perpetual American call and similar perpetual options.

5.1. Sufficient conditions for the solution for the perpetual call-like options, in the class $\mathcal{M}_0$ of hitting times $\tau(a)$ of segments $[a, +\infty)$. The substitutions $-x$ for x and the dual process $\tilde{X}$ for X transform a problem for an RLPE X on $(h, +\infty)$ into a problem for an RLPE $\tilde{X}$ on $(-\infty, -h)$. Therefore, all the statements and proofs for call-like options are obtained from the corresponding statements for put-like options by changing the direction on the real axis and the reflection of the complex plane w.r.t. the real axis. The boundedness conditions (4.9) and (4.10) allow for the growth of the payoff in the direction to $-\infty$, so the growth of the payoff for the call option is not a problem, insofar as the main considerations in the proof of the analogue of Theorem 4.4 can be restricted to a strip $\Im\xi \in (\sigma_-, \sigma_+)$, where the real part of $a(\xi) = q + \psi(\xi)$ is positive. In the case of calls, the payoff is $g(x) = e^x - K$; hence we need $a(-i) = q + \psi(-i)$ to be positive. If there is no dividend, $q = r$, and the EMM condition for the measure means that $q + \psi(-i) = 0$. This provides the formal explanation of why we need the condition $q > r$ in the case of calls; standard considerations can be used to show that in the no-dividend case, it is nonoptimal to exercise the call option ever.

The analogue of Lemma 4.1 is obtained by inverting signs in all the inequalities there, and we will not state it explicitly. The reformulation of Theorem 4.4 is also straightforward; in addition to changing the signs of inequalities and reflections, the condition $\omega'_- < \omega'_+ < \sigma_+$ must be replaced with $\sigma_- < \omega'_- < \omega'_+$.

Notice only that we consider a problem

$$(q - L)V(x) = 0, \qquad x < h, \tag{5.1}$$

$$V(x) = g(x), \qquad x \geq h, \tag{5.2}$$

in the class of measurable functions bounded on $(-\infty, h]$, and its solution is

$$V = \phi_q^+(D)\mathbf{1}_{(h,+\infty)}\phi_q^+(D)^{-1}g. \tag{5.3}$$

The solution can be written as

$$V(h,\cdot) := V(\cdot) = qU_M^q\mathbf{1}_{(h,+\infty)}w(\cdot), \tag{5.4}$$

where

$$w = U_N^q(q - L)g \tag{5.5}$$

and U_M^q and U_N^q are the resolvents of the supremum and infimum processes, respectively.

THEOREM 5.1. *Assume that $\nu_1 = \nu - 1$ in (2.15), and that (4.9)–(4.10) hold with $\sigma_- < \omega'_- < \omega'_+$ and $m = 2$, and let there exist $\tilde{h}_1 \leq \tilde{h}_2$ such that the following conditions are satisfied:*

$$w(x) < 0 \qquad \forall\ x < \tilde{h}_1, \tag{5.6}$$

$$w(x) = 0 \qquad \forall\ \tilde{h}_1 \leq x \leq \tilde{h}_2, \tag{5.7}$$

$$w(x) > 0 \qquad \forall\ x > \tilde{h}_2. \tag{5.8}$$

Then for any $\tilde{h} \in [\tilde{h}_1, \tilde{h}_2]$, $\tau(\tilde{h})$ is an optimal stopping time in the class $\mathcal{M}_0$.

5.2. Perpetual American call. For calls $g(x) = e^x - K$, (4.9) and (4.10) hold with $\omega'_+ = 0$ and $\omega'_- = -1$, respectively, and any m. So, we have to assume that $\sigma_- < -1$, which is equivalent to $q > r$.

Here $\tilde{h}$ is defined as the solution to the equation $\phi_q^+(-i)^{-1}e^x - K = 0$, that is,

$$e^{\tilde{h}} = K\phi_q^+(-i) = KqE\left[\int_0^\infty e^{-qt+M_t}dt \mid M_0 = 0\right]. \tag{5.9}$$

When an optimal $\tilde{h}$ is found, we can calculate the rational price by using the explicit formulas for ϕ_q^+:

$$V(\tilde{h}, x) = (2\pi)^{-1}\int_{-\infty+i\sigma}^{+\infty+i\sigma}\phi_q^+(\xi)\hat{u}(\tilde{h},\xi)d\xi, \tag{5.10}$$

where $\sigma \in (\lambda_-, -1)$ is arbitrary and $\hat{u}(\tilde{h},\xi)$ is the Fourier transform of $u(\tilde{h}, x) := \mathbf{1}_{(\tilde{h},+\infty)}(x)w(x)$:

$$\hat{u}(\tilde{h},\xi) = \int_{\tilde{h}}^{+\infty} e^{-ix\xi}(\phi_q^+(-i)^{-1}e^x - K)dx = \frac{Ke^{-i\tilde{h}\xi}}{(-i\xi)(1 - i\xi)} = \frac{-Ke^{-i\tilde{h}\xi}}{\xi(\xi + i)}.$$

By substituting into (5.10), we obtain the formula for the rational perpetual call price

$$V(\tilde{h}, x) = -\frac{K}{2\pi}\int_{-\infty+i\sigma}^{+\infty+i\sigma}\frac{\exp[i(x - \tilde{h})\xi]\phi_q^+(\xi)}{\xi(\xi + i)}d\xi, \tag{5.11}$$

where $\sigma \in (\lambda_-, -1)$ is arbitrary. One can easily calculate $\hat{u}$ for payoffs of the form (1.3) and obtain the analogue of (5.11).

Assume that ϕ in (2.13) admits the analytic continuation into the lower half-plane $\Im\xi < 0$ with the cut $(-i\infty, i\lambda_-]$ and satisfies (2.20) there. (If $\phi(\xi) = a\xi^2 + \phi_1(\xi)$, then we require that ϕ_1 satisfies (2.20) with $\nu_1 < 2$, in the lower half-plane with the cut.) Assume also that $q+\psi$ has the only zero $-i\beta_+$ in the lower half-plane, $\lambda_- < -\beta_+ < 0$; by Lemma 2.7, these conditions are satisfied for model processes. Then ϕ_q^+ admits the analytic continuation into the lower half-plane with one simple pole at $-i\beta_+$, and the cut $(-i\infty, i\lambda_-]$, by

$$\phi_q^+(\xi) = q(q + \psi(\xi))^{-1}\phi_q^-(\xi)^{-1}. \tag{5.12}$$

For $z \in (\lambda_+, +\infty)$, set

$$\Phi_q^+(z) = iq[(q + \psi(iz - 0))^{-1} - (q + \psi(iz + 0))^{-1}]\phi_q^-(iz)^{-1}. \tag{5.13}$$

By transforming the contour in (5.11) into the integral over the banks of the cut $(-i\infty, i\lambda_-]$, we meet the simple pole, which gives the first term in (5.14) below; in the integral over the banks of the contour, we make the change of variables $\xi = iz$, and, finally, obtain for $x < \tilde{h}$

$$V(\tilde{h}, x) = \frac{iqK\exp[\beta_+(x - \tilde{h})]}{\psi'(-i\beta_+)\phi_q^-(-i\beta_+)\beta_+(\beta_+ - 1)} + (2\pi)^{-1}\int_{-\infty}^{\lambda_-} \frac{K\Phi_q^+(z)\exp[-(x - \tilde{h})z]}{z(1 + z)}dz. \tag{5.14}$$

As the empirical studies of financial markets reveal, $-\lambda_-$ is usually large; hence, the second term in (5.14) is small. Therefore, one may calculate it with a large relative error. This observation facilitates the numerical implementation of (5.14). The leading term is a decaying exponential function, as in the Gaussian case, when there is no cut at all, and the second term in (5.14) is zero.

6. Reduction to the free boundary value problem and verification of optimality in the class $\mathcal{M}$.

6.1. Main lemma. Consider the following free boundary value problem: Given a nonnegative continuous function g, find an open set $\mathcal{C}$ and a function V such that

$$(q - L)V(x) = 0, \qquad x \in \mathcal{C}, \tag{6.1}$$

$$V(x) = g(x), \qquad x \notin \mathcal{C}, \tag{6.2}$$

$$V(x) \geq g(x), \qquad x \in \mathcal{C}, \tag{6.3}$$

$$(q - L)V(x) \geq 0, \qquad x \notin \overline{\mathcal{C}}. \tag{6.4}$$

LEMMA 6.1. *Let $(\tilde{\mathcal{C}}, \tilde{V})$ be a solution of* (6.1)–(6.4)*, let τ_* be the hitting time of $\mathcal{C}$, and let*

$$\tilde{W} := (q - L)\tilde{V} \quad \textit{be universally measurable}; \tag{6.5}$$

$$U^q\tilde{W} = \tilde{V}. \tag{6.6}$$

Then τ_* and $V_* = \tilde{V}$ solve the optimization problem (1.1).

Proof. Due to (6.1) and (6.4), $\tilde{W}$ is nonnegative, and by (6.5), it is universally measurable; therefore, for any stopping time τ, (4.7) holds, and by substituting from

(6.6), we obtain (4.8). From (4.8), (6.4), and (6.1), we conclude that for any stopping time τ,

$$\tilde{V}(x) \geq E^x \left[e^{-q\tau} \tilde{V}(X_\tau) \right], \tag{6.7}$$

and from (4.8) and (6.1), for a chosen stopping time τ_*,

$$\tilde{V}(x) = E^x \left[e^{-q\tau_*} \tilde{V}(X_{\tau_*}) \right]. \tag{6.8}$$

By using (6.3) and (6.2), we deduce from (6.7) and (6.8)

$$\tilde{V}(x) \geq E^x[e^{-q\tau} g(X_\tau)],$$

$$\tilde{V}(x) = E^x[e^{-q\tau_*} g(X_{\tau_*})].$$

But this means that a pair (τ_*, V_*), where $V_* = \tilde{V}$, is the optimal stopping time and the rational price. □

6.2. Verification of conditions of Lemma 6.1 for puts and options with payoffs (1.3). Assume that the conditions of Theorem 4.4 hold. Let $\tilde{h}$ be defined by conditions (4.17)–(4.19), and set $\mathcal{C} = (\tilde{h}, +\infty)$. Define $\tilde{V}(x) = V(h, x)$ by (4.11). Then (6.1)–(6.2) hold by Theorem 4.4, and by repeating a part of the proof of Theorem 4.6, we see that (6.3) holds. It remains to verify (6.4)–(6.6). We formulate sufficient conditions, which hold for puts and many other payoffs of the form (1.3).

Of the process, we require the following:

A. The function ϕ in (2.14) admits the analytic continuation into the lower half-plane with the cut $(-i\infty, i\lambda_-]$, and admits the bound (2.20) in this half-plane, outside a vicinity of $i\lambda_-$. If $\nu = 2$, there exist c and $\nu_1 < 2$ such that $\phi(\xi) - c\xi^2$ satisfy (2.20) with ν_1 instead of ν.

B. In a neighborhood of $i\lambda_-$, ϕ may have a weak singularity:

$$|\phi(\xi)| \leq C|\xi - i\lambda_-|^{-\alpha} \tag{6.9}$$

for some $\alpha < 1$.

C. For any $z \in (-\infty, \lambda_-)$, the limit

$$\Psi_-(z) = i[\psi(iz - 0) - \psi(iz + 0)] \tag{6.10}$$

exists and is nonpositive.

LEMMA 6.2. *Let X be a mixture of independent BM, HP, NIG, normal tempered stable (NTS) Lévy, and KoBoL. Then* A–C *hold.*

Proof. For the proof, see the appendix. □

THEOREM 6.3. *Let X be an RLPE of exponential type $[\lambda_-, \lambda_+]$, let* A–C *hold, and let*

$$g(x) = K - \sum_{j=1}^{l} c_j e^{\gamma_j x}, \tag{6.11}$$

where $K > 0$, $c_j > 0$, and $-i\gamma_j \in [i\sigma_-, 0)$, $j = 1, \ldots, l$. Then

(a) *the solution of the optimal stopping problem* (1.1) *in the class $\mathcal{M}$ exists and belongs to $\mathcal{M}_0$;*

(b) *the optimal exercise price, $\tilde{h}$, is the solution to the equation*

$$K = \sum_{j=1}^{l} c_j \phi_q^-(-i\gamma_j)^{-1} e^{\gamma_j h}; \tag{6.12}$$

(c) *the price of the option can be calculated from* (4.27) *with any $\sigma \in (0, \sigma_+)$ and*

$$\hat{u}(\xi) = K(-i\xi)^{-1} - \sum_{j=1}^{l} c_j \phi_q^-(-i\gamma_j)^{-1}(\gamma_j - i\xi)^{-1}. \tag{6.13}$$

Proof. Notice that $\hat{u}$ admits the meromorphic continuation into the complex plane with a finite number of simple poles at points $\{0, -i\gamma_1, \dots, -i\gamma_l\} \subset (i\lambda_-, i\sigma)$, and it has the following asymptotics, as $\xi \to \infty$:

$$\hat{u}(\xi) = c_1\xi^{-1} + c_2\xi^{-2} + O(|\xi|^{-3}). \tag{6.14}$$

Under condition (6.12), $c_1 = 0$, which means that the candidate for the optimal solution is more smooth than a solution for a generic h. Hence, much simpler considerations than in the proof of Lemma 4.5 for a generic h, and without additional conditions, show that $(q - L)V \in UL$ (in fact, it turns out to be even bounded), and therefore, (6.5) and (6.6) hold. It remains to verify (6.4). The conditions (6.4) and (6.12) are evidently "additive" w.r.t. g in the sense that if g_1 and g_2 satisfy (6.4) (resp., (6.12)), then $g_1 + g_2$ satisfies (6.4) (resp., (6.12)) as well. Hence, it suffices to prove that if g is a payoff of the form

$$g(x) = A - Be^{\gamma x}, \tag{6.15}$$

where $A, B > 0$, $\sigma_- \le \gamma < 0$, and

$$A - B\phi_q^-(-i\gamma)^{-1} e^{\gamma h} = 0, \tag{6.16}$$

then $V = \phi_q^-(D)\mathbf{1}_{(-\infty,h)}\phi_q^-(D)^{-1}g$ satisfies $(q - L)V \ge 0$. The payoff (6.15) being essentially the same as the one for puts, the calculations leading to (4.29) give

$$V(x) = \frac{A}{2\pi} \int_{-\infty+i\sigma}^{+\infty+i\sigma} \frac{\exp[i(x-h)\xi]\phi_q^-(\xi)}{(-i\xi)(\gamma - i\xi)} d\xi, \tag{6.17}$$

where $\sigma \in (0, \lambda_+)$ is arbitrary. By applying $(q - L) = q + \psi(D)$ to (6.17), we see that it suffices to prove that the following function is nonnegative on $(0, +\infty)$:

$$W(x) = (2\pi)^{-1} \int_{-\infty+i\sigma}^{+\infty+i\sigma} \frac{\exp[ix\xi](q + \psi(\xi))\phi_q^-(\xi)}{(-i\xi)(\gamma - i\xi)} d\xi.$$

By using A–C, we can transform the contour of integration and reduce to the integral over the banks of the cut $(-i\infty, i\lambda_-]$. In the process of the transformation, the contour crosses two simple poles at $\xi = 0$ and $\xi = -i\gamma$ (if $-i\gamma$ is a root of $q + \psi(\xi)$, as in the case of puts on a nondividend-paying stock, there is no second pole, but there is no need to consider this case separately: The corresponding term below will automatically be 0), which gives the first two terms; in the integral over the banks of the cut, we make the change of the variable $\xi = iz$. The result is

$$W(x) = q/\gamma - (1/\gamma)(q + \psi(-i\gamma))\phi^-(-i\gamma)e^{\gamma x} + (2\pi)^{-1} \int_{-\infty}^{\lambda_-} \frac{\Psi_-(z)\phi_q^-(iz)\exp[-zx]}{z(z+\gamma)} dz. \tag{6.18}$$

For $z \le 0$, $\phi_q^-(iz) > 0$, and since $\lambda_- < -\gamma$, the denominator of the integrand is positive. From (6.10), the integrand is negative; hence it is a decreasing function of x on $(-\infty, 0)$. Since $0 < \gamma \le -\sigma_-$, $(1/\gamma)(q + \psi(-i\gamma))\phi^-(-i\gamma) \ge 0$. It follows that W is decreasing on $(-\infty, 0)$, and hence it suffices to show that $W(+0) \ge 0$.

If $\kappa_+ < 1$, the integrand is absolutely integrable uniformly in $x \in (-\infty, 0]$, and therefore

$$W(+0) = W(0) = (q/\gamma) - (1/\gamma)(q + \psi(-i\gamma))\phi^-(-i\gamma) + (2\pi)^{-1} \int_{-\infty}^{\lambda_-} \frac{\Psi_-(z)\phi_q^-(iz)}{z(z+\gamma)} dz.$$

By transforming the contour of integration back, and taking into account that $(q + \psi(\xi))\phi_q^-(\xi) = q\phi_q^+(\xi)^{-1}$, we arrive at

$$W(+0) = W(0) = q(2\pi)^{-1} \int_{-\infty+i\sigma}^{+\infty+i\sigma} \frac{d\xi}{\phi_q^+(\xi)(-i\xi)(\gamma - i\xi)}.$$

The integrand is holomorphic in the upper half-plane $\Im\xi > 0$ and admits an estimate via $C(1 + |\xi|)^{-2+\kappa_+}$ for $\Im\xi \ge \sigma > 0$. Hence, we can push the line of integration up, and in the limit $\sigma \to +\infty$ obtain zero. This finishes the proof in the case $\kappa_+ < 1$.

If $\kappa_+ = 1$, we can represent $(q + \psi(\xi))\phi_q^-(\xi) = q\phi_q^+(\xi)^{-1}$ in the form

$$(q + \psi(\xi))\phi_q^-(\xi) = qa_+(0)^{-1}(-i\xi) + \chi(\xi),$$

where $a_+(0) > 0$ and χ enjoys all the properties of $(q + \psi(\xi))\phi_q^-(\xi)$ in the case $\kappa_+ < 1$ that have been used above. Hence, if we use χ instead of $(q + \psi(\xi))\phi_q^-(\xi)$ in the constructions above, we obtain a nonnegative function. To finish the proof, it remains to notice that

$$\begin{aligned} W_1(x) &:= (2\pi)^{-1} \int_{-\infty+i\sigma}^{+\infty+i\sigma} \frac{\exp[ix\xi] qa_+(0)^{-1}(-i\xi)}{(-i\xi)(\gamma - i\xi)} d\xi \\ &= qa_+(0)^{-1} \mathbf{1}_{(-\infty,0]}(x) e^{\gamma x} \ge 0. \qquad \square \end{aligned}$$

Now we consider the general case of payoffs of the form (1.3).

THEOREM 6.4. *Let the following conditions be satisfied:*

(1) *the equation*

$$\sum_{j=1}^{m} c_j \phi_q^-(-i\gamma_j)^{-1} e^{\gamma_j h} = 0 \tag{6.19}$$

has the unique solution, call it $\tilde{h}$*;*

(2) g *can be represented in the form*

$$g(x) = \sum_{k=1}^{l} c_k^+ \exp[\gamma_k^+ x] - \sum_{k=1}^{l} c_k^- \exp[\gamma_k^- x], \tag{6.20}$$

where $c_k^\pm$ *are positive,* $\gamma_k^\pm \in (-\sigma_+, -\sigma_-]$ *are not necessarily different and satisfy*

$$\gamma_k^+ < \gamma_k^- \qquad \forall\, k, \tag{6.21}$$

$$c_k^+ \phi_q^-(-i\gamma_k^+)^{-1} e^{\gamma_k^+ h} = c_k^- \phi_q^-(-i\gamma_k^-)^{-1} e^{\gamma_k^- h} \qquad \forall\, k. \tag{6.22}$$

Then

(a) *the solution of the optimal stopping problem* (1.1) *in the class* $\mathcal{M}$ *exists and belongs to* $\mathcal{M}_0$*, the* $\tilde{h}$ *being the optimal exercise price;*
(b) *the price of the option can be calculated from* (4.27) *with any* $\sigma \in (\max_j(-\gamma_j), \sigma_+)$*, and*

$$\hat{u}(\xi) = \sum_{j=1}^{l} c_j \phi_q^-(-i\gamma_j)^{-1}(\gamma_j - i\xi)^{-1}. \tag{6.23}$$

Proof. First, notice that for payoffs (6.11), (2) follows from (1).

Second, the conditions (1) and (6.4) are evidently "additive" w.r.t. g in the sense that if g_1 and g_2 satisfy (1) (resp., (6.4)), then $g_1 + g_2$ also satisfies (1) (resp., (6.4)). Condition (2) allows one to reduce to the case of the payoff of the form

$$g(x) = Ae^{\gamma^+ x} - Be^{\gamma^- x},$$

where $A, B > 0$ and $\sigma_- \le -\gamma^+ < -\gamma^- < \sigma_+$. Further, by using (4.42), we can reduce the verification to the case of the payoff (6.15), where $\gamma = \gamma^- - \gamma^+ > 0$, and $\psi(\cdot - i\gamma^+)$ instead of ψ. Since the conditions A–C are invariant under such a shift in the argument of the characteristic exponent, we can repeat the end of the proof of Theorem 6.3. □

6.3. Verification of conditions of Lemma 6.1 for calls and options with payoffs (1.3). In all formulations and proofs above, change the signs and make the reflection w.r.t. the origin.

7. The smooth fit principle. Consider the case of put options.

THEOREM 7.1. *Let* ϕ_q^- *satisfy*

$$\int_{-\infty+i\sigma}^{+\infty+i\sigma} |\phi_q^-(\xi)(1 - i\xi)^{-1}| d\xi < +\infty \tag{7.1}$$

for some $\sigma \in (0, \sigma_+)$*. Then the price of the perpetual American put satisfies the smooth fit principle.*

Proof. By differentiating under the integral sign in (4.27), we obtain $V' = v$, where

$$v(x) := -\frac{K}{2\pi} \int_{-\infty+i\sigma}^{+\infty+i\sigma} e^{i(x-\tilde{h})\xi} \phi_q^-(\xi)(1 - i\xi)^{-1} d\xi < +\infty.$$

Hence, V is smooth if and if only v is continuous. Under condition (7.1), the Fourier transform of v is of the class $L_1(\mathbf{R})$; hence v is continuous, and the smooth fit principle holds. □

THEOREM 7.2. *Let* ϕ_q^- *admit a representation* $\phi_q^-(\xi) = c + \chi(\xi)$*, where* $c \ne 0$ *and* χ *satisfies* (7.1)*. Then the smooth fit principle fails.*

Proof. This time we obtain that $v = v_1 + v_2$, where v_1 is continuous, and

$$v_2(x) = -\frac{cK}{2\pi} \int_{-\infty+i\sigma}^{+\infty+i\sigma} \frac{\exp[i(x - \tilde{h})\xi] d\xi}{1 - i\xi} = -cK\mathbf{1}_{(-\infty,\tilde{h})}(x) e^{x-\tilde{h}},$$

which is discontinuous. □

Notice that for RLPE, condition (7.1) fails if and only if $\mu > 0$ and $\nu \in (0, 1)$.

As our results in sections 4 and 6 show, the natural candidate is determined from the equation $w(x) := \phi_q^-(D)^{-1}g(x) = 0$, and it can be singled out formally in one of the following forms:

I. *If there is a unique h such that $V(h;\cdot)$ is continuous, this h is the candidate; if $V(h;\cdot)$ is continuous for all h, then the candidate is chosen by the standard smooth fit principle.* (This observation was used in [30] for a jump process with the drift.)

II. *If there is an h such that $V'_x(h; H\pm 0)$ are finite, then h is the optimal boundary.*

The second principle works for purely non-Gaussian RLPE, i.e., for RLPE of order $\nu < 2$.

In all cases, one may say that the optimal choice of h makes $V(h,\cdot)$ "more smooth" at h than generically.

8. Appendix.

8.1. Proof of Lemma 6.2. The verification of A for NIG (and more generally, normal tempered stable Lévy processes (NTSLP)) and KoBoL is trivial due to the simplicity of the analytic expressions (2.18) and (2.17). In both cases, the characteristic exponents are continuous at the ends of the cuts, and there is no singularity mentioned in B.

Verification of C for NTSLP: Here $\lambda_- = -\alpha+\beta$, and for $z < -\alpha+\beta$,

$$\begin{aligned}
\Psi_-(z) &= i\delta[(\alpha^2-(\beta+i(iz-0)^2)^{\nu/2}-(\alpha^2-(\beta+i(iz+0)^2)^{\nu/2}]\\
&= i\delta[((\alpha+\beta-z-i0)(\alpha-\beta+z+i0))^{\nu/2}-((\alpha+\beta-z+i0)(\alpha-\beta+z-i0))^{\nu/2}]\\
&= i\delta(\alpha+\beta-z)^{\nu/2}[(\alpha-\beta+z+i0)^{\nu/2}-(\alpha-\beta+z+i0)^{\nu/2}]\\
&= i\delta(\alpha+\beta-z)^{\nu/2}(-\alpha+\beta-z)^{\nu/2}[e^{i\pi\nu/2}-e^{-i\pi\nu/2}]\\
&= -\delta(\alpha+\beta-z)^{\nu/2}(-\alpha+\beta-z)^{\nu/2}2\sin[\pi\nu/2]<0.
\end{aligned}$$

Verification for KoBoL: For $z<\lambda_-$,

$$\begin{aligned}
\Psi_-(z) &= ic\Gamma(-\nu)[-(-\lambda_- - i(iz-0))^\nu+(-\lambda_- - i(iz+0))^\nu]\\
&= -ic\Gamma(-\nu)[(-\lambda_-+z+i0)^\nu-(-\lambda_-+z-i0)^\nu]\\
&= -ic\Gamma(-\nu)(-z+\lambda_-)^\nu[e^{i\pi\nu}-e^{-i\pi\nu}]\\
&= c\Gamma(-\nu)(-z+\lambda_-)^\nu 2\sin(\pi\nu)<0,
\end{aligned}$$

since $\Gamma(-\nu)\sin(\pi\nu)<0$.

Equation (2.19) being more involved, the verification of A–C for HP is rather long, and we omit it here to save space.

8.2. Proof of Lemma 2.7. Part (a) for the lower half-plane is a part of the statement A of Lemma 6.2 proven above, so it remains to prove that there are no roots of $q+\psi(\xi)$ outside the imaginary axis. Take a large R and small $\epsilon>0$ so that all roots of $q+\psi(\xi)$ on $(i\lambda_-, i\lambda_+)$ lie on $(i(\lambda_-+2\epsilon), i(\lambda_+-2\epsilon))$, and construct a contour

$$\mathcal{L}_{\epsilon,R} = \mathcal{L}^+_{\epsilon,R}\cup\mathcal{L}^l_{\epsilon,R}\cup\mathcal{L}^-_{\epsilon,R}\cup\mathcal{L}^r_{\epsilon,R},$$

where

$$\begin{aligned}
\mathcal{L}^+_{\epsilon,R} &= \{\xi \mid |\xi|\le R,\ \Im\xi\ge\lambda_+-\epsilon,\ \operatorname{dist}(\xi,[i\lambda_+,+i\infty))=\epsilon\},\\
\mathcal{L}^-_{\epsilon,R} &= \{\xi \mid |\xi|\le R,\ \Im\xi\le\lambda_-+\epsilon,\ \operatorname{dist}(\xi,(-i\infty,i\lambda_-])=\epsilon\},\\
\mathcal{L}^l_{\epsilon,R} &= \{\xi \mid |\xi|= R,\ \Re\xi<0,\ \operatorname{dist}(\xi,[i\lambda_+,+i\infty)\cup(-i\infty,i\lambda_-])\ge\epsilon\},\\
\mathcal{L}^r_{\epsilon,R} &= \{\xi \mid |\xi|= R,\ \Re\xi>0,\ \operatorname{dist}(\xi,(-i\infty,i\lambda_-]\cup[i\lambda_+,+i\infty))\ge\epsilon\}.
\end{aligned}$$

Let $U_{\epsilon,R}$ be a part of the complex plane, bounded by $\mathcal{L}_{\epsilon,R}$, and $N \in \{0,1,2\}$ (resp., $N_{\epsilon,R}$) the number of roots of $q+\psi(\xi)$ on $(i(\lambda_- + 2\epsilon), i(\lambda_+ - 2\epsilon))$ (resp., on $U_{\epsilon,R}$). Since the complex plane with cuts $(-i\infty, i\lambda_-]$ and $[i\lambda_+, +i\infty)$ is a union of all $U_{\epsilon,R}$, and since $U_{\epsilon,R} \subset U_{\epsilon',R'}$ whenever $\epsilon \geq \epsilon'$ and $R \leq R'$, it suffices to show that for sufficiently large R and small $\epsilon > 0$, $N = N_{\epsilon,R}$.

We will do this for KoBoL; the other cases can be considered similarly. First we check that

$$q + \psi(\xi) \neq 0 \tag{8.1}$$

for any $\xi \in \mathcal{L}_{\epsilon,R}$, provided that $\epsilon > 0$ is sufficiently small and R is sufficiently large:

(1) For $\xi \in \mathcal{L}_{\epsilon,R}$ such that $\Im\xi \in [\lambda_-, \lambda_+]$, (8.1) holds by continuity of ψ for all $\epsilon \in (0, \epsilon_0)$ and $R \geq R_0$, provided that $\epsilon_0 > 0$ is sufficiently small and R_0 large.

(2) As $R \to +\infty$ and $|\xi| = R$,

$$q + \psi(\xi) \sim -i\mu\xi + o(|\xi|), \quad \nu \in (0,1), \tag{8.2}$$

and if $\nu \in (1,2)$ or $\mu = 0$ and $\nu \in (0,1)$,

$$q + \psi(\xi) \sim c\Gamma(-\nu)[(-i\xi)^\nu + (i\xi)^\nu] + o(|\xi|^\nu); \tag{8.3}$$

hence (8.1) holds for these R and ξ.

(3) Fix such R; then on parts of $\mathcal{L}_{\epsilon,R}$ near the cuts, (8.1) holds since the limits of the imaginary part of $q+\psi(\xi)$ are nonzero, namely, for $z > \lambda_+$,

$$\Im(q + \psi(iz \mp 0)) = -c\Gamma(-\nu)(z - \lambda_+)^\nu \sin(\mp\pi\nu), \tag{8.4}$$

and for $z < \lambda_-$,

$$\Im(q + \psi(iz \mp 0)) = -c\Gamma(-\nu)(-z + \lambda_-)^\nu \sin(\pm\pi\nu). \tag{8.5}$$

Thus, (8.1) has been proven, and now, to show that $N = N_{\epsilon,R}$, it suffices to verify an equality

$$\frac{1}{2\pi} \int_{\partial U_{\epsilon,R}} d\arg(q + \psi(\xi)) = N. \tag{8.6}$$

One can check (8.6) by considering various N and $\nu \in (0,1)$, $\nu \in (1,2)$; if $N = 1$, one has to distinguish cases $q + \psi(i\lambda_-) > 0$, $q + \psi(i\lambda_-) < 0$, and if $\nu \in (0,1)$, cases $\mu = 0$, $\mu > 0$, and $\mu < 0$.

We write down the argument for two cases; the other cases can be considered similarly.

1. If $\nu \in (1,2)$ or $\nu \in (0,1)$ and $\mu = 0$, and $N = 2$, then at $\xi = i(\lambda_- + \epsilon)$ and $\xi = i(\lambda_+ - \epsilon)$, $q + \psi(\xi)$ is negative and (8.3) holds. When ξ moves from $i(\lambda_+ - \epsilon)$ along $\mathcal{L}_{\epsilon,R}$ counterclockwise till an intersection point with a circle $|\xi| = R$, and $\epsilon > 0$ is small enough, $q + \psi(\xi)$ moves to the right half-plane due to (8.3), passing below the origin in the complex plane due to (8.4) and an inequality $-\Gamma(-\nu)\sin(-\pi\nu) < 0$. At the intersection point, it is (approximately) equal to $2c\Gamma(-\nu)\cos(\pi\nu)R^\nu$, due to (8.3). When ξ moves along $\mathcal{L}^l_{\epsilon,R}$ till the intersection with a line $\Re\xi = -\epsilon$, $q + \psi(\xi)$ remains in an angle of less than 2π and arrives at approximately the starting point $2c\Gamma(-\nu)\cos(\pi\nu)R^\nu$, due to (8.3). After that, ξ moves to $i(\lambda_- + \epsilon)$; due to (8.5),

$q + \psi(\xi)$ passes above the origin till a point on the negative real axis. In the result, we obtain

$$\frac{1}{2\pi} \int_{\xi \in \mathcal{L}_{\epsilon,R}, \Re\xi \le 0} d \arg(q + \psi(\xi)) = 1. \tag{8.7}$$

Similarly, we obtain (8.7) with $\Re\xi \ge 0$, and by adding the two integrals, we finish the proof of (8.6).

2. Let $N = 1$, $q + \psi(i\lambda_-) < 0$, and $\nu \in (0,1)$, $\mu > 0$. Then $q + \psi(i\lambda_+) > 0$, and therefore the first part of the journey described above is in the right half-plane $\Re\xi > 0$, due to (8.2) and an assumption $\mu > 0$. During the second part of the journey, $q + \psi(\xi)$ moves above the origin and arrives at (approximately) $-\mu R$, and after that moves to $i(i\lambda_- + \epsilon)$, remaining in the left half-plane. Thus, this time we obtain (8.7) with 1/2 in the RHS, and after completing the full circle, we obtain (8.6) with $N = 1$.

Acknowledgments. The second author thanks A.N. Shiryaev, A. Melnikov, O.E. Barndorff-Nielsen, Ken-Iti Sato, G. Peskir, participants of the Europhysics conference in Dublin (July 1999), the Probability Theory seminar at MGU (December 1999), and the Workshop on Lévy processes at the University of Aarhus (January 2000), for useful discussions. The first part of the work was done at the University of Aarhus, and the second author is grateful to O.E. Barndorff-Nielsen and the University of Aarhus for the hospitality.

We are also thankful to the anonymous referee for valuable remarks on the first version of the paper.

REFERENCES

[1] O. E. Barndorff-Nielsen, *Exponentially decreasing distributions for the logarithm of particle size*, Proc. Roy. Soc. London. Ser. A, 353 (1977), pp. 401–419.

[2] O. E. Barndorff-Nielsen, *Processes of normal inverse Gaussian type*, Finance and Stochastics, 2 (1998), pp. 41–68.

[3] O. E. Barndorff-Nielsen and W. Jiang, *An Initial Analysis of Some German Stock Price Series*, Working Paper Series, 15, CAF University of Aarhus/Aarhus School of Business, Aarhus, Denmark, 1998.

[4] O. E. Barndorff-Nielsen and S. Levendorskiĭ, *Feller processes of normal inverse Gaussian type*, Quantitative Finance, 1 (2001), pp. 1–14.

[5] J. Bertoin, *Lévy Processes*, Cambridge Tracts in Math. 121, Cambridge University Press, Cambridge, UK, 1996.

[6] J-P. Bouchaud and M. Potters, *Theory of Financial Risk*, Cambridge University Press, Cambridge, UK, 2000.

[7] S. I. Boyarchenko and S. Z. Levendorskiĭ, *Models of Investment under Uncertainty When Shocks Are Non-Gaussian*, Working Paper Series EERC, 98/02, EERC/Eurasia Foundation, Moscow, 1998.

[8] S. I. Boyarchenko and S. Z. Levendorskiĭ, *On rational pricing of derivative securities for a family of non-Gaussian processes*, Preprint 98/7, Universität Potsdam, Insititut für Mathematik, Potsdam, Germany, 1998.

[9] S. I. Boyarchenko and S. Z. Levendorskiĭ, *Generalizations of the Black-Scholes Equation for Truncated Lévy Processes*, manuscript, 1999.

[10] S. I. Boyarchenko and S. Z. Levendorskiĭ, *Pricing of a Perpetual American Put for Truncated Lévy Processes*, manuscript, 1999.

[11] S. I. Boyarchenko and S. Z. Levendorskiĭ, *Option pricing for truncated Lévy processes*, in Applications of Physics in Financial Analysis (Europhysics conference abstracts, Dublin, 1999), P. Alstrom, K. B. Lauritsen, eds., European Physical Society, Mulhouse, France, 1999.

[12] S. I. Boyarchenko and S. Z. Levendorskiĭ, *Option pricing for truncated Lévy processes*, Internat. J. Theoret. Appl. Finance, 3 (2000), pp. 549–552.

[13] A. A. Borovkov, *Stochastic Processes in Queuing Theory*, Nauka, Moscow, 1972 (translation Springer, New York, 1976).
[14] R. Cont, M. Potters, and J.-P. Bouchaud, *Scaling in stock market data: Stable laws and beyond*, in Scale Invariance and Beyond (Proceedings of the CNRS Workshop on Scale Invariance, Les Houches, France, 1997), Springer, Berlin, 1997, pp. 75–85.
[15] D. A. Darling, T. Liggett, and H. M. Taylor, *Optimal stopping for partial sums*, Ann. Math. Statist., 43 (1972), pp. 1363–1368.
[16] F. Delbaen and W. Schachermayer, *A general version of the fundamental theorem of asset pricing*, Math. Ann., 300 (1994), pp. 463–520.
[17] E. Eberlein and U. Keller, *Hyperbolic distributions in finance*, Bernoulli, 1 (1995), pp. 281–299.
[18] E. Eberlein, U. Keller, and K. Prause, *New insights into smile, mispricing and value at risk: The hyperbolic model*, J. Business, 71 (1998), pp. 371–406.
[19] E. Eberlein and S. Raible, *Term structure models driven by general Lévy processes*, Math. Finance, 9 (1999), pp. 31–53.
[20] G. I. Eskin, *Boundary Problems for Elliptic Pseudo-Differential Equations*, Nauka, Moscow, 1973 (Transl. Math. Monogr. 52, AMS, Providence, RI, 1980).
[21] H. P. Mc Kean, Jr., *Appendix: A free boundary problem for the heat equation arising from a problem in mathematical economics*, Indust. Mang. Rev., 6 (1965), pp. 32–39.
[22] I. Koponen, *Analytic approach to the problem of convergence of truncated Lévy flights towards the Gaussian stochastic process*, Phys. Rev. E, 52 (1995), pp. 1197–1199.
[23] D. B. Madan, P. Carr, and E. C. Chang, *The variance Gamma process and option pricing*, European Finance Review, 2 (1998), pp. 79–105.
[24] R. N. Mantegna and H. E. Stanley, *Stochastic process with ultraslow convergence to a Gaussian: The truncated Lévy flight*, Phys. Rev. Lett., 73 (1994), pp. 2946–2949.
[25] R. N. Mantegna and H. E. Stanley, *Physics investigation of financial markets*, in Proceedings of the International School of Physics "Enrico Fermi," Course CXXXIV, F. Mallamace and H. E. Stanley, eds., IOS Press, Amsterdam, 1997, pp. 473–489.
[26] A. Matacz, *Financial modeling and option theory with the truncated Lévy process*, Internat. J. Theor. Appl. Finance, 3 (2000), pp. 143–160.
[27] R. C. Merton, *The theory of rational option pricing*, Bell J. Economics, 4 (1973), pp. 141–183.
[28] E. Mordecki, *Optimal stopping for a diffusion with jumps*, Finance Stochast., 3 (1999), pp. 227–236.
[29] E. Mordecki, *Optimal Stopping and Perpetual Options for Lévy Processes*, talk presented at the 1st World Congress of the Bachelier Finance Society, Paris, 2000.
[30] G. Peskir and A. N. Shiryaev, *Sequential testing problems for Poisson processes*, Ann. Statist., 28 (2000), pp. 837–859.
[31] N. U. Prabhu, *Stochastic Storage Processes. Queues, Insurance Risks and Dams*, Springer, New York, 1980.
[32] P. Protter, *Stochastic Integration and Differential Equations. A New Approach*, Springer-Verlag, Berlin, 1990.
[33] K. Sato, *Lévy Processes and Infinitely Divisible Distributions*, Cambridge University Press, Cambridge, UK, 1999.
[34] A. N. Shiryaev, *Essentials of Stochastic Finance. Facts, Models, Theory*, World Scientific, Singapore, River Edge, NJ, London, Hong Kong, 1999.
[35] N. Wiener and E. Hopf, *Über Eine Classe Singulärer Integralgleichungen*, Sitzber. Deutsch. Akad. Wiss. Berlin, Kl. Math. Phys. Tech., 1931, pp. 696–706.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 6, pp. 1697–1723

NORMAL FORMS AND BIFURCATIONS OF DISCRETE TIME NONLINEAR CONTROL SYSTEMS*

ARTHUR J. KRENER† AND LONG LI†

Abstract. The quadratic and cubic normal forms of discrete time nonlinear control systems are presented. These are the normal forms with respect to the group of state coordinate changes and invertible state feedbacks. We introduce the concept of a control bifurcation for such systems. A control bifurcation takes place at an equilibrium where there is a loss of linear stabilizability in contrast to a classical bifurcation, which typically takes place at an equilibrium where there is a loss of linear stability. We present the analogous control bifurcations to the well-known classical bifurcations; the fold, the transcritical, the flip, and the Neimark–Sacker bifurcations. When the loop is closed, a control bifurcation can lead to a classical bifurcation.

Key words. discrete time nonlinear control system, normal forms, bifurcations, control bifurcations

AMS subject classifications. 93C10, 93C55, 37G05, 37G10, 37G15

PII. S036301290037898X

1. Introduction.

The theory of normal forms and bifurcations of nonlinear difference equations is well known [1], [5], [9], [13]. Briefly, it is as follows. Consider two smooth (C^4) n dimensional difference equations with equilibrium points

$$\begin{aligned} x^+ &= f(x), \\ 0 &= f(0) \end{aligned} \tag{1.1}$$

and

$$\begin{aligned} z^+ &= g(z), \\ 0 &= g(0), \end{aligned} \tag{1.2}$$

where $x^+(t) = x(t+1)$. These are *locally diffeomorphic* if there exists a local diffeomorphism

$$\begin{aligned} z &= \phi(x), \\ 0 &= \phi(0) \end{aligned} \tag{1.3}$$

which carries (1.1) to (1.2),

$$g(\phi(x)) = \phi(f(x)).$$

Such a local diffeomorphism carries trajectories $x(t)$ in its domain onto trajectories $z(t)$ in its range,

$$z(t) = \phi(x(t));$$

*Received by the editors October 3, 2000; accepted for publication (in revised form) October 30, 2001; published electronically February 14, 2002.
http://www.siam.org/journals/sicon/40-6/37898.html
†Department of Mathematics, University of California, Davis, CA 95616-8633 (ajkrener@ucdavis.edu, long21ST@yahoo.com). The first author's research was supported in part by NSF 9970998.

hence the two dynamics are locally smoothly equivalent.

There is a weaker notion of equivalence; (1.1) is *locally topologically conjugate* to (1.2) if there is a local homeomorphism (1.3) which carries trajectories $x(s)$ in its domain onto trajectories $z(t)$ in its range while preserving the orientation of time, but not the exact time.

The *linear approximation* of (1.1) around the fixed point $x = 0$ is

$$\delta x^+ = \frac{\partial f}{\partial x}(0)\ \delta x, \tag{1.4}$$

and this is a *hyperbolic fixed point* if $\frac{\partial f}{\partial x}(0)$ has no eigenvalues on the unit circle. The discrete time *Grobman–Hartman* theorem states that if the equilibrium $x = 0$ of (1.1) is hyperbolic, then it is locally topologically conjugate to its linear approximation (1.4). A related theorem is that two hyperbolic equilibria are locally topologically conjugate if their linear approximations have the same number of eigenvalues strictly inside the unit circle, the signs of their products are the same, and the same number of eigenavalues strictly outside and the signs of their products are the same [9].

A parametrized system

$$x^+ = f(x, \mu) \tag{1.5}$$

can have a locus of equilibria

$$x_e = f(x_e, \mu_e).$$

It undergoes a *local bifurcation* at an equilibrium $x_e,\ \mu_e$ that is not locally topologically conjugate to every nearby equilibrium. In light of the above, such a bifurcation can happen only if one or more eigenvalues of the linearized system cross the unit circle, or the sign of the product of the strictly stable eigenvalues changes, or the sign of the product of the strictly unstable eigenvalues changes.

A standard approach to analyzing the behavior of the parametrized system (1.5) around a bifurcation point is to add the parameter as an additional state with trivial dynamics

$$\mu^+ = \mu \tag{1.6}$$

and then compute the center manifold through the bifurcation point and the dynamics restricted to this manifold [3], [9]. The center manifold is an invariant manifold of the extended difference equation (1.5)–(1.6), which is tangent at the bifurcation point to the eigenspace of the eigenvalues on the unit circle. In practice, one does not compute the center manifold and its dynamics exactly; in most cases of interest, an approximation of degree two or three suffices. If the other eigenvalues are off the unit circle, then this part of the dynamics cannot affect the local topological conjugacy around the bifurcation point. If at the bifurcation point all of the eigenvalues of the linear approximation are inside or on the unit circle, then the bifurcation point will be locally asymptotically stable for the complete dynamics iff the dynamics on the center manifold is locally asymptotically stable. Of course, at some nearby equilibria the dynamics may be unstable.

The next step is to compute the Poincaré normal form of the center manifold dynamics. This is a normal form under smooth changes of coordinates

$$z = \phi(x) = Tx - \phi^{[2]}(x) - \phi^{[3]}(x) - \cdots, \tag{1.7}$$

where $\phi^{[d]}(x)$ denotes a vector field that is a homogeneous polynomial of degree d in x. The linear part of the change of coordinates T puts the linear part of the center manifold dynamics in Jordan form. The quadratic, cubic, and higher parts of the change of coordinates $\phi^{[2]}$ and $\phi^{[3]}$ simplify the quadratic, cubic, and higher parts of the center manifold dynamics by putting them in Poincaré normal form. From its normal form the bifurcation is recognized and understood. Examples are the fold (or saddle-node), the flip, and the Neimark–Sacker bifurcations. The first depends on the normal form of degree two, and the last two depend on the normal form of degree three. These are the only ones that are generic and of codimension 1, i.e., depend on a single parameter, so these are the most important.

Kang and Krener [6] developed a quadratic normal form for continuous time nonlinear systems whose linear part is controllable. This was extended to discrete time systems by Barbot, Monaco, and Normand-Cyrot [2]. These authors considered a larger group of transformations to bring the system to normal form, including invertible state feedback as well as change of state coordinates. Kang [7], [8] also developed a quadratic normal form for continuous time nonlinear systems whose linear part may have uncontrollable modes. Krener, Kang, and Chang [10], [4] described the quadratic and cubic normal forms of continuous time nonlinear control systems and also their bifurcations.

In this paper, we will develop quadratic and cubic normal forms for discrete time nonlinear control systems of the form

$$\begin{aligned} x^+ = f(x,u) = Ax + Bu + f^{[2]}(x,u) \\ +f^{[3]}(x,u) + O(x,u)^4, \end{aligned} \tag{1.8}$$

where x, u are of dimensions $n, 1$ and $f^{[d]}(x,u)$ denotes a vector field that is a homogeneous polynomial of degree d in x, u. We do not assume that the linear part of the system is controllable. Moreover, our linear and quadratic normal forms differ from that of [2] for linearly controllable systems.

We also describe some of the simplest bifurcations of discrete time nonlinear control systems. A control system does not need a parameter to bifurcate; the control can play the same role. The equilibria of a controlled difference equation,

$$x^+ = f(x,u), \tag{1.9}$$

are those values of x_e, u_e such that $f(x_e, u_e) = x_e$. The equilibria are conveniently parametrized by u or one of the state variables. Two key facts differentiate bifurcations of a control system (1.8) from that of a parametrized system (1.5). The first is that for the latter the structural stability of the equilibria is the crucial issue, but for the former the stabilizability by state feedback is the crucial issue. A control system (1.8) is linearly controllable (linearly stabilizable) at x_e, u_e if the local linear approximation

$$\delta x^+ = \frac{\partial f}{\partial x}(x_e, u_e)\ \delta x + \frac{\partial f}{\partial u}(x_e, u_e)\ \delta u$$

is controllable (stabilizable). If the linear approximation is stabilizable, then the nonlinear system is locally stabilizable. If the linear approximation is not stabilizable, then the nonlinear system may or may not be locally stabilizable, depending on higher degree terms. A *control bifurcation* of (1.8) takes place at an equilibrium where the linear approximation loses stabilizability. Notice that this is different from the bifurcation of a parametrized system (1.5), which takes place at an equilibrium where there

is a loss of structural stability with respect to parameter variations. To emphasize this distinction, we shall refer to the latter as a *classical bifurcation*.

The other difference between control and classical bifurcations is that when bringing the control system into normal form, a different group of transformations is used. For classical bifurcations, we use parameter dependent change of state coordinates and change of parameter coordinates, but for control bifurcations we use change of state coordinates and state dependent change of control coordinates (invertible state feedback) to simplify the dynamics.

2. Quadratic normal form. Consider a smooth (C^3) system of the form (1.8) under the action of linear and quadratic change of state coordinates and state feedback

$$z = \phi(x) = Tx - \phi^{[2]}(x), \tag{2.1}$$

$$v = \alpha(x,u) = Kx + Lu - \alpha^{[2]}(x,u), \tag{2.2}$$

where $T,\ L$ are invertible.

It is well known that there exist a linear change of coordinates T and a linear feedback $K,\ L$ that transform the system into the linear normal form

$$\begin{aligned}\begin{bmatrix} x_1^+ \\ x_2^+ \end{bmatrix} &= \begin{bmatrix} f_1(x_1,x_2,u) \\ f_2(x_1,x_2,u) \end{bmatrix} \\ &= \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ B_2 \end{bmatrix} u \\ &\quad + \begin{bmatrix} f_1^{[2]}(x_1,x_2,u) \\ f_2^{[2]}(x_1,x_2,u) \end{bmatrix} + O(x_1,x_2,u)^3,\end{aligned} \tag{2.3}$$

where $x_1,\ x_2$ are $n_1,\ n_2$ dimensional, $n_1 + n_2 = n$, A_1 is in Jordan form, and A_2, B_2 are in controller (Brunovsky) form:

$$A_2 = \begin{bmatrix} 0 & 1 & \dots & 0 \\ & \ddots & \ddots & \\ 0 & 0 & \dots & 1 \\ 0 & 0 & \dots & 0 \end{bmatrix}, \qquad B_2 = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

The following result generalizes [11].

THEOREM 2.1. *Consider the system* (2.3)*, where A_1 is diagonal and A_2, B_2 are in Brunovsky form. There exist a quadratic change of coordinates and a quadratic feedback*

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} \phi_1^{[2]}(x_1,x_2) \\ \phi_2^{[2]}(x_1,x_2) \end{bmatrix},$$

$$v = u - \alpha^{[2]}(x_1,x_2,u)$$

which transform the system (2.3) *into the quadratic normal form*

$$\begin{aligned}\begin{bmatrix} z_1^+ \\ z_2^+ \end{bmatrix} &= \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \begin{bmatrix} 0 \\ B_2 \end{bmatrix} v \\ &\quad + \begin{bmatrix} \tilde f_1^{[2;0]}(z_1;z_2,v) & + & \tilde f_1^{[1;1]}(z_1;z_2,v) & + & \tilde f_1^{[0;2]}(z_1;z_2,v) \\ 0 & + & 0 & + & \tilde f_2^{[0;2]}(z_1;z_2,v) \end{bmatrix} \\ &\quad + O(z_1,z_2,v)^3,\end{aligned} \tag{2.4}$$

where $\tilde{f}_i^{[d_1;d_2]}(z_1; z_2, v)$ *is a polynomial vector field homogeneous of degree* d_1 *in* z_1 *and homogeneous of degree* d_2 *in* z_2, v. *For notational convenience, we define* $z_{2,n_2+1} = v$.

The vector field $\tilde{f}_1^{[2;0]}$ *is in the quadratic normal form of Poincaré,*

$$\tilde{f}_1^{[2;0]} = \sum_{\lambda_i=\lambda_j\lambda_k} \beta_i^{jk}\ \mathbf{e}_1^i\ z_{1,j}\ z_{1,k}, \tag{2.5}$$

where $\lambda_1, \ldots, \lambda_{n_1}$ *are the eigenvalues of* A_1, $\mathbf{e}_r^i$ *is the* ith *unit vector in* z_r *space, and* $z_{r,i}$ *is the* ith *component of* z_r. *The other vector fields are as follows:*

$$\begin{aligned}\tilde{f}_1^{[1;1]} = &\sum_{\lambda_i=0}\sum_{\lambda_j=0}\sum_{k=1}^{n_2+1} \gamma_i^{jk}\ \mathbf{e}_1^i\ z_{1,j}\ z_{2,k} \\ &+ \sum_{\lambda_i\neq 0}\sum_{\lambda_j\neq 0} \gamma_i^{j1}\ \mathbf{e}_1^i\ z_{1,j}\ z_{2,1},\end{aligned} \tag{2.6}$$

$$\tilde{f}_1^{[0;2]} = \sum_{\lambda_i\neq 0}\sum_{k=1}^{n_2+1} \delta_i^{1k}\ \mathbf{e}_1^i\ z_{2,1}\ z_{2,k}, \tag{2.7}$$

$$\tilde{f}_2^{[0;2]} = \sum_{i=1}^{n_2-1}\sum_{k=i+2}^{n_2+1} \epsilon_i^{1k}\ \mathbf{e}_2^i\ z_{2,1}\ z_{2,k}. \tag{2.8}$$

The normal form is unique; that is, each system (2.3) *can be transformed into only one such normal form* (2.4)–(2.8) *by a quadratic change of coordinates* (2.1) *and quadratic feedback* (2.2). *This follows from the fact that the numbers in the above,* $\beta_i^{jk}, \gamma_i^{jk}, \delta_i^{1k}, \epsilon_i^{1k}$ *for the indicated indices, are moduli, i.e., continuous invariants of the system* (2.3) *under a quadratic change of coordinates and quadratic feedback. Let* $\sigma_{jk} = 2$ *if* $j = k$ *and* $\sigma_{jk} = 1$ *otherwise. The moduli are defined as follows:*

$$\begin{aligned}\beta_i^{jk} = &\frac{1}{\sigma_{jk}}\ \frac{\partial^2 f_{1,i}}{\partial x_{1,j}\partial x_{1,k}}(0,0,0) \\ &\text{for } 1 \le i,\ j,k \le n_1,\ \text{and } \lambda_i = \lambda_j\lambda_k,\end{aligned} \tag{2.9}$$

$$\begin{aligned}\gamma_i^{jk} = &\frac{\partial^2 f_{1,i}}{\partial x_{1,j}\partial x_{2,k}}(0,0,0) \\ &\text{for } 1 \le i,\ j \le n_1,\ 1 \le k \le n_2+1,\ \text{and } \lambda_i = \lambda_j = 0,\end{aligned} \tag{2.10}$$

$$\begin{aligned}\gamma_i^{j1} = &\sum_{l=0}^{n_2} \left(\frac{\lambda_i}{\lambda_j}\right)^l \frac{\partial^2 f_{1,i}}{\partial x_{1,j}\partial x_{2,l+1}}(0,0,0) \\ &\text{for } 1 \le i,\ j \le n_1,\ \text{and } \lambda_i\lambda_j \neq 0,\end{aligned} \tag{2.11}$$

$$\begin{aligned}\delta_i^{1k} = &\frac{1}{\sigma_{1k}} \sum_{l=0}^{n_2-k+1} \lambda_i^l \frac{\partial^2 f_{1,i}}{\partial x_{2,1+l}\partial x_{2,k+l}}(0,0,0) \\ &\text{for } 1 \le i \le n_1,\ 1 \le k \le n_2+1,\ \text{and } \lambda_i \neq 0,\end{aligned} \tag{2.12}$$

$$\begin{aligned}\epsilon_i^{1k} = &\sum_{l=0}^{n_2-k+1} \frac{\partial^2 f_{2,i+l}}{\partial x_{2,1+l}\partial x_{2,k+l}}(0,0,0) \\ &\text{for } 1 \le i \le n_2-1 \text{ and } i+2 \le k \le n_2+1.\end{aligned} \tag{2.13}$$

Remarks. If some of the eigenvalues of A_1 are complex, then a linear complex change of coordinates is required to bring it to Jordan form. In this case, some of

the coordinates of z_1 are complex conjugate pairs, and some of the coefficients in the normal form are complex. These complex coefficients occur in conjugate pairs so that the real dimension of the coefficient space of the normal form is unchanged.

In the normal form of Poincaré (2.5), the eigenvalues satisfying $\lambda_i = \lambda_j \lambda_k$ are said to be in quadratic resonance.

We defer the proof to a later section as it is quite lengthy.

3. Cubic normal form. We present the cubic normal form of a system that is already in linear and quadratic normal form.

THEOREM 3.1. *Consider a smooth* (C^4) *system*

$$\begin{aligned}\begin{bmatrix} x_1^+ \\ x_2^+ \end{bmatrix} &= \begin{bmatrix} f_1(x_1, x_2, u) \\ f_2(x_1, x_2, u) \end{bmatrix} \\ &= \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ B_2 \end{bmatrix} u \\ &\quad + \begin{bmatrix} f_1^{[2;0]}(x_1; x_2, u) \\ 0 \end{bmatrix} + \begin{bmatrix} f_1^{[1;1]}(x_1; x_2, u) \\ 0 \end{bmatrix} + \begin{bmatrix} f_1^{[0;2]}(x_1; x_2, u) \\ f_2^{[0;2]}(x_1; x_2, u) \end{bmatrix} \\ &\quad + \begin{bmatrix} f_1^{[3]}(x_1; x_2, u) \\ f_2^{[3]}(x_1; x_2, u) \end{bmatrix} + O(x_1, x_2, u)^4, \end{aligned} \tag{3.1}$$

where A_1 *is diagonal,* A_2, B_2 *are in Brunovsky form, and the quadratic terms are in the normal form of Theorem* 2.1. *There exist a cubic change of coordinates and a cubic feedback*

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} \phi_1^{[3]}(x_1, x_2) \\ \phi_2^{[3]}(x_1, x_2) \end{bmatrix},$$

$$v = u - \alpha^{[3]}(x_1, x_2, u)$$

which transform the system (3.1) *into the cubic normal form*

$$\begin{aligned}\begin{bmatrix} z_1^+ \\ z_2^+ \end{bmatrix} &= \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \begin{bmatrix} 0 \\ B_2 \end{bmatrix} v \\ &\quad + \begin{bmatrix} f_1^{[2;0]}(z_1; z_2, v) \\ 0 \end{bmatrix} + \begin{bmatrix} f_1^{[1;1]}(z_1; z_2, v) \\ 0 \end{bmatrix} + \begin{bmatrix} f_1^{[0;2]}(z_1; z_2, v) \\ f_2^{[0;2]}(z_1; z_2, v) \end{bmatrix} \\ &\quad + \begin{bmatrix} \tilde{f}_1^{[3;0]}(z_1; z_2, v) \\ 0 \end{bmatrix} + \begin{bmatrix} \tilde{f}_1^{[2;1]}(z_1; z_2, v) \\ 0 \end{bmatrix} \\ &\quad + \begin{bmatrix} \tilde{f}_1^{[1;2]}(z_1; z_2, v) \\ \tilde{f}_2^{[1;2]}(z_1; z_2, v) \end{bmatrix} + \begin{bmatrix} \tilde{f}_1^{[0;3]}(z_1; z_2, v) \\ \tilde{f}_2^{[0;3]}(z_1; z_2, v) \end{bmatrix} + O(z_1, z_2, v)^4. \end{aligned} \tag{3.2}$$

The vector field $\tilde{f}_1^{[3;0]}$ *is in the cubic normal form of Poincaré,*

$$\tilde{f}_1^{[3;0]} = \sum_{\lambda_i = \lambda_j \lambda_k \lambda_l} \beta_i^{jkl} \, \mathbf{e}_1^i \; z_{1,j} \; z_{1,k} \; z_{1,l}, \tag{3.3}$$

and the other vector fields are as follows:

$$\tilde{f}_1^{[2;1]} = \sum_{\lambda_i=0} \sum_{\lambda_j\lambda_k=0} \sum_{l=1}^{n_2+1} \gamma_i^{jkl}\ \mathbf{e}_1^i\ z_{1,j}\ z_{1,k}\ z_{2,l}$$

$$+ \sum_{\lambda_i\neq 0} \sum_{\lambda_j\lambda_k\neq 0} \gamma_i^{jk1}\ \mathbf{e}_1^i\ z_{1,j}\ z_{1,k}\ z_{2,1}, \tag{3.4}$$

$$\tilde{f}_1^{[1;2]} = \sum_{\lambda_i=0} \sum_{\lambda_j=0} \sum_{k=1}^{n_2+1} \sum_{l=k}^{n_2+1} \delta_i^{jkl}\ \mathbf{e}_1^i\ z_{1,j}\ z_{2,k}\ z_{2,l}$$

$$+ \sum_{\lambda_i\neq 0} \sum_{\lambda_j\neq 0} \sum_{l=1}^{n_2+1} \delta_i^{j1l}\ \mathbf{e}_1^i\ z_{1,j}\ z_{2,1}\ z_{2,l}, \tag{3.5}$$

$$\tilde{f}_1^{[0;3]} = \sum_{\lambda_i\neq 0} \sum_{k=1}^{n_2+1} \sum_{l=k}^{n_2+1} \epsilon_i^{1kl}\ \mathbf{e}_1^i\ z_{2,1}\ z_{2,k}\ z_{2,l}, \tag{3.6}$$

$$\tilde{f}_2^{[1;2]} = \sum_{i=1}^{n_2-1} \sum_{\lambda_j\neq 0} \sum_{l=i+2}^{n_2+1} \zeta_i^{j1l}\ \mathbf{e}_2^i\ z_{1,j}\ z_{2,1}\ z_{2,l}, \tag{3.7}$$

$$\tilde{f}_2^{[0;3]} = \sum_{i=1}^{n_2-1} \sum_{l=i+2}^{n_2+1} \sum_{k=1}^{l} \eta_i^{1kl}\ \mathbf{e}_2^i\ z_{2,1}\ z_{2,k}\ z_{2,l}. \tag{3.8}$$

The normal form is unique; that is, each system (3.1) *can be transformed into only one such normal form* (3.2)–(3.8). *This follows from the fact that the numbers in the above,* $\beta_i^{jkl}, \gamma_i^{jkl}, \delta_i^{1kl}, \epsilon_i^{1kl}, \zeta_i^{j1l}, \eta_i^{1kl}$ *for the indicated indices, are moduli of the system* (2.3) *under a cubic change of coordinates and cubic feedback. Let* $\sigma_{jkl} = 6$ *if* $j = k = l$ *and* $\sigma_{jkl} = \sigma_{jk}\sigma_{kl}\sigma_{jl}$ *otherwise. These moduli are defined as follows:*

$$\beta_i^{jkl} = \frac{1}{\sigma_{jkl}}\ \frac{\partial^3 f_{1,i}}{\partial x_{1,j}\partial x_{1,k}\partial x_{1,l}}(0,0,0) \tag{3.9}$$

for $1 \le i,\ j,k,l \le n_1$, *and* $\lambda_i = \lambda_j\lambda_k\lambda_l$,

$$\gamma_i^{jkl} = \frac{1}{\sigma_{jk}}\ \frac{\partial^3 f_{1,i}}{\partial x_{1,j}\partial x_{1,k}\partial x_{2,l}}(0,0,0) \tag{3.10}$$

for $1 \le i \le n_1,\ 1 \le j \le k \le n_1,\ 1 \le l \le n_2+1$,
and $\lambda_i = \lambda_j\lambda_k = 0$,

$$\gamma_i^{jk1} = \frac{1}{\sigma_{jk}} \sum_{r=0}^{n_2-k+1} \left(\frac{\lambda_i}{\lambda_j\lambda_k}\right)^r \frac{\partial^3 f_{1,i}}{\partial x_{1,j}\partial x_{1,k}\partial x_{2,r+1}}(0,0,0) \tag{3.11}$$

for $1 \le i \le n_1,\ 1 \le j \le k \le n_1$, *and* $\lambda_i\lambda_j\lambda_k \neq 0$,

$$\delta_i^{jkl} = \frac{1}{\sigma_{kl}}\ \frac{\partial^3 f_{1,i}}{\partial x_{1,j}\partial x_{2,k}\partial x_{2,l}}(0,0,0) \tag{3.12}$$

for $1 \le i,\ j \le n_1,\ 1 \le k \le l \le n_2+1$, *and* $\lambda_i = \lambda_j = 0$,

$$\delta_i^{j1l} = \frac{1}{\sigma_{1l}} \sum_{r=0}^{n_2-l+1} \left(\frac{\lambda_i}{\lambda_j}\right)^l \frac{\partial^3 f_{1,i}}{\partial x_{1,j}\partial x_{2,1+r}\partial x_{2,l+r}}(0,0,0) \tag{3.13}$$

for $1 \le i,\ j \le n_1,\ 1 \le k \le n_2+1$, *and* $\lambda_i\lambda_j\lambda_k \neq 0$,

$$\epsilon_i^{1kl} = \frac{1}{\sigma_{1kl}} \sum_{r=0}^{n_2-l+1} \lambda_i^r \frac{\partial^3 f_{1,i}}{\partial x_{2,1+r}\partial x_{2,k+r}\partial x_{2,l+r}}(0,0,0) \tag{3.14}$$

$$\text{for } 1 \le i \le n_1,\ 1 \le k \le l,\ i+2 \le l \le n_2+1,\ \text{and } \lambda_i \neq 0,$$

$$\zeta_i^{j1l} = \frac{1}{\sigma_{1l}} \sum_{r=0}^{n_2-l+1} \lambda_j^{-r} \frac{\partial^3 f_{2,i+r}}{\partial x_{1,j} \partial x_{2,1+r} \partial x_{2,l+r}}(0,0,0) \tag{3.15}$$

$$\text{for } 1 \le i \le n_2-1,\ i+2 \le l \le n_2+1,\ \text{and } \lambda_j \neq 0,$$

$$\eta_i^{1kl} = \frac{1}{\sigma_{1kl}} \sum_{r=0}^{n_2-l+1} \frac{\partial^3 f_{2,i+r}}{\partial x_{2,1+r} \partial x_{2,k+r} \partial x_{2,l+r}}(0,0,0) \tag{3.16}$$

$$\text{for } 1 \le i \le n_2-1,\ 1 \le k \le l,\ \text{and } i+2 \le l \le n_2+1.$$

Remarks. Once again, if some of the eigenvalues of A_1 are complex, then a linear complex change of coordinates is required to bring it to Jordan form. In this case, some of the coordinates of z_1 are complex conjugate pairs, and some of the coefficients in the normal form are complex. These complex coefficients occur in conjugate pairs so that the real dimension of the coefficient space of the normal form is unchanged.

In the normal form of Poincaré (3.3), the eigenvalues satisfying $\lambda_i = \lambda_j \lambda_k \lambda_l$ are said to be in cubic resonance.

We defer the proof to a later section as it is quite lengthy.

4. Control bifurcations. In the above theorems, there are many more details than are necessary to understand the types of bifurcations that are possible. Recall that, in the bifurcation theory of a parametrized system of difference equations, the interesting part of the dynamics is that restricted to the center manifold. This leads to a great reduction in the dimension of space that must be explored. A similar fact holds true when studying control bifurcations. In most applications, one will ultimately use state feedback in an attempt to stabilize the system so the coordinates that are linearly stabilizable can be ignored to a large extent. If there are modes which are neutrally stable and are not linearly stabilizable, then the particular choice of feedback will influence the shape of the center manifold of the closed loop system and the dynamics thereon. It might be possible to achieve asymptotically stable center manifold dynamics by the proper choice of feedback although it will not be exponentially stable. We now discuss some important bifurcations of control systems.

4.1. Fold control bifurcation. Just as with classical bifurcations of discrete time dynamical systems, the simplest control bifurcation is the fold. The uncontrollable part is one dimensional and unstable with $A_1 > 1$. Because the linearly controllable part of the quadratic normal form (2.4) is in Brunovsky form, the equilibria z_e, v_e are conveniently parametrized by $\mu = v_e$. The equilibria $z_e(\mu), v_e(\mu)$ are given by

$$z_{e1} = \mu^2 (1 - A_1)^{-1} \tilde{\delta} + O(\mu)^3, \tag{4.1}$$

$$z_{e2,i} = \mu + O(\mu)^2, \qquad i = 1, \ldots, n_2, \tag{4.2}$$

$$v_e = \mu, \tag{4.3}$$

where

$$\tilde{\delta} = \sum_{k=1}^{n_2+1} \delta_1^{1k}.$$

The local linearization around z_e, v_e is

$$\begin{aligned}\begin{bmatrix} \tilde{z}_1^+ \\ \tilde{z}_2^+ \end{bmatrix} &= \left(\begin{bmatrix} A_1 + \mu\gamma_1^{11} & \mu\Delta \\ 0 & A_2 \end{bmatrix} + O(\mu)^2 \right) \begin{bmatrix} \tilde{z}_1 \\ \tilde{z}_2 \end{bmatrix} \\ &\quad + \left(\begin{bmatrix} \mu B_1 \\ B_2 \end{bmatrix} + O(\mu)^2 \right) \tilde{v}, \end{aligned} \tag{4.4}$$

where $\tilde{z} = z - z_e(\mu)$, $\tilde{v} = v - v_e(\mu)$, and

$$\begin{aligned}\Delta &= \begin{bmatrix} \tilde{\delta} + \delta_1^{11} & \delta_1^{12} & \dots & \delta_1^{1n_2} \end{bmatrix}, \\ B_1 &= \delta_1^{1,n_2+1}. \end{aligned}$$

If the transversality condition

$$\tilde{\delta} + \delta_1^{11} + A_1\delta_1^{12} + \cdots + A_1^{n_2}\delta_1^{1,n_2+1} \neq 0 \tag{4.5}$$

is satisfied, then the system is linearly controllable and hence stabilizable about any equilibrium except $\mu = 0$. Consider a parametrized family of feedbacks

$$\begin{aligned} v &= \kappa(z, \mu), \\ \tilde{v} &= K_1(\mu)\tilde{z_1} + K_2(\mu)\tilde{z_2}. \end{aligned} \tag{4.6}$$

Ideally one would like to find a smooth family of feedbacks that makes the family of equilibria asymptotically stable, i.e., for each small μ, the closed loop system

$$z^+ = \tilde{f}(z, \kappa(z, \mu))$$

is asymptotically stable to $z_e(\mu)$. Notice that the lowest degree terms of more general smooth feedbacks will be like (4.6). We restrict our attention to smooth feedbacks for practical and mathematical reasons. Smooth feedbacks are easy to implement, and they allow an analysis of the closed loop system based on the low degree terms.

Clearly the z_2 subsystem is stabilizable for all μ by the proper choice of K_2, and this feedback gain can be chosen independent of μ. The question is: Can we find $K_1(\mu)$ which stabilizes the z_1 coordinate?

Since the linear approximations are stabilizable for $\mu \neq 0$, it is certainly possible to find a stabilizing feedback at each such μ. The linear approximation at $\mu = 0$ has an uncontrollable, unstable mode, so it is not possible to stabilize it. But is it possible to stabilize the approximations for $\mu \neq 0$ with a feedback that is bounded through $\mu = 0$? The answer is no for systems with a fold control bifurcation. For any bounded feedback, the closed loop system will be unstable in some neighborhood of $\mu = 0$.

To see this, note that the closed loop linear approximation

$$\begin{bmatrix} \tilde{z}_1^+ \\ \tilde{z}_2^+ \end{bmatrix} = \left(\begin{bmatrix} A_1 + \mu\left(\gamma_1^{11} + B_1K_1\right) & \mu\left(\Delta + B_1K_2\right) \\ B_2K_1 & A_2 + B_2K_2 \end{bmatrix} + O(\mu)^2 \right) \begin{bmatrix} \tilde{z}_1 \\ \tilde{z}_2 \end{bmatrix} \tag{4.7}$$

is clearly unstable at $\mu = 0$ since it has an eigenvalue $A_1 > 1$. Furthermore, if the feedback $v = K_2(\mu)z_2$ stabilizes the z_2 subsystem, then A_1 is a simple root of the characteristic polynomial of the closed loop system when $\mu = 0$. Hence there is a simple root near A_1 of the characteristic polynomial for all small $|\mu|$.

By using larger and larger gain, it is possible to stabilize the system closer and closer to $\mu = 0$. But if the feedback gain is continuous, at best it will stabilize the closed loop system for only some small but not too small $\mu > 0$ or only some small

but not too small $\mu < 0$. The controllability of z_1 reverses direction (folds) at $\mu = 0$, so a continuous choice of feedback gain cannot stabilize on both sides of $\mu = 0$. If a smooth family of feedbacks (4.6) does stabilize the system for some small $\mu > 0$, the parametrized closed loop system generically undergoes a classical fold bifurcation at some smaller $\mu > 0$. A classical fold bifurcation is also called a limit point bifurcation, a saddle-node bifurcation, or a turning point bifurcation.

We illustrate this with a simple example in normal form:

$$\begin{aligned} z_1^+ &= 2z_1 - z_2^2, \\ z_2^+ &= v. \end{aligned}$$

The equilibria are $z_{e,1} = \mu^2,\ z_{e,2} = \mu,\ v_e = \mu$. Under the feedback $v = K_1(\mu)\tilde{z}_1 + K_2(\mu)\tilde{z}_2$, the closed loop linear approximation is

$$\begin{bmatrix} \tilde{z}_1^+ \\ \tilde{z}_2^+ \end{bmatrix} = \begin{bmatrix} 2 & -2\mu \\ K_1(\mu) & K_2(\mu) \end{bmatrix} \begin{bmatrix} \tilde{z}_1 \\ \tilde{z}_2 \end{bmatrix},$$

where $\tilde{z} = z - z_e(\mu),\ \tilde{v} = v - v_e(\mu)$. If $K(\mu)$ is bounded, then as $\mu \to 0$ one eigenvalue converges to 2, so the system is unstable for small $|\mu|$. If we choose $K_1 = 15/2$ and $K_2 = -1/2$, then the closed loop linear approximation is stable for $|\mu| > 0.1$ and unstable for $|\mu| < 0.1$. It undergoes a fold bifurcation at $\mu = 0.1$.

To see this, consider the closed loop nonlinear system under this feedback in coordinates centered at the bifurcation $\bar{z}_1 = z_1 - 0.01,\ \bar{z}_2 = z_2 - 0.1,\ \bar{\mu} = \mu - 0.1$,

$$\begin{aligned} \bar{z}_1^+ &= 2\bar{z}_1 - 0.2\bar{z}_2 - \bar{z}_2^2, \\ \bar{z}_2^+ &= 7.5\bar{z}_1 - 0.5\bar{z}_2 - 7.5\bar{\mu}^2. \end{aligned}$$

It is convenient to reparametrize by $\nu = 7.5\bar{\mu}^2 \geq 0$. The center manifold is given by

$$\bar{z}_2 = -2\nu + 5\bar{z}_1 + 440\nu^2 - 600\nu\bar{z}_1 + 250\bar{z}_1^2 + O(\bar{z}_1, \nu)^3,$$

and the center manifold dynamics is

$$\bar{z}_1^+ = 0.4\nu + \bar{z}_1 - 92\nu^2 + 140\nu\bar{z}_1 - 75\bar{z}_1^2 + O(\bar{z}_1, \nu)^3$$

or, in the variables $\hat{z}_1 = \sqrt{75}(\bar{z}_1 - 0.9333\nu),\ \hat{\nu} = 0.4\nu - 26.667\nu^2$,

$$\hat{z}_1^+ = \hat{\nu} + \hat{z}_1 - \hat{z}_1^2 + O(\hat{z}_1, \hat{\nu})^3,$$

the familiar form of a discrete time fold bifurcation [9].

4.2. Transcritical control bifurcation. A degenerate form of the above bifurcation occurs when the uncontrollable part is one dimensional and neutrally stable, $A_1 = 1$. The equilibria z_e, v_e depend on roots z_1, μ of the quadratic form

$$0 = \beta_1^{11} z_1^2 + \gamma_1^{11} z_1 \mu + \tilde{\delta}\mu^2.$$

If this form is positive or negative definite, then there is only an isolated equilibrium $z_1 = z_{2,1} = \cdots z_{2,n_2} = v = 0$.

If this form is indefinite but not degenerate, i.e., if it has a positive and a negative eigenvalue, then there are two curves of equilibria that cross at $z_1 = z_{2,1} = \cdots z_{2,n_2} =$

$v = 0$. Suppose that $z_1 = c_k\mu,\ k = 1, 2$, are the two lines of roots of the quadratic form; then the equilibria z_e, v_e are given by

$$\begin{aligned} z_1 &= c_k\mu + O(\mu)^2, && k = 1, 2, \\ z_{e2,i} &= \mu + O(\mu)^2, && i = 1, \ldots, n_2, \\ v_e &= \mu. \end{aligned}$$

Suppose $z_e(\mu),\ v_e(\mu)$ is one curve of equilibria, and one chooses a parametrized family of smooth feedbacks (4.6), where $\tilde{z} = z - z_e(\mu),\ \tilde{v} = v - v_e(\mu)$. Notice that the closed loop system has a single curve of equilibria. The closed loop approximation (4.7) has $\lambda = A_1 = 1$ as an eigenvalue at $\mu = 0$. This eigenvalue is a function $\lambda = \lambda(\mu)$ and

$$\frac{d\lambda}{d\mu}(0) = \gamma_1^{11} + B_1K_1(0),$$

so generically the eigenvalues pass through the unit circle at $\mu = 0$, and the closed loop system goes from stable to unstable through a classical fold bifurcation.

Suppose one chooses a parametrized family of smooth feedbacks that preserves both curves of equilibria,

$$\begin{aligned} &v = \kappa(z, \mu), \\ &\tilde{v} = K_1(\mu)(z_1 - c_1\mu + O(\mu)^2)(z_1 - c_2\mu + O(\mu)^2) + K_2(\mu)\tilde{z}_2. \end{aligned}$$

Then generically the closed loop system undergoes a classical transcritical bifurcation.

If the quadratic form is degenerate, then the locus or loci of equilibria may depend on cubic and higher terms.

4.3. Flip control bifurcation. The next simplest control bifurcation of a discrete time system is the flip. The uncontrollable part is again one dimensional and unstable, but now $A_1 \leq -1$. The equilibria z_e, v_e are conveniently parametrized by $\mu = v_e$. The equilibria $z_e(\mu), v_e(\mu)$ are given by (4.1), and the local linearizations are given by (4.4). If the transversality condition (4.5) is satisfied, then these are controllable when $\mu \neq 0$ but unstabilizable when $\mu = 0$.

One can find a parametrized family of smooth feedbacks (4.6) which will stabilize the z_2 modes for all μ and the z_1 mode for some range of $\mu \neq 0$. If $A_1 < -1$, then as $\mu \to 0$ it requires larger and larger gain to stabilize the latter. To see this, note that the closed loop linear approximation (4.7) is clearly unstable at $\mu = 0$ since $A_1 < -1$. Furthermore, if the feedback $v = K_2(\mu)z_2$ stabilizes the z_2 subsystem, then A_1 is a simple root of the characteristic polynomial of the closed loop system when $\mu = 0$. Hence there is a simple root near A_1 of the characteristic polynomial for all small $|\mu|$.

By using larger and larger gain, it is possible to stabilize the system closer and closer to $\mu = 0$. But if the feedback gain is bounded, at best it will stabilize the closed loop system only for some small but not too small $\mu > 0$ and/or only some small but not too small $\mu < 0$. If a smooth family of feedbacks (4.6) does stabilize the system for some small $\mu > 0$, the parametrized closed loop system generically undergoes a classical flip bifurcation at some smaller $\mu > 0$.

We illustrate this with an example:

$$\begin{aligned} z_1^+ &= -2z_1 + z_2^2, \\ z_2^+ &= v. \end{aligned}$$

The equilibria are $z_{e,1} = \frac{1}{3}\mu^2,\ z_{e,2} = \mu,\ v_e = \mu$. Under the feedback $v = K_1(\mu)\tilde{z}_1 + K_2(\mu)\tilde{z}_2$, the closed loop linear approximation is

$$\begin{bmatrix} \tilde{z}_1^+ \\ \tilde{z}_2^+ \end{bmatrix} = \begin{bmatrix} -2 & 2\mu \\ K_1(\mu) & K_2(\mu) \end{bmatrix} \begin{bmatrix} \tilde{z}_1 \\ \tilde{z}_2 \end{bmatrix},$$

where $\tilde{z} = z - z_e(\mu),\ \tilde{v} = v - v_e(\mu)$. If $K(\mu)$ is bounded, then as $\mu \to 0$ one eigenvalue converges to -2, so the system is unstable for small $|\mu|$. If we choose $K_1 = -15/2$ and $K_2 = 1/2$, then the closed loop linear approximation is stable for $|\mu| > 0.1$ and unstable for $|\mu| < 0.1$. It undergoes a classical flip bifurcation at $\mu = 0.1$.

To see this, consider the closed loop nonlinear system under this feedback in coordinates centered at the the bifurcation $\bar{z}_1 = z_1 - 1/300,\ \bar{z}_2 = z_2 - 1/10,\ \bar{\mu} = \mu - 1/10$,

$$\begin{aligned} \bar{z}_1^+ &= -2\bar{z}_1 + 0.2\bar{z}_2 + \bar{z}_2^2, \\ \bar{z}_2^+ &= -7.5\bar{z}_1 + 0.5\bar{z}_2 + \bar{\mu} + 2.5\bar{\mu}^2. \end{aligned}$$

It is convenient to reparametrize by $\nu = \bar{\mu} + 2.5\bar{\mu}^2$. The center manifold is given by

$$\bar{z}_2 = 0.67\nu + 5\bar{z}_1 - 10.37\nu^2 + 111.11\nu\bar{z}_1 - 83.33\bar{z}_1^2 + O(\bar{z}_1, \nu)^3,$$

and the center manifold dynamics is

$$\begin{aligned} \bar{z}_1^+ = {} & 0.13\nu - \bar{z}_1 - 1.63\nu^2 + 2.89\nu\bar{z}_1 + 8.33\bar{z}_1^2 \\ & -123.12\nu^3 + 1763.0\nu^2\bar{z}_1 - 111.11\nu\bar{z}_1^2 - 1944.4\bar{z}_1^3 + O(\bar{z}_1, \nu)^4, \end{aligned}$$

or, in the variables

$$\begin{aligned} \hat{\nu} &= 3\nu + 172.50\nu^2, \\ \hat{z}_1 &= -2.87\nu + 43.30\bar{z}_1 - 8.02\nu^2 + 24.06\nu\bar{z}_1 - 180.42\bar{z}_1^2 \\ &\quad +48.11\nu^3 - 360.84\nu^2\bar{z}_1 + 2706.3\nu\bar{z}_1, \end{aligned}$$

the parametrized closed loop system is

$$\begin{aligned} \hat{\nu}^+ &= \hat{\nu}, \\ \hat{z}_1^+ &= -\hat{z}_1 + \hat{\nu}\hat{z}_1 - \hat{z}_1^3 + O(\hat{\nu}, \hat{z}_1)^4, \end{aligned}$$

a familiar form of a discrete time flip bifurcation [9].

4.4. Neimark–Sacker control bifurcation. The discrete time analogue of a classical Hopf bifurcation is called a Neimark–Sacker bifurcation. We present the control analogue of this bifurcation. The uncontrollable modes are a nonzero complex conjugate pair,

$$A_1 = \begin{bmatrix} \lambda & 0 \\ 0 & \bar{\lambda} \end{bmatrix},$$

where $\lambda = \rho e^{i\theta},\ \bar{\lambda} = \rho e^{-i\theta},\ \theta \neq 0, \pi/2, \pi, 3\pi/2$. The equilibria $z_e(\mu), v_e(\mu)$ are given by

$$\begin{bmatrix} z_{e1,1} \\ z_{e1,2} \end{bmatrix} = \mu^2(I - A_1)^{-1} \begin{bmatrix} \tilde{\delta}_1 \\ \tilde{\delta}_2 \end{bmatrix} + O(\mu)^3, \tag{4.8}$$

$$z_{e2,i} = \mu + O(\mu)^2, \qquad i = 1, \ldots, n_2, \tag{4.9}$$

$$v_e = \mu, \tag{4.10}$$

where

$$\tilde{\delta}_i = \sum_{k=1}^{n_2+1} \delta_i^{1k}.$$

The local linearization around z_e, v_e is

$$\begin{aligned}\begin{bmatrix} \tilde{z}_1^+ \\ \tilde{z}_2^+ \end{bmatrix} &= \left(\begin{bmatrix} A_1 + \mu\Gamma & \mu\Delta \\ 0 & A_2 \end{bmatrix} + O(\mu)^2 \right) \begin{bmatrix} \tilde{z}_1 \\ \tilde{z}_2 \end{bmatrix} \\ &\quad + \left(\begin{bmatrix} \mu B_1 \\ B_2 \end{bmatrix} + O(\mu)^2 \right) \tilde{v}, \end{aligned} \tag{4.11}$$

where $\tilde{z} = z - z_e(\mu)$, $\tilde{v} = v - v_e(\mu)$, and

$$\Gamma = \begin{bmatrix} \gamma_1^{11} & \gamma_1^{21} \\ \gamma_2^{11} & \gamma_2^{21} \end{bmatrix},$$

$$\Delta = \begin{bmatrix} \tilde{\delta}_1 + \delta_1^{11} & \delta_1^{12} & \dots & \delta_1^{1n_2} \\ \tilde{\delta}_2 + \delta_2^{11} & \delta_2^{12} & \dots & \delta_2^{1n_2} \end{bmatrix},$$

$$B_1 = \begin{bmatrix} \delta_1^{1,n_2+1} \\ \delta_2^{1,n_2+1} \end{bmatrix}.$$

If the transversality condition

$$\begin{bmatrix} \tilde{\delta}_1 + \delta_1^{11} \\ \tilde{\delta}_2 + \delta_2^{11} \end{bmatrix} + A_1 \begin{bmatrix} \delta_1^{12} \\ \delta_2^{12} \end{bmatrix} + \cdots + A_1^{n_2} \begin{bmatrix} \delta_1^{1,n_2+1} \\ \delta_2^{1,n_2+1} \end{bmatrix} \neq 0 \tag{4.12}$$

is satisfied, then the system is linearly controllable and hence stabilizable about any equilibrium except $\mu = 0$. Consider a parametrized family of feedbacks (4.6).

If $\rho < 1$, then the system is stabilizable about any equilibrium, but if $\rho \geq 1$, then the system is not stabilizable when $\mu = 0$. The case $\rho \geq 1$ is called a Neimark–Sacker control bifurcation. We distinguish two subcases, $\rho > 1$ and $\rho = 1$.

If $\rho > 1$, then it requires larger and larger gain to stabilize the system closer and closer to $\mu = 0$. Since the feedback (4.6) is smooth, it will stabilize only for some small $\mu > 0$ or for some small $\mu < 0$ but not both. At $\mu = 0$, the poles of the closed loop system are $\lambda, \bar{\lambda}$ and the poles of $A_2 + B_2 K_2(0)$. The latter can be made stable, but the former are unstable. Since the feedback is bounded, as $\mu \to 0$ the poles converge to these. The system is controllable for $\mu \neq 0$, so the poles can be placed arbitrarily by feedback. The poles associated primarily with the z_2 subsystem can be kept stable, but the two poles associated primarily with the z_1 subsystem will leave the unit disk at some small value(s) of μ. Depending on the choice of feedback, they will leave one at a time as real poles, leave together through ± 1, or leave together as a nonzero complex conjugate pair. If they leave separately as real poles, then generically the closed loop system undergoes a fold or flip bifurcation as the first pole leaves through ± 1. If they leave together as a complex conjugate pair that is neither real nor imaginary, then generically the system undergoes a Neimark–Sacker bifurcation. If they leave together through ± 1, the situation can be quite complicated and will not be discussed here.

If $\rho = 1$ and the feedback (4.6) is continuous, then generically the system undergoes a Neimark–Sacker bifurcation at $\mu = 0$ provided that $e^{ik\theta} \neq 1$ for $k = 1, 2, 3, 4$. We illustrate this with an example:

$$z_{1,1}^+ = e^{i\pi/4} z_{1,1} + z_2^2,$$

$$\begin{aligned} z_{1,2}^+ &= e^{-i\pi/4} z_{1,2} + z_2^2, \\ x_2 &= u. \end{aligned}$$

The equilibria are

$$\begin{aligned} z_{e1,1} &= c\mu^2, \\ z_{e1,2} &= \bar{c}\mu^2, \\ x_{e2} &= \mu, \\ u_e &= \mu, \end{aligned}$$

where $c = \left(1 - e^{i\pi/4}\right)^{-1}$. The linear approximations are

$$\begin{aligned} \tilde{z}_{1,1}^+ &= e^{i\pi/4} \tilde{z}_{1,1} + 2\mu \tilde{z}_2, \\ \tilde{z}_{1,2}^+ &= e^{-i\pi/4} \tilde{z}_{1,2} + 2\mu \tilde{z}_2, \\ \tilde{x}_2^+ &= \tilde{u}, \end{aligned}$$

where $\tilde{z}_{1,1} = z_{1,1} - c\mu^2$, $\tilde{z}_{1,2} = z_{1,2} - \bar{c}\mu^2$, $\tilde{x}_z = z_2 - \mu$, $\tilde{u} = u - \mu$. The linear approximations are controllable except at $\mu = 0$.

The feedback

$$u = \mu + 0.5(z_{1,1} - c\mu^2) + 0.5(z_{1,2} - \bar{c}\mu^2) + 0.5(x_2 - \mu)$$

places the poles of the closed loop system inside the open unit disk at $0.7953 \pm 0.5743i$, 0.3957 at $\mu = 0.1$. A pair of poles leaves the unit disk at $e^{\pm i\pi/4}$ when $\mu = 0$.

The closed loop dynamics undergoes a Neimark–Sacker classical bifurcation at $\mu = 0$. The discrete time analogue of the first Lyapunov coefficient is found in Kuznetsov [9, p. 186, formula (5.74)]. For this example, its value is 46.8, which indicates that the system undergoes a subcritical Neimark–Sacker bifurcation at $\mu = 0$. For small $\mu > 0$, the equilibrium is exponentially stable, but there is an unstable invariant closed curve nearby. For small $\mu < 0$, the equilibrium is unstable as is the bifurcation equilibrium $\mu = 0$.

5. Proof of the quadratic normal form. We can expand the change of coordinates and feedback as follows:

$$\begin{aligned} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} &- \begin{bmatrix} \phi_1^{[2;0]}(x_1; x_2) \\ \phi_2^{[2;0]}(x_1; x_2) \end{bmatrix} \\ &- \begin{bmatrix} \phi_1^{[1;1]}(x_1; x_2) \\ \phi_2^{[1;1]}(x_1; x_2) \end{bmatrix} - \begin{bmatrix} \phi_1^{[0;2]}(x_1; x_2) \\ \phi_2^{[0;2]}(x_1; x_2) \end{bmatrix}, \end{aligned}$$

$$\begin{aligned} v = u &- \alpha^{[2;0]}(x_1; x_2, u) \\ &- \alpha^{[1;1]}(x_1; x_2, u) - \alpha^{[0;2]}(x_1; x_2, u). \end{aligned}$$

These do not change the linear part of the dynamics. The quadratic part of the dynamics is changed to

$$\begin{aligned} \tilde{f}_i^{[d_1;d_2]}(z_1; z_2, v) = f_i^{[d_1;d_2]}(z_1; z_2, v) \\ - \phi_i^{[d_1;d_2]}(A_1 z_1; A_2 z_2) \\ + A_i \phi_i^{[d_1;d_2]}(z_1; z_2) \\ - B_i \alpha^{[d_1;d_2]}(z_1; z_2, v), \end{aligned}$$

where $B_1 = 0$, so the proof splits into six cases, $i = 1, 2$; $d_1 = 0, 1, 2$; $d_2 = 2 - d_1$.

The normal form of $\tilde{f}_2^{[0;2]}(z_1; z_2, v)$. We start by showing that $\tilde{f}_2^{[0;2]}(z_1; z_2, v)$ can be brought into the above form. There are two basic operations, *pull up* and *push down*, which are used to achieve this. Consider a part of the dynamics

$$\begin{aligned} z_{2,i-1}^+ &= z_{2,i} + \cdots, \\ z_{2,i}^+ &= z_{2,i+1} + cz_{2,j}z_{2,k} + \cdots, \\ z_{2,i+1}^+ &= z_{2,i+2} + \cdots, \end{aligned}$$

where $1 < i \le n_2$, $1 \le j \le k \le n_2 + 1$; recall that $z_{2,n+1} = v$.

If $1 < j$, we can *pull up* the quadratic term by defining

$$\bar{z}_{2,i} = z_{2,i} - cz_{2,j-1}z_{2,k-1},$$

and then the dynamics becomes

$$\begin{aligned} z_{2,i-1}^+ &= \bar{z}_{2,i} + cz_{2,j-1}z_{2,k-1} + \cdots, \\ \bar{z}_{2,i}^+ &= z_{2,i+1} + \cdots, \\ z_{2,i+1}^+ &= z_{2,i+2} + \cdots, \end{aligned}$$

and all the other quadratic terms remain the same. Notice that if $i = 1$, we can still pull up, and the term disappears. By pulling up all the quadratic terms until $j = 1$, we obtain

$$z_{2,i}^+ = z_{2,i+1} + cz_{2,1}z_{2,k} + \cdots. \tag{5.1}$$

The other operation on the dynamics is *push down*. If $k \le n_2$, define

$$\bar{z}_{2,i+1} = z_{2,i+1} + cz_{2,j}z_{2,k}$$

yielding

$$\begin{aligned} z_{2,i-1}^+ &= z_{2,i} + \cdots, \\ z_{2,i}^+ &= \bar{z}_{2,i+1} + \cdots, \\ \bar{z}_{2,i+1}^+ &= z_{2,i+2} + cz_{2,j+1}z_{2,k+1} + \cdots, \end{aligned}$$

and all the other quadratic terms remain unchanged. Notice that if $i + 1 = n_2$, then we can absorb the quadratic term into the control using feedback. From (5.1) we push down every term where $k \le i + 1$. These terms can be pushed all the way down and absorbed in the control. The result is (2.8).

Next we show that the number ϵ_i^{1k} (2.13) is an invariant. Clearly ϵ_i^{1k} is potentially changed only by $\phi_2^{[0;2]}(x_1; x_2)$ and $\alpha_2^{[0;2]}(x_1; x_2, u)$. Therefore, we need only consider coordinate changes of the form

$$\bar{x}_{2,\rho} = x_{2,\rho} + cx_{2,\sigma}x_{2,\tau},$$

where $1 \le \rho \le n_2$, $1 \le \sigma \le \tau \le n_2$, and feedbacks of the form

$$\bar{u} = u + cx_{2,\sigma}x_{2,\tau},$$

where $1 \le \sigma \le \tau \le n_2 + 1$ with $x_{2,n+1} = u$. More general coordinate changes and feedbacks are just compositions of these. The coordinate change affects only a piece of the dynamics (2.3),

$$\begin{aligned} x^+_{2,\rho-1} &= x_{2,\rho} + f^{[2]}_{2,\rho-1}(x_1, x_2, u) + O(x_1, x_2, u)^3, \\ x^+_{2,\rho} &= x_{2,\rho+1} + f^{[2]}_{2,\rho}(x_1, x_2, u) + O(x_1, x_2, u)^3 \end{aligned}$$

is transformed to

$$\begin{aligned} x^+_{2,\rho-1} &= \bar{x}_{2,\rho} + f^{[2]}_{2,\rho-1}(x_1, x_2, u) - c x_{2,\sigma} x_{2,\tau} + O(x_1, x_2, u)^3, \\ \bar{x}^+_{2,\rho} &= x_{2,\rho+1} + f^{[2]}_{2,\rho}(x_1, x_2, u) + c x_{2,\sigma+1} x_{2,\tau+1} + O(x_1, x_2, u)^3, \end{aligned}$$

and ϵ^{1k}_i is unchanged. The feedback affects only

$$x^+_{2,n_2} = u + f^{[2]}_{2,n_2}(x_1, x_2, u) + O(x_1, x_2, u)^3,$$

transforming it into

$$x^+_{2,n_2} = \bar{u} + f^{[2]}_{2,n_2}(x_1, x_2, u) - c x_{2,\sigma} x_{2,\tau} + O(x_1, x_2, u)^3,$$

and again ϵ^{1k}_i is unchanged because $i + l \le n_2 - 1$.

The normal form of $\tilde{f}^{[1;1]}_2(z_1; z_2, v)$. The two basic operations, pull up and push down, are slightly different. Consider a part of the dynamics

$$\begin{aligned} z^+_{2,i-1} &= z_{2,i} + \cdots, \\ z^+_{2,i} &= z_{2,i+1} + c z_{1,j} z_{2,k} + \cdots, \\ z^+_{2,i+1} &= z_{2,i+2} + \cdots, \end{aligned}$$

where $1 < i \le n_2$, $1 \le j \le n_1$, $1 \le k \le n_2 + 1$.

If $\lambda_j \ne 0$ and $1 < k$, we can pull up the quadratic term by defining

$$\bar{z}_{2,i} = z_{2,i} - \frac{c}{\lambda_j} z_{1,j} z_{2,k-1};$$

then the dynamics becomes

$$\begin{aligned} z^+_{2,i-1} &= \bar{z}_{2,i} + \frac{c}{\lambda_j} z_{1,j} z_{2,k-1} + \cdots, \\ \bar{z}^+_{2,i} &= z_{2,i+1} + \cdots, \\ z^+_{2,i+1} &= z_{2,i+2} + \cdots, \end{aligned}$$

and all the other quadratic terms remain the same. Again, if $i = 1$, we can still pull up, and the term disappears. So by pulling up all quadratic terms where $\lambda_j \ne 0$ until $k = 1$, we obtain

$$z^+_{2,i} = z_{2,i+1} + c z_{1,j} z_{2,1} + \cdots.$$

Repeated pushing down eliminates this term. Define

$$\bar{z}_{2,i+1} = z_{2,i+1} + c z_{1,j} z_{2,1},$$

yielding

$$\begin{aligned} z^+_{2,i-1} &= z_{2,i} + \cdots, \\ z^+_{2,i} &= \bar{z}_{2,i+1} + \cdots, \\ \bar{z}^+_{2,i+1} &= z_{2,i+2} + c\lambda_j z_{1,j} z_{2,2} + \cdots, \end{aligned}$$

and all the other quadratic terms remain unchanged. If $\lambda_j = 0$, then the term drops out. If $\lambda_j \neq 0$, then we can continue to push down until $i+1 = n_2$ and the quadratic term can be absorbed into the control using feedback. The result is $\tilde{f}_2^{[1;1]}(z_1; z_2, v) = 0$.

The normal form of $\tilde{f}_2^{[2;0]}(z_1; z_2, v)$. Consider a part of the dynamics

$$\begin{aligned} z^+_{2,i} &= z_{2,i+1} + c z_{1,j} z_{1,k} + \cdots, \\ z^+_{2,i+1} &= z_{2,i+2} + \cdots, \end{aligned}$$

where $1 \leq i \leq n_2$, $1 \leq j \leq k \leq n_1$.

Pushing down one or more times eliminates this term. Define

$$\bar{z}_{2,i+1} = z_{2,i+1} + c z_{1,j} z_{1,k},$$

yielding

$$\begin{aligned} z^+_{2,i-1} &= z_{2,i} + \cdots, \\ z^+_{2,i} &= \bar{z}_{2,i+1} + \cdots, \\ \bar{z}^+_{2,i+1} &= z_{2,i+2} + c\lambda_j\lambda_k z_{1,j} z_{1,k} + \cdots, \end{aligned}$$

and all the other quadratic terms remain unchanged. If $\lambda_j \lambda_k = 0$, then the term drops out. Otherwise, the quadratic term can be pushed down repeatedly until it is absorbed in the control. The result is $\tilde{f}_2^{[2;0]}(z_1; z_2, v) = 0$.

The normal form of $\tilde{f}_1^{[2;0]}(z_1; z_2, v)$. This is just the quadratic normal form of Poincaré as described in the introduction, and β_i^{jk} are the invariants. See [1], [5], [9], or [13]. Consider a part of the dynamics

$$z^+_{1,i} = \lambda_i z_{1,i} + c z_{1,j} z_{1,k} + \cdots,$$

where $1 \leq i \leq n_1$, $1 \leq j \leq k \leq n_1$.

If $\lambda_i \neq \lambda_j \lambda_k$, then define

$$\bar{z}_{1,i} = z_{1,i} - \frac{c}{(\lambda_j \lambda_k - \lambda_i)} z_{1,j} z_{1,k}$$

so that

$$\bar{z}^+_{1,i} = \lambda_i \bar{z}_{1,i} + \cdots.$$

Next we show that the numbers β_i^{jk} (2.9) are invariants. Clearly β_i^{jk} is potentially changed only by $\phi_1^{[2;0]}(x_1; x_2)$. Therefore, we need only consider coordinate changes of the form

$$\bar{x}_{1,\rho} = x_{1,\rho} + c x_{1,\sigma} x_{1,\tau},$$

where $1 \le \rho \le n_1$, $1 \le \sigma \le \tau \le n_1$, because more general ones are just compositions of these. This coordinate change affects only a piece of the dynamics (2.3), and

$$x^+_{1,\rho} = \lambda_\rho x_{1,\rho} + f^{[2]}_{1,\rho}(x_1, x_2, u) + O(x_1, x_2, u)^3$$

is transformed to

$$\bar{x}^+_{1,\rho} = \lambda_\rho \bar{x}_{1,\rho} + f^{[2]}_{1,\rho}(x_1, x_2, u) + c(\lambda_\sigma \lambda_\tau - \lambda_\rho) x_{1,\sigma} x_{1,\tau} + O(x_1, x_2, u)^3.$$

Clearly, if $\lambda_\rho = \lambda_\sigma \lambda_\tau$, then β^{jk}_i (2.9) is unchanged.

The normal form of $\tilde{f}^{[1;1]}_1(z_1; z_2, v)$. Consider a part of the dynamics

$$z^+_{1,i} = \lambda_i z_{1,i} + c z_{1,j} z_{2,k} + \cdots,$$

where $1 \le i \le n_1$, $1 \le j \le n_1$, $1 \le k \le n_2 + 1$.

If $\lambda_j \neq 0$ and $k > 1$, then we can pull up by defining

$$\bar{z}_{1,i} = z_{1,i} - \frac{c}{\lambda_j} z_{1,j} z_{2,k-1}$$

so that

$$\bar{z}^+_{1,i} = \lambda_i \bar{z}_{1,i} + \frac{c\lambda_i}{\lambda_j} z_{1,j} z_{2,k-1} + \cdots.$$

If $\lambda_i = 0$, then the term disappears; otherwise, we can continue to pull up until $k = 1$.

If $\lambda_i \neq 0$, then we can push down by defining

$$\bar{z}_{1,i} = z_{1,i} + \frac{c}{\lambda_i} z_{1,j} z_{2,k};$$

then

$$\bar{z}^+_{1,i} = \lambda_i \bar{z}_{1,i} + \frac{c\lambda_j}{\lambda_i} z_{1,j} z^+_{2,k} + \cdots.$$

If $\lambda_j = 0$, then the term disappears.

If $\lambda_i = \lambda_j = 0$, then we cannot pull up or push down. The result is (2.7).

Next we show that the numbers γ^{jk}_i (2.10)–(2.11) are invariants. Clearly γ^{jk}_i is potentially changed only by $\phi^{[1;1]}_1(x_1; x_2)$. Therefore, we need only consider coordinate changes of the form

$$\bar{x}_{1,\rho} = x_{1,\rho} + c x_{1,\sigma} x_{2,\tau},$$

where $1 \le \rho, \sigma \le n_1$, $1 \le \tau \le n_2$, because more general ones are just compositions of these. This coordinate change affects only a piece of the dynamics (2.3), and

$$x^+_{1,\rho} = \lambda_\rho x_{1,\rho} + f^{[2]}_{1,\rho}(x_1, x_2, u) + O(x_1, x_2, u)^3$$

is transformed to

$$\bar{x}^+_{1,\rho} = \lambda_\rho \bar{x}_{1,\rho} + f^{[2]}_{1,\rho}(x_1, x_2, u) + c\lambda_\sigma x_{1,\sigma} x_{2,\tau+1} - c\lambda_\rho x_{1,\sigma} x_{2,\tau} + O(x_1, x_2, u)^3.$$

Clearly, if $\lambda_\rho = \lambda_\sigma = 0$, then γ^{jk}_i (2.10) is unchanged. A simple calculation shows that if $\lambda_\rho \lambda_\sigma \neq 0$, then γ^{j1}_i (2.11) is unchanged.

The normal form of $\tilde{f}_1^{[0;2]}(z_1; z_2, v)$. Consider a part of the dynamics

$$z_{1,i}^+ = \lambda_i z_{1,i} + c z_{2,j} z_{2,k} + \cdots,$$

where $1 \le i \le n_1$, $1 \le j \le k \le n_2$.

If $j > 1$, then we can pull up by defining

$$\bar{z}_{1,i} = z_{1,i} - c z_{2,j-1} z_{2,k-1};$$

then

$$\bar{z}_{1,i}^+ = \lambda_i \bar{z}_{1,i} + c\lambda_i z_{2,j-1} z_{2,k-1} + \cdots.$$

If $\lambda_i = 0$, then the term disappears; otherwise, we can continue to pull up until $j = 1$. The result is (2.7).

Finally, we show that the numbers δ_i^{1k} (2.12) are invariants. Clearly, δ_i^{1k} is potentially changed only by $\phi_1^{[0;2]}(x_1; x_2)$. Therefore, we need only consider coordinate changes of the form

$$\bar{x}_{1,\rho} = x_{1,\rho} + c x_{2,\sigma} x_{2,\tau},$$

where $1 \le \rho \le n_1$, $1 \le \sigma \le \tau \le n_2$, because more general ones are just compositions of these. This change of coordinates affects only a piece of the dynamics (2.3), and

$$x_{1,\rho}^+ = \lambda_\rho x_{1,\rho} + f_{1,\rho}^{[2]}(x_1, x_2, u) + O(x_1, x_2, u)^3$$

is transformed to

$$\bar{x}_{1,\rho}^+ = \lambda_\rho \bar{x}_{1,\rho} + f_{1,\rho}^{[2]}(x_1, x_2, u) + c x_{2,\sigma+1} x_{2,\tau+1} - c\lambda_\rho x_{2,\sigma} x_{2,\tau} + O(x_1, x_2, u)^3.$$

Clearly, if $\lambda_\rho \ne 0$, then δ_i^{1k} (2.12) is unchanged.

6. Proof of the cubic normal form. Cubic changes of coordinates and cubic feedbacks do not change the linear and quadratic parts of the system. Their effect on the cubic part of the system splits into cases, this time eight cases, $i = 1, 2$; $d_1 = 0, 1, 2, 3$; $d_2 = 3 - d_1$.

The normal form of $\tilde{f}_2^{[0;3]}(z_1; z_2, v)$. We again use the two basic operations pull up and push down. Consider a part of the dynamics

$$\begin{aligned} z_{2,i-1}^+ &= z_{2,i} + f_{2,i-1}^{[2]}(z_1, z_2, v) + \cdots, \\ z_{2,i}^+ &= z_{2,i+1} + f_{2,i}^{[2]}(z_1, z_2, v) + c z_{2,j} z_{2,k} z_{2,l} + \cdots, \\ z_{2,i+1}^+ &= z_{2,i+2} + f_{2,i+1}^{[2]}(z_1, z_2, v) + \cdots, \end{aligned}$$

where $1 < i \le n_2$, $1 \le j \le k \le l \le n_2 + 1$; recall that $z_{2,n+1} = v$.

If $1 < j$, we can *pull up* the cubic term by defining

$$\bar{z}_{2,i} = z_{2,i} - c z_{2,j-1} z_{2,k-1} z_{2,l-1};$$

then the dynamics becomes

$$\begin{aligned} z_{2,i-1}^+ &= \bar{z}_{2,i} + f_{2,i-1}^{[2]}(z_1, z_2, v) + c z_{2,j-1} z_{2,k-1} z_{2,l-1} + \cdots, \\ \bar{z}_{2,i}^+ &= z_{2,i+1} + f_{2,i}^{[2]}(z_1, z_2, v) + \cdots, \\ z_{2,i+1}^+ &= z_{2,i+2} + f_{2,i+1}^{[2]}(z_1, z_2, v) + \cdots, \end{aligned}$$

and all the other cubic terms remain the same. Notice that if $i = 1$, we can still pull up, and the term disappears. By pulling up all cubic terms until $j = 1$, we obtain that

$$z^+_{2,i} = z_{2,i+1} + f^{[2]}_{2,i}(z_1, z_2, v) + cz_{2,1}z_{2,k}z_{2,l} + \cdots. \tag{6.1}$$

The other operation on the dynamics is *push down*. If $l \le n_2$, define

$$\bar{z}_{2,i+1} = z_{2,i+1} + cz_{2,j}z_{2,k}z_{2,l},$$

yielding

$$\begin{aligned} z^+_{2,i-1} &= z_{2,i} + f^{[2]}_{2,i-1}(z_1, z_2, v) + \cdots, \\ z^+_{2,i} &= \bar{z}_{2,i+1} + f^{[2]}_{2,i}(z_1, z_2, v) + \cdots, \\ \bar{z}^+_{2,i+1} &= z_{2,i+2} + f^{[2]}_{2,i+1}(z_1, z_2, v) + cz_{2,j+1}z_{2,k+1}z_{2,l+1} + \cdots, \end{aligned}$$

and all the other cubic terms remain unchanged. Notice that if $i + 1 = n_2$, then we can absorb the cubic term into the control. From (6.1) we push down every term where $l \le i + 1$. These terms can be pushed all the way down and absorbed in the control. The result is (3.8).

Next we show that the number η^{1kl}_i (3.16) is an invariant. Clearly η^{1kl}_i is potentially changed only by $\phi^{[0;3]}_2(x_1; x_2)$ and $\alpha^{[0;3]}_2(x_1; x_2, u)$. Therefore, we need only consider coordinate changes of the form

$$\bar{x}_{2,\rho} = x_{2,\rho} + cx_{2,\sigma}x_{2,\tau}x_{2,\upsilon},$$

where $1 \le \rho \le n_2$, $1 \le \sigma \le \tau \le \upsilon \le n_2$, and feedbacks of the form

$$\bar{u} = u + cx_{2,\sigma}x_{2,\tau}x_{2,\upsilon},$$

where $1 \le \rho \le n_2, 1 \le \sigma \le \tau \le \upsilon \le n_2 + 1$ with $x_{2,n+1} = u$ because more general ones are just compositions of these. The coordinate change affects only a piece of the dynamics (3.1),

$$\begin{aligned} x^+_{2,\rho-1} &= x_{2,\rho} + f^{[2]}_{2,\rho-1}(x_1, x_2, u) + f^{[3]}_{2,\rho-1}(x_1, x_2, u), \\ x^+_{2,\rho} &= x_{2,\rho+1} + f^{[2]}_{2,\rho}(x_1, x_2, u) + f^{[3]}_{2,\rho}(x_1, x_2, u) \end{aligned}$$

is transformed to

$$\begin{aligned} x^+_{2,\rho-1} &= \bar{x}_{2,\rho} + f^{[2]}_{2,\rho-1}(x_1, x_2, u) + f^{[3]}_{2,\rho-1}(x_1, x_2, u) - cx_{2,\sigma}x_{2,\tau}x_{2,\upsilon}, \\ \bar{x}^+_{2,\rho} &= x_{2,\rho+1} + f^{[2]}_{2,\rho}(x_1, x_2, u) + f^{[3]}_{2,\rho}(x_1, x_2, u) + cx_{2,\sigma+1}x_{2,\tau+1}x_{2,\upsilon+1}, \end{aligned}$$

and η^{1kl}_i is unchanged. The feedback affects only

$$x^+_{2,n_2} = u + f^{[2]}_{2,n_2}(x_1, x_2, u) + f^{[3]}_{2,n_2}(x_1, x_2, u),$$

transforming it into

$$x^+_{2,n_2} = \bar{u} + f^{[2]}_{2,n_2}(x_1, x_2, u) + f^{[3]}_{2,n_2}(x_1, x_2, u) - cx_{2,\sigma}x_{2,\tau}x_{2,\upsilon},$$

and again η^{1kl}_i is unchanged because $i + r \le n_2 - 1$.

The normal form of $\tilde{f}_2^{[1;2]}(z_1; z_2, v)$. The two basic operations, pull up and push down, are slightly different. Consider a part of the dynamics

$$\begin{aligned} z_{2,i-1}^+ &= z_{2,i} + f_{2,i-1}^{[2]}(z_1, z_2, v) + \cdots, \\ z_{2,i}^+ &= z_{2,i+1} + f_{2,i}^{[2]}(z_1, z_2, v) + cz_{1,j}z_{2,k}z_{2,l} + \cdots, \\ z_{2,i+1}^+ &= z_{2,i+2} + f_{2,i}^{[2]}(z_1, z_2, v) + \cdots, \end{aligned}$$

where $1 < i \le n_2$, $1 \le j \le n_1$, $1 \le k \le l \le n_2 + 1$.

If $\lambda_j \neq 0$ and $1 < k$, we can pull up the cubic term by defining

$$\bar{z}_{2,i} = z_{2,i} - \frac{c}{\lambda_j} z_{1,j} z_{2,k-1} z_{2,l-1};$$

then the dynamics becomes

$$\begin{aligned} z_{2,i-1}^+ &= \bar{z}_{2,i} + \frac{c}{\lambda_j} z_{1,j} z_{2,k-1} z_{2,l-1} + \cdots, \\ \bar{z}_{2,i}^+ &= z_{2,i+1} + \cdots, \\ z_{2,i+1}^+ &= z_{2,i+2} + \cdots, \end{aligned}$$

and all the other cubic terms remain the same. Again, if $i = 1$, we can still pull up, and the term disappears. So by pulling up all cubic terms where $\lambda_j \neq 0$ until $k = 1$, we obtain

$$z_{2,i}^+ = z_{2,i+1} + f_{2,i}^{[2]}(z_1, z_2, v) + cz_{1,j}z_{2,1}z_{2,l} + \cdots.$$

If $l \le n_2$, we can also push down by defining

$$\bar{z}_{2,i+1} = z_{2,i+1} + cz_{1,j}z_{2,1}z_{2,l},$$

yielding

$$\begin{aligned} z_{2,i-1}^+ &= z_{2,i} + f_{2,i-1}^{[2]}(z_1, z_2, v) + \cdots, \\ z_{2,i}^+ &= \bar{z}_{2,i+1} + f_{2,i}^{[2]}(z_1, z_2, v) + \cdots, \\ \bar{z}_{2,i+1}^+ &= z_{2,i+2} + f_{2,i+1}^{[2]}(z_1, z_2, v) + c\lambda_j z_{1,j} z_{2,2} z_{2,l+1} + \cdots, \end{aligned}$$

and all the other cubic terms remain unchanged. If $\lambda_j = 0$, then the term drops out. If $\lambda_j \neq 0$ and $l \le i + 1$, then the cubic term can be pushed down repeatedly and absorbed in the control. The result is (3.7).

Next we show that the number ζ_i^{j1l} (3.15) is an invariant. Clearly ζ_i^{j1l} is potentially changed only by $\phi_2^{[1;2]}(x_1; x_2)$ and $\alpha_2^{[1;2]}(x_1; x_2, u)$. Therefore, we need only consider coordinate changes of the form

$$\bar{x}_{2,\rho} = x_{2,\rho} + cx_{1,\sigma}x_{2,\tau}x_{2,\upsilon},$$

where $1 \le \rho \le n_2$, $1 \le \sigma \le n_1$, $1 \le \tau \le \upsilon \le n_2$ and feedbacks of the form

$$\bar{u} = u + cx_{1,\sigma}x_{2,\tau}x_{2,\upsilon},$$

where $1 \le \sigma \le n_1$, $1 \le \tau \le \upsilon \le n_2 + 1$ with $x_{2,n+1} = u$ because more general ones are just compositions of these. The coordinate change affects only a piece of the dynamics (3.1),

$$\begin{aligned} x^+_{2,\rho-1} &= x_{2,\rho} + f^{[2]}_{2,\rho-1}(x_1, x_2, u) + f^{[3]}_{2,\rho-1}(x_1, x_2, u), \\ x^+_{2,\rho} &= x_{2,\rho+1} + f^{[2]}_{2,\rho}(x_1, x_2, u) + f^{[3]}_{2,\rho}(x_1, x_2, u) \end{aligned}$$

is transformed to

$$\begin{aligned} x^+_{2,\rho-1} &= \bar{x}_{2,\rho} + f^{[2]}_{2,\rho-1}(x_1, x_2, u) + f^{[3]}_{2,\rho-1}(x_1, x_2, u) - c x_{1,\sigma} x_{2,\tau} x_{2,\upsilon}, \\ \bar{x}^+_{2,\rho} &= x_{2,\rho+1} + f^{[2]}_{2,\rho}(x_1, x_2, u) + f^{[3]}_{2,\rho}(x_1, x_2, u) + c\lambda_\sigma x_{1,\sigma} x_{2,\tau+1} x_{2,\upsilon+1}, \end{aligned}$$

and ζ_i^{j1l} is unchanged. The feedback affects only

$$x^+_{2,n_2} = u + f^{[2]}_{2,n_2}(x_1, x_2, u) + f^{[3]}_{2,n_2}(x_1, x_2, u),$$

transforming it into

$$x^+_{2,n_2} = \bar{u} + f^{[2]}_{2,n_2}(x_1, x_2, u) + f^{[3]}_{2,n_2}(x_1, x_2, u) - c x_{1,\sigma} x_{2,\tau} x_{2,\upsilon},$$

and again ζ_i^{j1l} is unchanged.

The normal form of $\tilde{f}_2^{[2;1]}(z_1; z_2, v)$. Consider a part of the dynamics

$$\begin{aligned} z^+_{2,i-1} &= z_{2,i} + f^{[2]}_{2,i-1}(z_1, z_2, v) + \cdots, \\ z^+_{2,i} &= z_{2,i+1} + f^{[2]}_{2,i}(z_1, z_2, v) + c z_{1,j} z_{1,k} z_{2,l} + \cdots, \\ z^+_{2,i+1} &= z_{2,i+2} + f^{[2]}_{2,i+1}(z_1, z_2, v) + \cdots, \end{aligned}$$

where $1 < i \le n_2$, $1 \le j \le k \le n_1$, $1 \le l \le n_2 + 1$.

If $\lambda_j \lambda_k \neq 0$ and $1 < l$, we can pull up the cubic term by defining

$$\bar{z}_{2,i} = z_{2,i} - \frac{c}{\lambda_j \lambda_k} z_{1,j} z_{1,k} z_{2,l-1};$$

then the dynamics becomes

$$\begin{aligned} z^+_{2,i-1} &= \bar{z}_{2,i} + f^{[2]}_{2,i-1}(z_1, z_2, v) + \frac{c}{\lambda_j \lambda_k} z_{1,j} z_{1,k} z_{2,l-1} + f^{[2]}_{2,i}(z_1, z_2, v) + \cdots, \\ \bar{z}^+_{2,i} &= z_{2,i+1} + f^{[2]}_{2,i}(z_1, z_2, v) + \cdots, \\ z^+_{2,i+1} &= z_{2,i+2} + f^{[2]}_{2,i+1}(z_1, z_2, v) + \cdots, \end{aligned}$$

and all the other cubic terms remain the same. Again, if $i = 1$, we can still pull up, and the term disappears. So by pulling up all cubic terms where $\lambda_j \lambda_k \neq 0$ until $l = 1$, we obtain

$$z^+_{2,i} = z_{2,i+1} + c z_{1,j} z_{1,k} z_{2,1} + \cdots.$$

Pushing down eliminates this term and any term with $\lambda_j \lambda_k = 0$. Define

$$\bar{z}_{2,i+1} = z_{2,i+1} + c z_{1,j} z_{1,k} z_{2,1},$$

yielding

$$\begin{aligned} z^+_{2,i-1} &= z_{2,i} + f^{[2]}_{2,i-1}(z_1, z_2, v) + \cdots, \\ z^+_{2,i} &= \bar{z}_{2,i+1} + f^{[2]}_{2,i}(z_1, z_2, v) + \cdots, \\ \bar{z}^+_{2,i+1} &= z_{2,i+2} + f^{[2]}_{2,i+1}(z_1, z_2, v) + c\lambda_j\lambda_k z_{1,j} z_{1,k} z_{2,2} + \cdots, \end{aligned}$$

and all the other cubic terms remain unchanged. If $\lambda_j\lambda_k = 0$, then the term drops out. If $\lambda_j\lambda_k \neq 0$, then we can push down repeatedly until the cubic term is absorbed in the control. The result is $\tilde{f}^{[2;1]}_2(z_1; z_2, v) = 0$.

The normal form of $\tilde{f}^{[3;0]}_2(z_1; z_2, v)$. Consider a part of the dynamics

$$\begin{aligned} z^+_{2,i} &= z_{2,i+1} + f^{[2]}_{2,i}(z_1, z_2, v) + c z_{1,j} z_{1,k} z_{1,l} + \cdots, \\ z^+_{2,i+1} &= z_{2,i+2} + f^{[2]}_{2,i+1}(z_1, z_2, v) + \cdots, \end{aligned}$$

where $1 \le i \le n_2$, $1 \le j \le k \le l \le n_1$.

Pushing down one or more times eliminates this term. Define

$$\bar{z}_{2,i+1} = z_{2,i+1} + c z_{1,j} z_{1,k} z_{1,l},$$

yielding

$$\begin{aligned} z^+_{2,i} &= \bar{z}_{2,i+1} + f^{[2]}_{2,i}(z_1, z_2, v) + \cdots, \\ \bar{z}^+_{2,i+1} &= z_{2,i+2} + f^{[2]}_{2,i+1}(z_1, z_2, v) + c\lambda_j\lambda_k\lambda_l z_{1,j} z_{1,k} z_{1,l} + \cdots, \end{aligned}$$

and all the other cubic terms remain unchanged. If $\lambda_j\lambda_k\lambda_l = 0$, then the term drops out. Otherwise, the term can be pushed down repeatedly until it is absorbed in the control. The result is $\tilde{f}^{[3;0]}_2(z_1; z_2, v) = 0$.

The normal form of $\tilde{f}^{[3;0]}_1(z_1; z_2, v)$. This is just the cubic normal form and invariants of Poincaré (see [1], [5], [9], and [13]). Consider a part of the dynamics

$$z^+_{1,i} = \lambda_i z_{1,i} + f^{[2]}_{1,i}(z_1, z_2, v) + c z_{1,j} z_{1,k} z_{1,l} + \cdots,$$

where $1 \le i \le n_1$, $1 \le j \le k \le l \le n_1$.

If $\lambda_i \neq \lambda_j\lambda_k\lambda_l$, then define

$$\bar{z}_{1,i} = z_{1,i} - \frac{c}{(\lambda_j\lambda_k\lambda_l - \lambda_i)} z_{1,j} z_{1,k} z_{1,l}$$

so that

$$\bar{z}^+_{1,i} = \lambda_i \bar{z}_{1,i} + f^{[2]}_{1,i}(z_1, z_2, v) + \cdots.$$

Next we show that the numbers β^{jkl}_i (3.9) are invariants. Clearly, β^{jkl}_i is potentially changed only by $\phi^{[3;0]}_1(x_1; x_2)$. Therefore, we need only consider coordinate changes of the form

$$\bar{x}_{1,\rho} = x_{1,\rho} + c x_{1,\sigma} x_{1,\tau} x_{1,\upsilon},$$

where $1 \le \rho \le n_1$, $1 \le \sigma \le \tau \le \upsilon \le n_1$ because more general ones are just compositions of these. This coordinate change affects only a piece of the dynamics (3.1), and

$$x^+_{1,\rho} = \lambda_\rho x_{1,\rho} + f^{[2]}_{1,\rho}(x_1, x_2, u) + f^{[3]}_{1,\rho}(x_1, x_2, u) + O(x_1, x_2, u)^3$$

is transformed to

$$\bar{x}^+_{1,\rho} = \lambda_\rho \bar{x}_{1,\rho} + f^{[2]}_{1,\rho}(x_1, x_2, u) + f^{[3]}_{1,\rho}(x_1, x_2, u) + c(\lambda_\sigma\lambda_\tau\lambda_\upsilon - \lambda_\rho)x_{1,\sigma}x_{1,\tau} + O(x_1, x_2, u)^3.$$

Clearly, if $\lambda_\rho = \lambda_\sigma\lambda_\tau\lambda_\upsilon$, then β^{jkl}_i (3.9) is unchanged.

The normal form of $\tilde{f}^{[2;1]}_1(z_1; z_2, v)$. Consider a part of the dynamics

$$z^+_{1,i} = \lambda_i z_{1,i} + f^{[2]}_{1,i}(z_1, z_2, v) + cz_{1,j}z_{1,k}z_{2,l} + \cdots,$$

where $1 \le i \le n_1$, $1 \le j \le k \le n_1$, $1 \le l \le n_2$.

If $\lambda_j\lambda_k \neq 0$ and $l > 1$, we can pull up by defining

$$\bar{z}_{1,i} = z_{1,i} - \frac{c}{\lambda_j\lambda_k} z_{1,j}z_{1,k}z_{2,l-1}$$

so that

$$\bar{z}^+_{1,i} = \lambda_i \bar{z}_{1,i} + f^{[2]}_{1,i}(z_1, z_2, v) + \frac{c\lambda_i}{\lambda_j\lambda_k} z_{1,j}z_{1,k}z_{2,l-1} + \cdots.$$

If $\lambda_i = 0$, then the term disappears; otherwise, we can continue to pull up until $l = 1$.

If $\lambda_i \neq 0$ and $\lambda_j\lambda_k = 0$, then the term disappears by pushing down

$$\bar{z}_{1,i} = z_{1,i} + \frac{c}{\lambda_i} z_{1,j}z_{1,k}z_{2,l}$$

so that

$$\begin{aligned}\bar{z}^+_{1,i} &= \lambda_i \bar{z}_{1,i} + f^{[2]}_{1,i}(z_1, z_2, v) + \frac{c\lambda_j\lambda_k}{\lambda_i} z_{1,j}z_{1,k}z^+_{2,l} + \cdots, \\ &= \lambda_i \bar{z}_{1,i} + f^{[2]}_{1,i}(z_1, z_2, v) + \cdots.\end{aligned}$$

If $\lambda_i = \lambda_j\lambda_k = 0$, then we cannot pull up or push down. The result is (3.4).

Next we show that the numbers γ^{jkl}_i (3.10)–(3.11) are invariants. Clearly, γ^{jkl}_i is potentially changed only by $\phi^{[2;1]}_1(x_1; x_2)$. Therefore, we need only consider coordinate changes of the form

$$\bar{x}_{1,\rho} = x_{1,\rho} + cx_{1,\sigma}x_{1,\tau}x_{2,\upsilon},$$

where $1 \le \rho \le n_1$, $1 \le \sigma \le \tau \le n_1$, $1 \le \upsilon \le n_2$ because more general ones are just compositions of these. This coordinate change affects only a piece of the dynamics (3.1), and

$$x^+_{1,\rho} = \lambda_\rho x_{1,\rho} + f^{[2]}_{1,\rho}(x_1, x_2, u) + f^{[3]}_{1,\rho}(x_1, x_2, u) + O(x_1, x_2, u)^4$$

is transformed to

$$\begin{aligned}\bar{x}^+_{1,\rho} = \lambda_\rho \bar{x}_{1,\rho} &+ f^{[2]}_{1,\rho}(x_1, x_2, u) + f^{[3]}_{1,\rho}(x_1, x_2, u) \\ &+ c\lambda_\sigma\lambda_\tau x_{1,\sigma}x_{1,\tau}x_{2,\upsilon+1} - c\lambda_\rho x_{1,\sigma}x_{1,\tau}x_{2,\upsilon} + O(x_1, x_2, u)^4.\end{aligned}$$

Clearly, if $\lambda_\rho = \lambda_\sigma\lambda_\tau = 0$, then γ^{jkl}_i (3.10) is unchanged. A simple calculation shows that if $\lambda_\rho\lambda_\sigma\lambda_\tau \neq 0$, then γ^{jk1}_i (3.11) is unchanged.

The normal form of $\tilde{f}_1^{[1;2]}(z_1; z_2, v)$. Consider a part of the dynamics

$$z_{1,i}^+ = \lambda_i z_{1,i} + f_{1,i}^{[2]}(z_1, z_2, v) + c z_{1,j} z_{2,k} z_{2,l} + \cdots,$$

where $1 \le i \le n_1$, $1 \le j \le n_1$, $1 \le k \le l \le n_2$.

If $\lambda_j \neq 0$ and $k > 1$, then we can pull up by defining

$$\bar{z}_{1,i} = z_{1,i} - \frac{c}{\lambda_j} z_{1.j} z_{2,k-1} z_{2,l-1};$$

then

$$\bar{z}_{1,i}^+ = \lambda_i \bar{z}_{1,i} + f_{1,i}^{[2]}(z_1, z_2, v) + \frac{c\lambda_i}{\lambda_j} z_{1,j} z_{2,k-1} z_{2,l-1} + \cdots.$$

If $\lambda_i = 0$, then the term disappears; otherwise, we can continue to pull up until $k = 1$.

If $\lambda_i \neq 0$ and $\lambda_j = 0$, then the term disappears by pushing down

$$\bar{z}_{1,i} = z_{1,i} + \frac{c}{\lambda_i} z_{1,j} z_{2,k} z_{2,l};$$

then

$$\begin{aligned} \bar{z}_{1,i}^+ &= \lambda_i \bar{z}_{1,i} + f_{1,i}^{[2]}(z_1, z_2, v) + \frac{c\lambda_j}{\lambda_i} z_{1,j} z_{1,k}^+ z_{2,l}^+ + \cdots, \\ &= \lambda_i \bar{z}_{1,i} + f_{1,i}^{[2]}(z_1, z_2, v) + \cdots. \end{aligned}$$

If $\lambda_i = \lambda_j = 0$, then we cannot pull up or push down. The result is (3.5).

Next we show that the numbers δ_i^{jkl} (3.12)–(3.13) are invariants. Clearly, δ_i^{jkl} is potentially changed only by $\phi_1^{[1;2]}(x_1; x_2)$. Therefore, we need only consider coordinate changes of the form

$$\bar{x}_{1,\rho} = x_{1,\rho} + c x_{1,\sigma} x_{2,\tau} x_{2,\upsilon},$$

where $1 \le \rho, \sigma \le n_1$, $1 \le \tau \le \upsilon \le n_2$ because more general ones are just compositions of these. This coordinate change affects only a piece of the dynamics (3.1), and

$$x_{1,\rho}^+ = \lambda_\rho x_{1,\rho} + f_{1,\rho}^{[2]}(x_1, x_2, u) + f_{1,\rho}^{[3]}(x_1, x_2, u) + O(x_1, x_2, u)^4$$

is transformed to

$$\begin{aligned} \bar{x}_{1,\rho}^+ = \lambda_\rho \bar{x}_{1,\rho} &+ f_{1,\rho}^{[2]}(x_1, x_2, u) + f_{1,\rho}^{[3]}(x_1, x_2, u) \\ &+ c\lambda_\sigma \lambda_\tau x_{1,\sigma} x_{1,\tau} x_{2,\upsilon+1} - c\lambda_\rho x_{1,\sigma} x_{1,\tau} x_{2,\upsilon} + O(x_1, x_2, u)^4. \end{aligned}$$

Clearly, if $\lambda_\rho = \lambda_\sigma \lambda_\tau = 0$, then δ_i^{jkl} (3.12) is unchanged. A simple calculation shows that if $\lambda_\rho \lambda_\sigma \lambda_\tau \neq 0$, then δ_i^{j1l} (3.13) is unchanged.

The normal form of $\tilde{f}_1^{[0;3]}(z_1; z_2, v)$. Consider a part of the dynamics

$$z_{1,i}^+ = \lambda_i z_{1,i} + f_{1,i}^{[2]}(z_1, z_2, v) + c z_{2,j} z_{2,k} z_{2,l} + \cdots,$$

where $1 \le i \le n_1$, $1 \le j \le k \le l \le n_2$.

If $j > 1$, we can pull up by defining

$$\bar{z}_{1,i} = z_{1,i} - c z_{2,j-1} z_{2,k-1} z_{2,l-1};$$

then

$$\bar{z}^+_{1,i} = \lambda_i \bar{z}_{1,i} + f^{[2]}_{1,i}(z_1, z_2, v) + c\lambda_i z_{2,j-1} z_{2,k-1} z_{2,l-1} + \cdots .$$

If $\lambda_i = 0$, then the term disappears; otherwise, we can continue to pull up until $j = 1$. The result is (3.6).

Finally, we show that the numbers ϵ^{1kl}_i (3.14) are invariants. Clearly, ϵ^{1kl}_i is potentially changed only by $\phi^{[0;3]}_1(x_1; x_2)$. Therefore, we need only consider coordinate changes of the form

$$\bar{x}_{1,\rho} = x_{1,\rho} + c x_{2,\sigma} x_{2,\tau} x_{2,\upsilon},$$

where $1 \le \rho \le n_1$, $1 \le \sigma \le \tau \le \upsilon \le n_2$ because more general ones are just compositions of these. This change of coordinates affects only a piece of the dynamics (2.3), and

$$x^+_{1,\rho} = \lambda_\rho x_{1,\rho} + f^{[2]}_{1,\rho}(x_1, x_2, u) + f^{[3]}_{1,\rho}(x_1, x_2, u)$$

is transformed to

$$\begin{aligned} \bar{x}^+_{1,\rho} = \lambda_\rho \bar{x}_{1,\rho} &+ f^{[2]}_{1,\rho}(x_1, x_2, u) + f^{[3]}_{1,\rho}(x_1, x_2, u) \\ &+ c x_{2,\sigma+1} x_{2,\tau+1} x_{2,\upsilon+1} - c\lambda_\rho x_{2,\sigma} x_{2,\tau} x_{2,\upsilon}. \end{aligned}$$

Clearly, if $\lambda_\rho \neq 0$, then ϵ^{1kl}_i (3.14) is unchanged.

7. Conclusion. We have developed a theory of quadratic and cubic normal forms for discrete time control systems. To avoid notational difficulties, we have restricted our attention to scalar input systems whose uncontrollable part is diagonalizable. But the basic operations of pull up and push down extend to more general systems. We have also shown the uniqueness of the normal forms.

We have introduced the concept of control bifurcation and have exhibited some simple examples.

REFERENCES

[1] V. I. Arnold, *Geometrical Methods in the Theory of Ordinary Differential Equations*, Springer-Verlag, New York, 1983.

[2] J.-P. Barbot, S. Monaco, and D. Normand-Cyrot, *Quadratic forms and feedback linearization in discrete time*, Internat. J. Control, 67 (1997), pp. 567–586.

[3] J. Carr, *Applications of the Centre Manifold Theory*, Springer-Verlag, New York, 1981.

[4] D.-E. Chang, W. Kang, and A. J. Krener, *Normal forms and bifurcations of control systems*, in Proceedings of the 39th IEEE Conference on Decision and Control, Sydney, Australia, 2000, pp. 1602–1607.

[5] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.

[6] W. Kang and A. J. Krener, *Extended quadratic controller normal form and dynamic state feedback linearization of nonlinear systems*, SIAM J. Control Optim., 30 (1992), pp. 1319–1337.

[7] W. Kang, *Bifurcation and normal form of nonlinear control systems, Part* I, SIAM J. Control Optim., 36 (1998), pp. 193–212.

[8] W. Kang, *Bifurcation and normal form of nonlinear control systems, Part* II, SIAM J. Control Optim., 36 (1998), pp. 213–232.

[9] Y. A. Kuznetsov, *Elements of Applied Bifurcation Theory*, Springer-Verlag, New York, 1998.

[10] A. J. Krener, W. Kang, and D.-E. Chang, *Control bifurcations*, IEEE Trans. Automat. Control, to appear.

[11] L. Li, *Normal Forms of Controllable and Uncontrollable Discrete Time Nonlinear Systems*, Master's Thesis, Graduate Group in Applied Mathematics, University of California, Davis, CA, 1999.

[12] L. Li and A. J. Krener, *Quadratic and cubic normal forms of discrete time nonlinear control systems*, in Anais do XIII Congresso Brasileiro de Automatica, CBA, Florianopolis, SC Brazil, 2000, pp. 19–25.

[13] S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, Springer-Verlag, New York, 1990.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 6, pp. 1724–1745

GENERALIZED SOLUTIONS IN NONLINEAR STOCHASTIC CONTROL PROBLEMS*

F. DUFOUR† AND BORIS M. MILLER‡

Abstract. An optimal stochastic control problem is considered for systems with unbounded controls satisfying an integral constraint. It is shown that there exists an optimal control within the class of generalized controls leading to impulse actions. Applying an approach of time transformation, developed recently for deterministic systems, the original control problem is shown to be equivalent to an optimal stopping problem. Moreover, the description of generalized solutions is given in terms of stochastic differential equations governed by a measure.

Key words. nonlinear stochastic systems, impulse control, generalized solutions, discontinuous time-change

AMS subject classifications. 49J30, 49N25, 93E20

PII. S0363012900374221

1. Introduction.

In this paper, the existence of an optimal control is discussed for the nonlinear stochastic system defined by the following equation:

$$x_t \doteq \zeta + \int_0^t A(s, x_s)ds + \int_0^t B(s, x_s)u_s ds + \int_0^t D(s, x_s)dW_s, \tag{1}$$

where the functions A, B, and D are deterministic, $\{W_t\}$ is a Brownian motion, and $\{u_t\}$ is the control. All the processes are assumed to be defined on a probability space $(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\})$. Let K be a closed convex cone. The class of admissible controls, labeled $\mathfrak{C}^a$, is defined by the class of K-valued, $\{\mathcal{F}_t\}$-predictable processes subject to the following constraint:

$$\int_0^T |u_s|ds \le M. \tag{2}$$

For an admissible control u, the cost is given by

$$J[u] = E[g(x_T)], \tag{3}$$

where g is a deterministic function and T is the terminal time.

When the control satisfies condition (2), it is easy to see that the optimal solution may not exist within the class of admissible control (see the example in section 3). Indeed, this constraint (2) implies that the admissible control can be chosen as close as desired to a control of impulsive type. An approach to solve this problem in a deterministic context, based on a time transformation, was originally suggested by Warga [21] and has been actively developed recently (see, for example, the survey [15]).

*Received by the editors June 19, 2000; accepted for publication (in revised form) November 1, 2001; published electronically February 14, 2002. This research was supported by a CNRS/Russian Academy of Sciences cooperation (PECO/NEI 9570) and in part by the Nonlinear Control Network and by Russian Basic Investigation Foundation grant 99-01-01-088.

http://www.siam.org/journals/sicon/40-6/37422.html

†LaBRI, Universite Bordeaux I, 351 cours de la Liberation, 33405 Talence Cedex, France (dufour@labri.u-bordeaux.fr).

‡Institute for Information Transmission Problems, 19 Bolshoy Karetny per., Moscow 101447, Russia (bmiller@iitp.ru).

In the stochastic context, this approach was introduced by Miller and Runggaldier in [17] to solve a special case of the problem studied in the present work. In this context, it appears necessary to introduce a new concept to describe the limit of a sequence of control processes subject to the constraint (2); this is the so-called *generalized* control. (For a more precise exposition, see Definition 3.1.) Similarly, the limit of a sequence of solutions of (1) is defined as a *generalized* solution. These definitions of generalized control and generalized solution are taken from the deterministic context (see, for example, [1, 16, 18]).

Our aim is to characterize the value of $\inf_{u\in\mathfrak{C}^a} J[u]$. By introducing the class of admissible generalized controls, labeled $\overline{\mathfrak{C}}^a$, it is shown that $\inf_{u\in\mathfrak{C}^a} J[u] = \inf_{u\in\overline{\mathfrak{C}}^a} J[u]$. The characterization of $\inf_{u\in\mathfrak{C}^a} J[u]$ will be completed when it is shown that the there exists an optimal generalized control $u^* \in \overline{\mathfrak{C}}^a$ such that $\inf_{u\in\overline{\mathfrak{C}}^a} J[u] = J[u^*]$. It proves that there exists an optimal generalized control for the original control problem, justifying, therefore, the introduction of this class of process, $\overline{\mathfrak{C}}^a$. This existence result is obtained by using a time transformation to convert the original control problem into an optimal stopping problem. Moreover, the representation of the generalized solution is given in terms of a stochastic differential equation governed by a measure. This important property enhances the link existing between this control problem and the class of singular control problems.

Singular stochastic control problems have recently received considerable attention in the literature (see [9, 10, 22] and the references therein). However, until now the theoretical basis for this kind of stochastic control problem was restricted to the class of systems where the gain of the singular control does not depend on the state process (see, for example, [9, 10] and the references therein). Therefore, our work can be considered as a first attempt to extend these results in the case where the gain of the singular control may depend on the state process. Other extensions of our approach are already planned, and in [4] it will be shown how this method can be applied to re-examine the singular problem studied in [9]. It must be pointed out that the control problem defined in (1)–(3) cannot be solved directly by using the results in [8, Theorem 4.7]. Our work can be generalized in several directions by adding soft constraints and considering the optimal stopping problem.

The paper is organized as follows. In section 2, we formulate the original control problem. The concept of *generalized* control is introduced in section 3 by analogy with the deterministic case. It is shown that the infimum of the expected cost over the class of admissible controls and the infimum over the class of admissible generalized controls are the same (see Proposition 3.2). Section 4 contains the description of the time transformation and introduces an auxiliary control problem that will be shown to be equivalent to the original one. On the basis of known results [8], the existence theorem is proved for the auxiliary problem. A consequence of this result is derived in section 5 and shows that there exists an optimal generalized control for the original control problem. Its representation is given in terms of a stochastic differential equation governed by a measure. In the appendix, some technical results are derived.

We introduce the following notation and terminology.

Notation. $\mathbb{N}_N$ is the set of the first N integers, that is, $\mathbb{N}_N = \{1, \dots, i, \dots, N\}$. $\mathbb{R}_+ \doteq \{x \in \mathbb{R} : x \geq 0\}$. The ith component of a vector M is denoted by M^i. The symbol $|.|$ is used to denote the norm of vectors and matrices. If X is a normed space, then for $R > 0$ the set $B_R(X)$ is defined by $\{x \in X : |x| < R\}$ and $\overline{B}_R(X) \doteq \{x \in X : |x| \leq R\}$. $(')$ denotes the transpose operation. $0_n \in \mathbb{R}^n$ is the zero vector. The indicator function of a set A is defined as $I_A(x)$. On a probability space $(\Omega, \mathcal{F}, P)$,

the mathematical expectation will be denoted by $E_P[.]$.

In order to define the state processes, let us introduce the following data:

- K is a subset of $\mathbb{R}^p$.
- $A : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}^n$.
- $B : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}^{n\times p}$.
- $D : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}^{n\times m}$.
- $g : \mathbb{R}^n \to \mathbb{R}_+$.
- ζ is a fixed vector in $\mathbb{R}^n$.
- T and M are fixed real numbers.
- $G : \mathbb{R}_+ \to \overline{\mathbb{R}}_+$ such that $G(T) = 0$ and $G(t) = \infty$ for $t \neq T$.

The following assumptions will be used in the paper.

(A.1) There are constants L_1 and L_2 such that for all $t, s \in \mathbb{R}_+$ and $x, y \in \mathbb{R}^n$

$$|A(t,x)| + |B(t,x)| + |D(t,x)| \leq L_1(1 + |x|),$$
$$|A(t,x) - A(s,y)| + |B(t,x) - B(s,y)| + |D(t,x) - D(s,y)| \leq L_2(|x-y| + |t-s|).$$

(A.2) The function g is continuous, and there exist a constant L_3 and a positive integer q such that

$$|g(x)|^2 \leq L_3(1 + |x|^q).$$

(A.3) K is a closed cone which is convex.

(A.4) For all $(t,x) \in [0,T) \times \mathbb{R}^n$, the set $K(t,x)$ defined by

$$K(t,x) \doteq \big\{((1-|\theta|)A(t,x) + B(t,x)\theta, (1-|\theta|)D(t,x)D(t,x)', |\theta|) \ : \ \theta \in \overline{B}_1(K)\big\}$$

is convex.

2. Problem statement. In this section, we formulate the stochastic control problem presented in the introduction using the formulation described in Haussmann and Lepeltier [8] and El Karoui, Nguyen, and Jeanblanc-Picqué [5].

Definition 2.1. *A control is defined by the term*

$$C \doteq (\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}, \{u_t\}, \{W_t\}, \{x_t\}),$$

where the following hold:

(i) $(\Omega, \mathcal{F}, P)$ *is a complete probability space with a right continuous complete filtration* $\{\mathcal{F}_t\}$.

(ii) $\{u_t\}$ *is a* K*-valued,* $\{\mathcal{F}_t\}$*-predictable process such that*

$$\int_0^T |u_s| ds \leq M. \tag{4}$$

(iii) $\{W_t\}$ *is an* $\{\mathcal{F}_t\}$ *standard* m*-dimensional Brownian motion.*

(iv) $\{x_t\}$ *is an* $\mathbb{R}^n$*-valued,* $\{\mathcal{F}_t\}$ *progressively measurable process such that*

$$(\forall t \in [0,T]), \ x_t \doteq \zeta + \int_0^t A(s,x_s)ds + \int_0^t B(s,x_s)u_s ds + \int_0^t D(s,x_s)dW_s. \tag{5}$$

We write $\mathfrak{C}$ *for the set of controls satisfying the previous conditions.*

The cost is given by

$$J[C] \doteq E_P[g(x_T)]. \tag{6}$$

The set $\mathfrak{C}^a$ *of admissible controls is defined by*

$$\mathfrak{C}^a \doteq \{C \in \mathfrak{C} : J[C] < \infty\}. \tag{7}$$

We shall consider as a control objective the minimization of $J[C]$ on $\mathfrak{C}^a$.

As already pointed out in the introduction, since we do not assume any conditions such as the coercivity condition (see (3.5) in [8]), the existence of an optimal control for the previous problem cannot be claimed using the approach described in [8]. Before presenting the concept of generalized control, let us derive the following technical lemma.

LEMMA 2.2. *The stochastic differential equation* (5)*, where* $\{u_t\}$ *satisfies item* (ii) *of Definition* 2.1*, has a unique solution such that*

$$(\forall q \in \mathbb{N}) \qquad E_P\left[\sup_{t\in[0,T]} |x_t|^{2q}\right] < D, \tag{8}$$

where D *is a constant.*

Proof. Using (A.1) and Theorem 7, page 197 in [19], the existence and the uniqueness of the solution are straightforward. The proof of (8) is given in the appendix. We cannot use standard arguments to derive it since the process $\{u_t\}$ may not be bounded but satisfies the inequality (4). □

3. Generalized controls. An optimal control may not exist within the class of ordinary admissible controls $\mathfrak{C}^a$. An example is now presented in order to illustrate this assertion. A deterministic problem is considered where $T = 1$, $M = 1$, the control $u_t \in K \doteq \mathbb{R}_+$, and the state satisfies the following equation:

$$(\forall t \in [0,1]) \quad x_t \doteq \int_0^t (u_s - x_s)ds.$$

The aim is to minimize the cost $J[C] = (1 - x_1)^2$. It is easy to show that $x_1 = \int_0^1 e^{(s-1)}u_s ds$, and, by using the fact that $\int_0^1 u_s ds \leq 1$, it follows that for any admissible control $J[C] > 0$.

Now let us introduce the sequence of admissible controls

$$u_t^n = \begin{cases} 0 & \text{for } 0 \leq t \leq 1 - \frac{1}{n}, \\ n & \text{for } 1 - \frac{1}{n} < t \leq 1. \end{cases}$$

Clearly, $\int_0^1 u_s^n ds = 1$, and the cost $J[C^n]$ associated to u^n is equal to $[1 - n(1 - e^{-\frac{1}{n}})]^2$. Consequently, $\lim_{n\to\infty} J[C^n] = 0$, showing that an optimal control does not exist within the class of ordinary admissible controls $\mathfrak{C}^a$. This is a consequence of the discontinuous behavior of the minimizing sequence $\{u_t^n\}$ at $t = 1$.

In order to characterize $\inf_{C\in\mathfrak{C}^a} J[C]$, we introduce the concept of generalized control, labeled C^g, and its associated class of admissible controls $\overline{\mathfrak{C}}^a$. Moreover, the correspondence between $\inf_{C\in\mathfrak{C}^a} J[C]$ and $\inf_{C^g\in\overline{\mathfrak{C}}^a} J[C^g]$ is given in Proposition 3.2, justifying the introduction of this new class of controls $\overline{\mathfrak{C}}^a$.

DEFINITION 3.1. *A generalized control is defined by the term*

$$C^g \doteq (\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}, \{U_t\}, \{W_t\}, \{X_t\}),$$

where the following hold:

(i) $(\Omega, \mathcal{F}, P)$ *is a complete probability space with a right continuous complete filtration* $\{\mathcal{F}_t\}$.

(ii) $\{U_t\}$ *is a* K*-valued, corlol,* $\{\mathcal{F}_t\}$ *progressively measurable process satisfying*

$$\operatorname*{Var}_{[0,T]}[U_t] \leq M, \qquad U_t - U_s \in K \quad \text{for } t \geq s. \tag{9}$$

(iii) $\{W_t\}$ *is an* $\{\mathcal{F}_t\}$ *standard* m*-dimensional Brownian motion.*

(iv) $\{X_t\}$ *is an* $\mathbb{R}^n$*-valued, corlol,* $\{\mathcal{F}_t\}$ *progressively measurable semimartingale such that the continuous part of* $\{X_t\}$ *satisfies*

$$\begin{aligned}(\forall t \in [0,T]) \quad X_t^c \doteq \zeta &+ \int_0^t A(s, X_s)ds + \int_0^t B(s, X_s)dU_s^c \\ &+ \int_0^t D(s, X_s)dW_s.\end{aligned} \tag{10}$$

(v) *There exists a sequence* $\{C^n\}_{n\in\mathbb{N}}$ *defined by*

$$C^n \doteq (\Omega, \mathcal{F}, P, \{\mathcal{F}_t^n\}, \{u_t^n\}, \{W_t^n\}, \{x_t^n\})$$

such that

$$(\forall n \in \mathbb{N}) \qquad C^n \in \mathfrak{C}^a$$

and

$$\begin{aligned}(\forall t \in [0,T)) \qquad X_t &= \limsup_{\substack{s\to t \\ s>t}} \lim_{n\to\infty} x_s^n, \quad P-a.s., \\ \text{and} \quad X_T &= \lim_{n\to\infty} x_T^n, \quad P-a.s.\end{aligned} \tag{11}$$

We write $\overline{\mathfrak{C}}$ *for the set of controls satisfying the previous conditions.*

The cost is given by

$$J[C^g] \doteq E_P[g(X_T)]. \tag{12}$$

The set $\overline{\mathfrak{C}}^a$ *of admissible controls is defined by*

$$\overline{\mathfrak{C}}^a \doteq \{C^g \in \overline{\mathfrak{C}} : J[C^g] < \infty\}. \tag{13}$$

Note that the discontinuous part of $\{X_t\}$ is generated by the discontinuous part of $\{U_t\}$.

The following result provides a correspondence between the sets of control $\mathfrak{C}^a$ and $\overline{\mathfrak{C}}^a$. Its proof is an immediate consequence of the definitions of $\mathfrak{C}^a$ and $\overline{\mathfrak{C}}^a$ and assumption (A.2).

PROPOSITION 3.2. *The set of control* $\mathfrak{C}^a$ *is a subset of* $\overline{\mathfrak{C}}^a$*, and*

$$\inf_{C\in\mathfrak{C}^a} J[C] = \inf_{C^g\in\overline{\mathfrak{C}}^a} J[C^g]. \tag{14}$$

4. Time transformation and the auxiliary control problem. In this section, we introduce an auxiliary control problem which is given in terms of an optimal stopping problem (see Definition 4.1). It is shown in Corollary 4.16 that this problem is equivalent to the initial one. A key property of the auxiliary control problem is that the controls take their values in a compact set.

DEFINITION 4.1. *An auxiliary control is defined by the term*

$$\Psi \doteq (\Omega, \mathcal{F}, P, \{\mathcal{G}_t\}, \{\theta_t\}, \{V_t\}, \{\Lambda_t\}, \gamma),$$

where the following hold:

(i) $(\Omega, \mathcal{F}, P)$ *is a complete probability space with a right continuous complete filtration* $\{\mathcal{G}_t\}$.

(ii) $\{\theta_t\}$ *is a* $\overline{B}_1(K)$*-valued,* $\{\mathcal{G}_t\}$*-predictable process.*

(iii) $\{V_t\}$ *is a* $\{\mathcal{G}_t\}$ *standard* m*-dimensional Brownian motion.*

(iv) γ *is a* $\{\mathcal{G}_t\}$ *stopping time such that*

$$\gamma \le T + M. \tag{15}$$

(v) $\{\Lambda_t \doteq (\eta_t, \xi_t')'\}$ *is an* $\mathbb{R}^{n+1}$*-valued,* $\{\mathcal{G}_t\}$ *progressively measurable process such that*

$$\eta_t \doteq t - \int_0^t |\theta_s| ds, \tag{16}$$

$$\xi_t \doteq \zeta + \int_0^t (1 - |\theta_s|) A(\eta_s, \xi_s) ds + \int_0^t B(\eta_s, \xi_s)\theta_s ds + \int_0^t \sqrt{1 - |\theta_s|}\, D(\eta_s, \xi_s) dV_s \tag{17}$$

for $t \in [0, \gamma]$.

We write $\overline{\Upsilon}$ *for the set of controls satisfying the previous conditions.*

The cost is given by

$$\mathcal{M}[\Psi] \doteq E_P[g(\xi_\gamma) + G(\eta_\gamma)]. \tag{18}$$

The set $\overline{\Upsilon}^a$ *of admissible auxiliary controls is defined by*

$$\overline{\Upsilon}^a \doteq \{\Psi \in \overline{\Upsilon} : \mathcal{M}[\Psi] < \infty\}. \tag{19}$$

Our aim is to show the equivalence between the auxiliary and the initial control problems. However, we first show the existence of an optimal control for the auxiliary problem.

THEOREM 4.2. *For the auxiliary control problem there exists an optimal control* Θ^**:*

$$\inf_{\Psi \in \overline{\Upsilon}^a} \mathcal{M}[\Psi] = \mathcal{M}[\Theta^*] \qquad \text{and} \qquad \Theta^* \in \overline{\Upsilon}^a. \tag{20}$$

Proof. Applying Corollary 4.8 in [8], it follows that there exist a probability space $(\Omega, \widetilde{\mathcal{F}}, \widetilde{P})$ and a filtration $\{\widetilde{\mathcal{G}}_t\}$ such that

- $\{\widetilde{\theta}_t\}$ is a $\overline{B}_1(K)$-valued, $\{\widetilde{\mathcal{G}}_t\}$ progressively measurable process,
- $\{V_t\}$ is a $\{\widetilde{\mathcal{G}}_t\}$ standard m-dimensional Brownian motion,

and

$$E_{\widetilde{P}}[g(\xi_\gamma)] \leq \inf_{\Psi\in\overline{\Upsilon}^a} \mathcal{M}[\Psi], \tag{21}$$

where γ is a $\{\widetilde{\mathcal{G}}_t\}$ stopping time and

$$\eta_t = t - \int_0^t |\widetilde{\theta}_s| ds,$$
$$\xi_t = \zeta + \int_0^t (1-|\widetilde{\theta}_s|)A(\eta_s,\xi_s)ds + \int_0^t B(\eta_s,\xi_s)\widetilde{\theta}_s ds + \int_0^t \sqrt{1-|\widetilde{\theta}_s|}D(\eta_s,\xi_s)dV_s.$$

In (21), we do not have an equality because in the control problem studied by Haussmann and Lepeltier [8] the set of admissible controls is defined on the set of progressively measurable processes and for an arbitrary probability space. In our case, the admissible controls are defined in the smaller set of predictable processes and on a probability space that must satisfy the usual hypotheses (completion and right continuity). However, using Lemmas A.1 and A.2, it can be shown that there exists a new probability space $(\Omega,\mathcal{F},P,\{\mathcal{G}_t\})$ satisfying the usual hypotheses based on a modification of $(\Omega,\widetilde{\mathcal{F}},\widetilde{P},\{\widetilde{\mathcal{G}}_t\})$. Moreover, the existence of a $\overline{B}_1(K)$-valued, $\{\mathcal{G}_t\}$-predictable process $\{\theta_t\}$ such that

$$\Theta^* \doteq (\Omega,\mathcal{F},P,\{\mathcal{G}_t\},\{\theta_t\},\{V_t\},\{(\eta_t,\xi_t')'\},\gamma) \in \overline{\Upsilon}^a \quad \text{and} \quad \mathcal{M}[\Theta^*] \leq \inf_{\Psi\in\overline{\Upsilon}^a} \mathcal{M}[\Psi]$$

is guaranteed by Lemma A.3.

Consequently, we have $\mathcal{M}[\Theta^*] = \inf_{\Psi\in\overline{\Upsilon}^a} \mathcal{M}[\Psi]$, which gives the result. □

In order to establish the correspondence between the auxiliary control problem and the initial one, we need to introduce the following subset of $\overline{\Upsilon}^a$, labeled Υ^a (see Definition 4.3). We prove in Theorem 4.9 that

$$\begin{aligned} \inf_{\Psi\in\Upsilon^a} \mathcal{M}[\Psi] &= \min_{\Psi\in\overline{\Upsilon}^a} \mathcal{M}[\Psi] \\ &= \mathcal{M}[\Theta^*]. \end{aligned} \tag{22}$$

Then it is shown in Theorem 4.15 that

$$\inf_{C\in\mathfrak{C}^a} J[C] = \inf_{\Psi\in\Upsilon^a} \mathcal{M}[\Psi]. \tag{23}$$

Therefore, combining (22) and (23), the main result of this section (see Corollary 4.16) will follow; that is,

$$\inf_{C\in\mathfrak{C}^a} J[C] = \mathcal{M}[\Theta^*].$$

The rest of this section is devoted to the proofs of relations (22) and (23).

First, in order to show that (22) holds, we prove that for any control $\Psi \in \overline{\Upsilon}^a$ there exists a sequence of controls $\{\Psi^n\}$ in Υ^a such that $\lim_{n\to\infty} \mathcal{M}[\Psi^n] = \mathcal{M}[\Psi]$. This is not a trivial consequence of the closure of Υ^a by $\overline{\Upsilon}^a$ since it is necessary to approximate the stopping time γ. Now we need the following definitions and technical results.

Definition 4.3. *Let us introduce the set* $\Upsilon \subset \overline{\Upsilon}$*:*

$$\Psi \doteq (\Omega,\mathcal{F},P,\{\mathcal{G}_t\},\{\theta_t\},\{V_t\},\{\Lambda_t\},\gamma) \in \Upsilon \Longleftrightarrow \begin{cases} \Psi \in \overline{\Upsilon} \\ \textit{and} \\ \{\theta_t\} \textit{ is a } B_1(K)\textit{-valued process} \end{cases}$$

and the corresponding set of admissible controls

$$\Upsilon^a = \Upsilon \cap \overline{\Upsilon}^a. \tag{24}$$

DEFINITION 4.4. *For* $\Psi \doteq (\Omega, \mathcal{F}, P, \{\mathcal{G}_t\}, \{\theta_t\}, \{V_t\}, \{\Lambda_t = (\eta_t, \xi_t')'\}, \gamma) \in \overline{\Upsilon}^a$, *define on* $(\Omega, \mathcal{F}, P)$

$$\nu^n \doteq \inf\left\{t \geq 0 : t - \int_0^t \frac{n}{n+1}|\theta_s|ds \geq \frac{nT}{n+1}\right\}, \tag{25}$$

$$\nu \doteq \inf\{t \geq 0 : \eta_t \geq T\}, \tag{26}$$

$$\alpha^n \doteq \frac{T + M - \nu^n - \frac{T}{n+1}}{T + M - \nu^n}. \tag{27}$$

LEMMA 4.5. *If* $\Psi \doteq (\Omega, \mathcal{F}, P, \{\mathcal{G}_t\}, \{\theta_t\}, \{V_t\}, \{\Lambda_t = (\eta_t, \xi_t')'\}, \gamma)$ *is an element of* $\overline{\Upsilon}^a$, *then* ν *and* ν^n *are* $\{\mathcal{G}_t\}$ *stopping times (for all* $n \in \mathbb{N}$*) and*

$$\lim_{n\to\infty} \nu^n = \nu, \quad P- \textit{a.s.}, \tag{28}$$

and

$$0 \leq \alpha^n < 1. \tag{29}$$

Proof. See the appendix. □

Using Lemma 4.5, we can now show that a sequence of control $\{\Psi^n\}$ in Υ^a can be constructed from any element Ψ in $\overline{\Upsilon}^a$ as described below.

PROPOSITION 4.6. *Assume that* $\Psi \doteq (\Omega, \mathcal{F}, P, \{\mathcal{G}_t\}, \{\theta_t\}, \{V_t\}, \{\Lambda_t = (\eta_t, \xi_t')'\}, \gamma)$ *is an element of* $\overline{\Upsilon}^a$. *Define the sequence* $\{\Psi^n\}_{n\in\mathbb{N}}$ *by*

$$\Psi^n \doteq \big(\Omega, \mathcal{F}, P, \{\mathcal{G}_t\}, \{\theta_t^n\}, \{V_t\}, \{\Lambda_t^n = (\eta_t^n, \xi_t^{n\prime})'\}, \gamma^n\big), \tag{30}$$

where

$$\theta_t^n \doteq \frac{n}{n+1}\theta_t I_{[\![0,\nu^n]\!]} + \alpha^n e_1 I_{]\!]\nu^n,\nu]\!]} + \alpha^n \theta_t I_{]\!]\nu,\gamma]\!]} \quad (\, e_1 \doteq (1, 0, \ldots, 0)' \in \mathbb{R}^p\,), \tag{31}$$

$$\eta_t^n \doteq t - \int_0^t |\theta_s^n|ds, \tag{32}$$

$$\begin{aligned}\xi_t^n \doteq \zeta &+ \int_0^t (1 - |\theta_s^n|)A(\eta_s^n, \xi_s^n)I_{\{s\leq\gamma^n\}}ds + \int_0^t B(\eta_s^n, \xi_s^n)\theta_s^n I_{\{s\leq\gamma^n\}}ds \\ &+ \int_0^t \sqrt{1 - |\theta_s^n|}D(\eta_s^n, \xi_s^n)I_{\{s\leq\gamma^n\}}dV_s,\end{aligned} \tag{33}$$

$$\gamma^n \doteq \inf\{t \geq 0 : \eta_t^n > T\}. \tag{34}$$

Then $\Psi^n \in \Upsilon^a$ *for all* $n \in \mathbb{N}$.

Proof. From Lemma 4.5 and assumption (A.2), it follows that for all $n \in \mathbb{N}$, $\{\theta_t^n\}$ is a $B_1(K)$-valued process. Moreover, using the fact that α^n is measurable with respect to $\mathcal{G}_{\nu^n}$, $\mathcal{G}_{\nu^n} \subset \mathcal{G}_\nu$, and Corollary 6.34 in [6], it follows easily that for all $n \in \mathbb{N}$, the process $\{\theta_t^n\}$ is $\{\mathcal{G}_t\}$-predictable. From the definitions of $\{\eta_t^n\}$ and γ^n, we obtain that $\eta_{T+M}^n \geq \eta_{\gamma^n}^n$. Therefore, we have that

$$\gamma^n \leq T + M \tag{35}$$

because $\{\eta_t^n\}$ is a strictly increasing process.

Now, applying Theorem 7, page 197 in [19], it is easy to see that (33) has a unique solution. Therefore, for all $n \in \mathbb{N}$ the control Ψ^n satisfies all of the conditions of Definition 4.3.

Clearly, we have $E_P[G(\eta^n_{\gamma^n})] = 0$ and

$$E_P\left[\sup_{s\leq T+M}|\xi_s^n|^p\right] \leq C \tag{36}$$

for a constant C depending on p but independent of n.

Combining hypothesis (A.2), the previous inequalities, and (35), we obtain that

$$\mathcal{M}[\Psi^n] = E_P[g(\xi^n_{\gamma^n})] \quad < \quad \infty, \tag{37}$$

and so $\Psi^n \in \Upsilon^a$ for all $n \in \mathbb{N}$. □

In order to derive the convergence of $\mathcal{M}[\Psi^n]$ to the cost function $\mathcal{M}[\Psi]$, we need the following technical lemma.

LEMMA 4.7. *Assume that* $\Psi \doteq (\Omega, \mathcal{F}, P, \{\mathcal{G}_t\}, \{\theta_t\}, \{V_t\}, \{\Lambda_t = (\eta_t, \xi_t')'\}, \gamma)$ *is an element of* $\overline{\Upsilon}^a$. *Then*

$$0 \leq \gamma^n - \gamma \leq \frac{T}{n+1} \tag{38}$$

and

$$E_P\left[\left|\int_0^\gamma |\theta_s - \theta_s^n|^2 ds\right|^2 + \int_0^\gamma |\eta_s - \eta_s^n|^2 ds\right] \leq C\left\{\frac{1}{(n+1)^2} + E_P[|\nu - \nu^n|^2]\right\} \tag{39}$$

for a constant C *independent of* n.

Proof. See the appendix. □

Finally, based on the previous lemma, we can prove that the sequence $\{\Psi^n\}$ satisfies the desired property.

PROPOSITION 4.8. *Assume that* $\Psi \doteq (\Omega, \mathcal{F}, P, \{\mathcal{G}_t\}, \{\theta_t\}, \{V_t\}, \{\Lambda_t = (\eta_t, \xi_t')'\}, \gamma)$ *is an element of* $\overline{\Upsilon}^a$. *Then the sequence* $\{\Psi^n\}$ *in* Υ^a *satisfies*

$$\lim_{n\to\infty} \mathcal{M}[\Psi^n] = \mathcal{M}[\Psi]. \tag{40}$$

Proof. Let us introduce the following equation:

$$\begin{aligned}\chi_t = \zeta &+ \int_0^t (1 - |\theta_s|)A(\eta_s, \chi_s)I_{\{s\leq\gamma\}}ds + \int_0^t B(\eta_s, \chi_s)\theta_s I_{\{s\leq\gamma\}}ds \\ &+ \int_0^t \sqrt{1 - |\theta_s|}D(\eta_s, \chi_s)I_{\{s\leq\gamma\}}dV_s.\end{aligned} \tag{41}$$

Applying Theorem 7, page 197 in [19], it is easy to see that this equation has a unique solution.

By using Doob's inequality and Gronwall's lemma, it is easy to show that there exists a constant $\overline{C}$ such that

$$(\forall t \in [0, T+M]) \qquad E_P\left[\sup_{s\leq t}|\chi_s - \xi_s^n|^2\right] \leq \overline{C}\left[\frac{1}{n+1} + \sqrt{E_P[|\nu^n - \nu|^2]}\right]. \tag{42}$$

Moreover, we clearly have $\chi_{T+M} = \xi_\gamma$ and $\xi^n_{T+M} = \xi^n_{\gamma^n}$, and so

$$E_P[|\xi_\gamma - \xi^n_{\gamma^n}|^2] \le \overline{C}\ \left[\frac{1}{n+1} + \sqrt{E_P[|\nu^n - \nu|^2]}\right]. \tag{43}$$

The sequence ν^n is bounded by $T + M$, and so it is uniformly integrable. Therefore, using Lemma 4.5, we have that $\lim_{n\to\infty} E_P[|\nu^n - \nu|^2]$. With (43), we obtain that $g(\xi^n_{\gamma^n}) \longrightarrow^P_{n\to\infty} g(\xi_\gamma)$ since the function g is continuous. Clearly, the sequence $\{g(\xi^n_{\gamma^n})\}$ is uniformly integrable, and so

$$\lim_{n\to\infty} E_P[g(\xi^n_{\gamma^n})] = E_P[g(\xi_\gamma)], \tag{44}$$

giving the result. □

In conclusion, we obtain the following result.

THEOREM 4.9. *Let $\Theta^* \in \overline{\Upsilon}^a$ be the optimal control for the auxiliary control problem. Then*

$$\inf_{\Psi\in\Upsilon^a} \mathcal{M}[\Psi] = \mathcal{M}[\Theta^*]. \tag{45}$$

Proof. The existence of Θ^* has been shown in Theorem 4.2. Since $\Upsilon^a \subset \overline{\Upsilon}^a$, we clearly have

$$\inf_{\Psi\in\Upsilon^a} \mathcal{M}[\Psi] \ge \mathcal{M}[\Theta^*].$$

However, using Proposition 4.8, the result follows. □

Now let us show that (23) holds. Its proof is given in Theorem 4.15 and is based on Propositions 4.12 and 4.14. Here we use two time transformations which establish the correspondence between $\mathfrak{C}^a$ and Υ^a.

Let us introduce the following time-change.

LEMMA 4.10. *Let $(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}, \{u_t\}, \{W_t\}, \{z_t = (y_t, x_t')'\})$ be an element of $\mathfrak{C}$, and let $\{\Gamma_t\}$ be the process defined by*

$$\Gamma_t \doteq t + \int_0^t |u_{s\wedge T}| ds. \tag{46}$$

Denote by $\{\Phi_t\}$ the right inverse of Γ:

$$\Phi_t \doteq \inf\{s \ge 0 : \Gamma_s > t\}. \tag{47}$$

Then $\{\Phi_t\}$ is a continuous time-change satisfying the following properties:

(i) $(\forall\ t \in \mathbb{R}_+)\quad \Phi_{\Gamma_t} = t \quad$ *and* $\quad \Gamma_{\Phi_t} = t.$

(ii) $(\forall\ t \in \mathbb{R}_+)\quad \Phi_t = \int_0^t \frac{1}{1+|u_{\Phi_s\wedge T}|} ds.$

Proof. Item (i) is obvious. Differentiating the second equality in (i) and using (46), item (ii) follows easily. □

REMARK 4.11. *An immediate consequence of the previous lemma is the following assertion:*

$$(\forall t \in [0, \Gamma_T]) \qquad \Phi_t = \int_0^t \frac{1}{1 + |u_{\Phi_s}|} ds, \tag{48}$$

which will be used repeatedly in what follows.

The following proposition shows that for any control $C \in \mathfrak{C}^a$, there exists a control $\Theta \in \Upsilon^a$ having the same cost.

PROPOSITION 4.12. *Assume that* $C \doteq (\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}, \{u_t\}, \{W_t\}, \{x_t\})$ *is an element of* $\mathfrak{C}^a$. *Write* Θ *for* $(\Omega, \mathcal{F}, P, \{\mathcal{F}_{\Phi_t}\}, \{\theta_t\}, \{V_t\}, \{\Delta_t\}, \Gamma_T)$, *where*

$$\theta_t \doteq \frac{u_{\Phi_t}}{1+|u_{\Phi_t}|}, \quad V_t \doteq \int_0^{\Phi_t} \sqrt{1+|u_s|}dW_s, \quad \Delta_t \doteq \begin{pmatrix} \Phi_t \\ x_{\Phi_t} \end{pmatrix}, \tag{49}$$

and Φ_t *(respectively,* Γ_t*) is defined by* (47) *(respectively,* (46)*). Then* Θ *belongs to* Υ^a *and*

$$\mathcal{M}[\Theta] = J[C]. \tag{50}$$

Proof. From Proposition 1.1 in [20, Chapter V], $\{\mathcal{F}_{\Phi_t}\}$ defines an increasing and right continuous filtration which is complete. Then assertion (i) of Definition 4.1 is satisfied. Now, using Theorem 3.52 in [11], it follows that $\{\frac{u_{\Phi_t}}{1+|u_{\Phi_t}|}\}$ is an $\{\mathcal{F}_{\Phi_t}\}$-predictable process. Moreover, for all $t \in \mathbb{R}_+$, $\frac{u_{\Phi_t}}{1+|u_{\Phi_t}|} \in B_1(K)$.

The process $\{N_t \doteq \int_0^t \sqrt{1+|u_s|}dW_s\}$ is an $\{\mathcal{F}_t\}$ continuous local martingale such that

$$(\forall t \in [0,T],\ \forall (i,j) \in \mathbb{N}_n^2) \quad \langle N^i, N^i\rangle_t = \Gamma_t,\ \langle N^i, N^j\rangle_t = 0 \quad (i \neq j). \tag{51}$$

Therefore, according to Theorem 4.13 in [12], $\{V_t = N_{\Phi_t}\}$ is an $\{\mathcal{F}_{\Phi_t}\}$ standard m-dimensional Brownian motion which gives item (iii) of Definition 4.1.

By Remark 2.9 and Theorem 2.33 in [6], the process $\{\Delta_t\}$ is adapted to $\{\mathcal{F}_{\Phi_t}\}$. Clearly, the process $\{\Delta_t\}$ is *corlol*. Consequently, $\{\Delta_t\}$ is progressively measurable with respect to $\{\mathcal{F}_{\Phi_t}\}$.

Using Proposition 1.1 in [20, Chapter V], Γ_T is an $\{\mathcal{F}_{\Phi_t}\}$ stopping time. With (4), we have that $\Gamma_T = T + \int_0^T |u_s|ds \leq T + M$. Therefore, item (iv) of Definition 4.1 is satisfied.

Now let us show that the components of $\{\Delta_t\}$ satisfy (16) and (17) on $[0, \Gamma_T]$. Using (48) and the definition of $\{\theta\}$, we have

$$(\forall t \in [0,\Gamma_T]) \qquad \Phi_t = t - \int_0^t |\theta_s|ds. \tag{52}$$

Therefore, the first component of the process $\{\Delta_t\}$ satisfies (16).

Now the process $\{x_{\Phi_t}\}$ satisfies (for all $t \in [0,\Gamma_T]$)

$$x_{\Phi_t} \doteq \zeta + \int_0^{\Phi_t} A(s,x_s)ds + \int_0^{\Phi_t} B(s,x_s)u_s ds + \int_0^{\Phi_t} D(s,x_s)dW_s. \tag{53}$$

Since $\{\Gamma_t\}$ is continuous, we can use Proposition 1.4 in [20, Chapter V] and Lemma 4.10 in order to obtain that

$$\begin{aligned} (\forall t \in [0,\Gamma_T]) \qquad \int_0^{\Phi_t} A(s,x_s)ds &= \int_0^t A(\Phi_s, x_{\Phi_s})d\Phi_s \\ &= \int_0^t A(\Phi_s, x_{\Phi_s})(1-|\theta_s|)ds. \end{aligned} \tag{54}$$

We can repeat the same argument to show that

$$(\forall t \in [0,\Gamma_T]) \qquad \int_0^{\Phi_t} B(s,x_s)u_s ds = \int_0^t B(\Phi_s, x_{\Phi_s})\theta_s ds. \tag{55}$$

Moreover,

$$(\forall t \in [0,\Gamma_T]) \qquad \int_0^{\Phi_t} D(s,x_s)dW_s = \int_0^{\Phi_t} D(s,x_s)\frac{1}{\sqrt{1+|u_s|}}dN_s$$

$$= \int_0^t D(\Phi_s, x_{\Phi_s})\sqrt{1-|\theta_s|}dV_s, \tag{56}$$

where the last equality is obtained by using Proposition 4.8 in [12].

Combining (53)–(56), we obtain that the process $\{x_{\Phi_t}\}$ satisfies

$$(\forall t \in [0,\Gamma_T]) \qquad x_{\Phi_t} = \zeta + \int_0^t (1-|\theta_s|)A(\Phi_s, x_{\Phi_s})ds + \int_0^t B(\Phi_s, x_{\Phi_s})\theta_s ds$$

$$+ \int_0^t \sqrt{1-|\theta_s|}D(\Phi_s, x_{\Phi_s})dV_s. \tag{57}$$

Therefore, assertion (iv) of Definition 4.1 is satisfied for the process $\{\Delta_t\}$ (see (49) for its definition). Finally, it follows that $\Theta \in \overline{\Upsilon}$. However, we have shown that $\{\theta_t\}$ is a $B_1(K)$-valued process. Consequently, $\Theta \in \Upsilon$.

Now the cost corresponding to Θ is given by $\mathcal{M}(\Theta) = E_P[g(x_{\Phi_{\Gamma_T}}) + G(\Phi_{\Gamma_T})]$.

However, $\Phi_{\Gamma_T} = T$ (see item (i) of Lemma 4.10). Therefore, we have $\mathcal{M}[\Theta] = E_P[g(x_T)] = J[C] < \infty$, implying that $\Theta \in \Upsilon^a$. □

The proof of the following lemma is similar to that of Lemma 4.10. Therefore, it is omitted.

LEMMA 4.13. *Let $(\Omega, \mathcal{F}, P, \{\mathcal{G}_t\}, \{\theta_t\}, \{V_t\}, \{\Lambda_t = (\eta_t, \xi_t')'\}, \gamma)$ be an element of Υ, and let $\{\psi_t\}$ be the right inverse of η:*

$$\psi_t \doteq \inf\{s \geq 0 : \eta_s > t\}. \tag{58}$$

The process $\{\psi_t\}$ is a continuous time-change satisfying the following properties:

(i) $(\forall\ t \in \mathbb{R}_+)$ $\psi_{\eta_t} = t$ *and* $\eta_{\psi_t} = t$.

(ii) $(\forall\ t \in \mathbb{R}_+)$ $\psi_t = \int_0^t \frac{1}{1-|\theta_{\psi_s}|}ds$.

Conversely to Proposition 4.12, we show that, for any control $\Psi \in \Upsilon^a$, there exists a control $S \in \mathfrak{C}^a$ having the same cost. The ideas to show this result are the same as the one used in the proof of Proposition 4.12. Consequently, this result is quoted without proof.

PROPOSITION 4.14. *Let $\Psi \doteq (\Omega, \mathcal{F}, P, \{\mathcal{G}_t\}, \{\theta_t\}, \{V_t\}, \{\Lambda_t = (\eta_t, \xi_t')'\}, \gamma)$ be an element of Υ^a. Write S for $(\Omega, \mathcal{F}, P, \{\mathcal{G}_{\psi_t}\}, \{u_t\}, \{W_t\}, \{\xi_{\psi_t}\})$, where*

$$u_t \doteq \frac{\theta_{\psi_t}}{1-|\theta_{\psi_t}|}, \qquad W_t \doteq \int_0^{\psi_t} \sqrt{1-|\theta_s|}dV_s, \tag{59}$$

and $\{\Psi_t\}$ is defined in (58).

Then S belongs to $\mathfrak{C}^a$, and

$$J[S] = \mathcal{M}[\Psi]. \tag{60}$$

Now we obtain the following result.

THEOREM 4.15. *The following property holds:*

$$\inf_{C \in \mathfrak{C}^a} J[C] = \inf_{\Psi \in \Upsilon^a} \mathcal{M}[\Psi]. \tag{61}$$

Proof. The result is an immediate consequence of Propositions 4.12 and 4.14. ∎

Finally, we derive an important characterization of $\inf_{C\in\mathfrak{C}^a} J[C]$.

COROLLARY 4.16. *Let $\Theta^* \in \overline{\Upsilon}^a$ be the optimal control for the auxiliary control problem. Then*

$$\inf_{C\in\mathfrak{C}^a} J[C] = \mathcal{M}[\Theta^*]. \tag{62}$$

Proof. It is a straightforward combination of Theorems 4.9 and 4.15. ∎

REMARK 4.17. *Let us denote by $\Theta^* \doteq (\Omega, \mathcal{F}, P, \{\mathcal{G}_t\}, \{\theta_t\}, \{V_t\}, \{(\eta_t, \xi_t')'\}, \gamma)$ the optimal control in $\overline{\Upsilon}^a$. There is no loss of generality to assume that*

$$\inf\{s : \eta_s > T\} = \gamma.$$

Indeed, if this is not the case, let $\widetilde{\Theta}$ be the control defined by

$$\begin{aligned}
\widetilde{\Theta} &\doteq (\Omega, \mathcal{F}, P, \{\mathcal{G}_t\}, \{\widetilde{\theta}_t\}, \{V_t\}, \{(\widetilde{\eta}_t, \widetilde{\xi}_t')'\}, \gamma),\\
\widetilde{\theta}_t &\doteq \theta_t I_{[0,\gamma]},\\
\widetilde{\eta}_t &\doteq t - \int_0^t |\widetilde{\theta}_s| ds,\\
\widetilde{\xi}_t &\doteq \zeta + \int_0^t (1 - |\widetilde{\theta}_s|) A(\widetilde{\eta}_s, \widetilde{\xi}_s) ds + \int_0^t B(\widetilde{\eta}_s, \widetilde{\xi}_s)\widetilde{\theta}_s ds\\
&\quad + \int_0^t \sqrt{1 - |\widetilde{\theta}_s|} D(\widetilde{\eta}_s, \widetilde{\xi}_s) dV_s.
\end{aligned}$$

Clearly, $\widetilde{\Theta} \in \Upsilon^a$ and $\inf\{s : \widetilde{\eta}_s > T\} = \gamma$. Moreover, it is easy to check that (for all $t \in [0, \gamma]$) $\widetilde{\theta}_t = \theta_t$, $\widetilde{\eta}_t = \eta_t$, and $\widetilde{\xi}_t = \xi_t$. Therefore,

$$\mathcal{M}(\widetilde{\Theta}) = \mathcal{M}(\Theta^*) = \min_{\Psi\in\overline{\Upsilon}^a} \mathcal{M}[\Psi].$$

5. Existence of an optimal generalized control. In this section, we obtain the last characterization of $\inf_{C\in\mathfrak{C}^a} J[C]$ in terms of an optimal generalized control.

THEOREM 5.1. *There exists a generalized control $C^{g*} \in \overline{\mathfrak{C}}^a$ such that*

$$\begin{aligned}
\inf_{C\in\mathfrak{C}^a} J[C] &= J[C^{g*}]\\
&= \min_{C^g\in\overline{\mathfrak{C}}^a} J[C^g].
\end{aligned}$$

Proof. Let us denote by

$$\Theta^* \doteq (\Omega, \mathcal{F}, P, \{\mathcal{G}_t\}, \{\theta_t\}, \{V_t\}, \{(\eta_t, \xi_t')'\}, \gamma)$$

the optimal control in $\overline{\Upsilon}^a$.

Define

$$\psi_t \doteq \inf\{s : \eta_s > t\}, \tag{63}$$

$$X_t \doteq \xi_{\psi_t}, \tag{64}$$

$$U_t \doteq \int_0^{\psi_t} \theta_s ds, \tag{65}$$

$$W_t \doteq \int_0^{\psi_t} \sqrt{1 - |\theta_s|} dV_s. \tag{66}$$

On the probability space $(\Omega, \mathcal{F}, P)$, $\{\psi_t\}$ is a time-change (see Proposition 1.1 in [20, Chapter V]). Moreover, $\{\mathcal{G}_{\psi_t}\}$ defines a right continuous complete filtration. Therefore, the processes $\{X_t\}$, $\{Y_t\}$, and $\{U_t\}$ are $\{\mathcal{G}_{\psi_t}\}$ progressively measurable (see Theorem T57, page 105 in [14]). Since $\{\xi_t\}$ is a continuous process, $\{X_t\}$, $\{Y_t\}$, and $\{U_t\}$ are *corlol*. Moreover, since K is a separable metric space satisfying assumption (A.3), it is easy to obtain that $\{U_t\}$ is a K-valued process and $U_t - U_s \in K$ for $t \geq s$.

According to Theorem 4.13 in [12], $\{W_t\}$ is a $\{\mathcal{G}_{\psi_t}\}$ standard m-dimensional Brownian motion.

Now, using Theorem 6.46 in [6], there exists a sequence $\{\tau_n\}$ of stopping times which exhausts the jumps of $\{\psi_t\}$. Clearly, we have

$$\overset{\infty}{\underset{n=1}{\cup}} [\![\psi_{\tau_n -}, \psi_{\tau_n}]\!] \subset \{(t,\omega) \in \mathbb{R}_+ \times \Omega : |\theta_t| = 1\}.$$

Define

$$\mathcal{D} \doteq \{(t,\omega) \in \mathbb{R}_+ \times \Omega : |\theta_t| = 1\} - \overset{\infty}{\underset{n=1}{\cup}} [\![\psi_{\tau_n -}, \psi_{\tau_n}]\!].$$

Consequently,

$$\begin{aligned}(\forall t \in [0,T]) \quad U_t &= \int_0^{\psi_t} I_{\{|\theta_s|<1\}} \theta_s ds + \int_0^{\psi_t} I_{\{|\theta_s|=1\}} \theta_s ds \\ &= \int_0^{\psi_t} [I_{\{|\theta_s|<1\}} + I_{\mathcal{D}}] \theta_s ds + \sum_{n \in \mathbb{N}} \int_{\psi_{\tau_n -}}^{\psi_{\tau_n}} \theta_s ds \, I_{[\![\tau_n, \infty[\![}.\end{aligned}$$

For $(t,\omega) \in \cup_{n=1}^{\infty} [\![\psi_{\tau_n -}, \psi_{\tau_n}]\!]$, we have $I_{\{|\theta_t|<1\}}(\omega) + I_{\mathcal{D}}(t,\omega) = 0$.

Therefore, $\{\int_0^t [I_{\{|\theta_s|<1\}} + I_{\mathcal{D}}] \theta_s ds\}$ is a $\{\psi_t\}$ continuous process. Consequently, the decomposition of the process $\{U_t\}$ is given by

$$\begin{aligned} U_t^c &= \int_0^{\psi_t} [I_{\{|\theta_s|<1\}} + I_{\mathcal{D}}] \theta_s ds, \\ U_t^d &= \sum_{n \in \mathbb{N}} \int_{\psi_{\tau_n -}}^{\psi_{\tau_n}} \theta_s ds \, I_{[\![\tau_n, \infty[\![}. \end{aligned}$$

From Lemma 1.37 in [11], we have

$$(\forall t \in \mathbb{R}_+) \qquad \eta_{\psi_t} = t. \tag{67}$$

Moreover, using Proposition 4.8 in [12], it follows that

$$(\forall t \in [0,T]) \quad \int_0^{\psi_t} \sqrt{1 - |\theta_s|} D(\eta_s, \xi_s) dV_s = \int_0^t D(s, \xi_{\psi_s}) dW_s.$$

Note that $\{\eta_t\}$ is a $\{\psi_t\}$ continuous process. Moreover, $\{\eta_t\}$ is a process of finite variation because it is absolutely continuous. Therefore, using Proposition 1.4 in [20, Chapter V] and (67), we obtain that

$$\begin{aligned}(\forall t \in [0,T)) \quad \int_0^{\psi_t} (1 - |\theta_s|) A(\eta_s, \xi_s) ds &= \int_0^{\psi_t} A(\eta_s, \xi_s) d\eta_s \\ &= \int_0^t A(s, \xi_{\psi_s}) ds.\end{aligned}$$

Again, using the fact that $\{\int_0^t [I_{\{|\theta_s|<1\}} + I_{\mathcal{D}}]\theta_s ds\}$ is a $\{\psi_t\}$ continuous process and Proposition 1.4 in [20, Chapter V], we have

$$\begin{aligned}\int_0^{\psi_t} B(\eta_s,\xi_s)\theta_s ds &= \int_0^{\psi_t} B(\eta_s,\xi_s)\left[I_{\{|\theta_s|<1\}} + I_{\mathcal{D}}\right]\theta_s ds \\ &\quad + \sum_{n\in\mathbb{N}} \int_{\psi_{\tau_n-}}^{\psi_{\tau_n}} B(\eta_s,\xi_s)\theta_s ds\, I_{[\![\tau_n,\infty[\![} \\ &= \int_0^t B(s,\xi_{\psi_s}) dU_s^c + \sum_{n\in\mathbb{N}} \int_{\psi_{\tau_n-}}^{\psi_{\tau_n}} B(\eta_s,\xi_s)\theta_s ds\, I_{[\![\tau_n,\infty[\![}.\end{aligned}$$

It follows that the process $\{X_t\}$ satisfies the following equation:

$$\begin{aligned}(\forall t\in[0,T]) \quad X_t &= \zeta + \int_0^t A(s,X_s)ds + \int_0^t B(s,X_s)dU_s^c + \int_0^t D(s,X_s)dW_s \\ &\quad + \sum_{n\in\mathbb{N}} \Delta X_{\tau_n}\ I_{[\![\tau_n,\infty[\![},\end{aligned} \tag{68}$$

where

$$\Delta X_{\tau_n} \doteq \int_{\psi_{\tau_n-}}^{\psi_{\tau_n}} B(\eta_s,\xi_s)\theta_s ds. \tag{69}$$

According to Proposition 4.8, there exists a sequence $\{\Psi^n\}$ such that $\Psi^n \in \Upsilon^a$ for all $n\in\mathbb{N}$. Write

$$\psi_t^n \doteq \inf\{s : \eta_s^n > t\}.$$

From Proposition 4.6, it follows that $\{\psi_t^n\}$ is a continuous, strictly increasing process and such that

$$\left(\forall t\in\left[0,\frac{nT}{n+1}\right)\right) \qquad \psi_t^n \le \psi_t^{n+1} \le \psi_t. \tag{70}$$

Therefore, for t in $[0,T))$, $\overline{\psi}_t \doteq \lim_{n\to\infty}\psi_t^n$ exists. Again, using (70), this limit is lower semicontinuous and increasing on $[0,T)$.

Using similar arguments as in the proof of Lemma 4.5, it can be shown easily that

$$(\forall t\in[0,T)) \qquad \eta_{\overline{\psi}_t} = t. \tag{71}$$

Combining (67) with (71), we obtain that

$$(\forall t\in\mathbb{R}_+) \qquad \sum_n I_{]\tau_n\wedge T,\tau_{n+1}\wedge T[\![}\psi_t = \sum_n I_{]\tau_n\wedge T,\tau_{n+1}\wedge T[\![}\overline{\psi}_t.$$

However, recalling that $\{\overline{\psi}_t\}$ is a lower semicontinuous, increasing process and $\{\psi_t\}$ is *corlol*, it follows that $\{\overline{\psi}_t\}$ is *collor* and

$$(\forall t\in[0,T)) \qquad \psi_t = \overline{\psi}_{t+}. \tag{72}$$

There is no loss of generality to assume that $\lim_{t\to 0 t<0}\psi_t = 0$, and so

$$(\forall t\in[0,T)) \qquad \overline{\psi}_t = \psi_{t-}. \tag{73}$$

Moreover, since $\psi^n \in \Upsilon^a$, we have that $\eta^n_{\gamma^n} = T$. Recalling that $\{\eta^n_t\}$ is strictly increasing and continuous, we obtain that $\psi^n_T = \gamma^n$. Note that $\psi_T = \gamma$ (see Remark 4.17). Using Lemma 4.7, it follows that

$$\overline{\psi}_T \doteq \lim_{n\to\infty} \psi^n_T = \psi_T = \gamma, \quad P-\text{a.s.} \tag{74}$$

Using (73) and the fact that $\{\psi_t\}$ is an increasing process, we have that, for all $t \in [0,T)$, $\overline{\psi}_t \le \psi_{T-} \le \psi_T = \overline{\psi}_T$. Consequently, $\{\overline{\psi}_t\}$ is increasing.

However, using similar arguments as in the proof of Proposition 4.8 (see (42)), it can be shown easily that there exists a subsequence, still denoted by n, such that (for all $t \in [0,T]$)

$$\lim_{n\to\infty} \xi^n_{\psi^n_t} = \xi_{\overline{\psi}_t}, \quad P-\text{a.s., and } \lim_{n\to\infty} \int_0^{\psi^n_t} |\theta^n_s| ds = \int_0^{\overline{\psi}_t} |\theta_s| ds, \quad P-\text{a.s.} \tag{75}$$

Define

$$C^n = \left(\Omega, \mathcal{F}, P, \{\mathcal{G}_{\psi^n_t}\}, \left\{\frac{\theta^n_{\psi^n_t}}{1-|\theta^n_{\psi^n_t}|}\right\}, \left\{\int_0^{\psi^n_t} \sqrt{1-|\theta^n_s|} dV_s\right\}, \{(\psi^n_t - t, \xi^{n\prime}_{\psi^n_t})'\}\right).$$

Since $\Psi^n \in \Upsilon^a$ for all $n \in \mathbb{N}$ and using Proposition 4.14, it follows that $C^n \in \mathfrak{C}^a$ for all $n \in \mathbb{N}$.

Using the fact that $\{\xi_t\}$ is continuous, (72), and (75), we obtain that

$$(\forall t \in [0,T)) \qquad X_t = \lim_{\substack{s\to t \\ s>t}} \lim_{n\to\infty} \xi^n_{\psi^n_s}, \quad P-\text{a.s.}$$

Moreover, using (74) and (75),

$$X_T = \xi_{\psi_T} = \xi_{\overline{\psi}_T} = \lim_{n\to\infty} \xi^n_{\psi^n_T}, \quad P-\text{a.s.} \tag{76}$$

From the definition of $\{\psi^n_t\}$ and since $\Psi^n \in \Upsilon^a$, we can use Lemma 4.13 in order to obtain

$$\begin{aligned}(\forall t \in [0,T]) \qquad \psi^n_t - t &= \int_0^t \frac{|\theta^n_{\psi^n_s}|}{1-|\theta^n_{\psi^n_s}|} ds \\ &= \int_0^t |\theta^n_{\psi^n_s}| d\psi^n_s.\end{aligned}$$

Using Proposition 1.4 in [20, Chapter V] and the fact that $\{\psi^n_t\}$ is a continuous process, we have

$$(\forall t \in [0,T]) \qquad \psi^n_t - t = \int_0^{\psi^n_t} |\theta^n_s| ds. \tag{77}$$

Therefore, combining (74), (75), and (77) yields

$$\int_0^{\psi_T} |\theta_s| ds = \gamma - T.$$

However,

$$\operatorname*{Var}_{[0,T]}[U_t] \le \operatorname*{Var}_{[0,T]}\left[\int_0^{\psi_t} |\theta_s| ds\right] = \int_0^{\psi_T} |\theta_s| ds.$$

Consequently, using (15), it follows that

$$\operatorname*{Var}_{[0,T]}[U_t] \leq M.$$

Finally, the generalized control C^{g*} defined by

$$C^{g*} \doteq (\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}, \{U_t\}, \{W_t\}, \{X_t\})$$

is an element of $\overline{\mathfrak{C}}^a$.

However, by hypothesis, $\inf_{C\in\mathfrak{C}^a} J[C] = E_P[g(\xi_\gamma)]$, and so we obtain with (76)

$$\inf_{C\in\mathfrak{C}^a} J[C] = E_P[g(X_T)] = J[C^{g*}].$$

Now, using Proposition 3.2, we obtain the result. □

Appendix. In this section, we prove some technical results.

Proof of Lemma 2.2. Let us consider $R > |\zeta|$ and define $\tau_R \doteq \inf\{t : |x_t| \geq R\}$. Clearly, the process $\{x_{t\wedge\tau_R}\}$ is solution of the following equation:

$$\begin{aligned} x_{t\wedge\tau_R} \doteq \zeta &+ \int_0^t A(s, x_{s\wedge\tau_R}) I_{\{s\leq\tau_R\}} ds + \int_0^t B(s, x_{s\wedge\tau_R}) u_s I_{\{s\leq\tau_R\}} ds \\ &+ \int_0^t D(s, x_{s\wedge\tau_R}) I_{\{s\leq\tau_R\}} dW_s. \end{aligned}$$

Using (A.1) and (4), it follows that

$$\begin{aligned} |x_{t\wedge\tau_R}| \leq |\zeta| + L_1(T+M) &+ \left| \int_0^t D(s, x_{s\wedge\tau_R}) I_{\{s\leq\tau_R\}} dW_s \right| \\ &+ \int_0^t |x_{s\wedge\tau_R}| L_1(1+|u_t|) ds. \end{aligned}$$

Using Gronwall's lemma, we obtain that

$$|x_{t\wedge\tau_R}| \leq M_1 + M_2 \sup_{s\leq t} \left| \int_0^s D(s, x_{s\wedge\tau_R}) I_{\{s\leq\tau_R\}} dW_s \right|,$$

where M_1 and M_2 are two constants. Using Theorem 6.5, page 87 in [7], assumption (A.1), and Gronwall's lemma, we finally have that

$$E_P\left[\sup_{t\leq T} |x_{t\wedge\tau_R}|^{2q}\right] \leq M$$

for a constant M.

Due to the continuity of $\{x_t\}$, $\tau_R \to \infty$ as $R \to \infty$. Therefore, using Fatou's lemma and the previous equation, the result follows. □

LEMMA A.1. *Suppose $(\Omega, \mathcal{F}, P)$ is a probability space with a filtration $\{\mathcal{G}_t\}$ and $\{V_t\}$ is a $\{\mathcal{G}_t\}$ standard Brownian motion. Then $\{V_t\}$ is a $\{\mathcal{G}_t^q\}$ standard Brownian motion on the probability space $\left(\Omega, \mathcal{F}^q, \overline{P}\right)$, where*

$$\begin{aligned} \mathcal{F}^q &\doteq \{A \subset \Omega : (\exists B \in \mathcal{F}) \text{ such that } A \triangle B \in \mathcal{N}\}, \\ \mathcal{G}_t^q &\doteq \{A \subset \Omega : (\exists B \in \mathcal{G}_t) \text{ such that } A \triangle B \in \mathcal{N}\}, \\ \mathcal{N} &\doteq \{A \subset \Omega : (\exists B \in \mathcal{F}) \text{ such that } A \subset B \text{ and } P(B) = 0\}, \end{aligned}$$

and the probability $\overline{P}$ *is defined by (for all* $A \in \mathcal{F}^q$*)* $\overline{P}(A) = P(B)$*, where* $B \in \mathcal{F}$ *and* $A \triangle B \in \mathcal{N}$.

Proof. From the definition of $\mathcal{G}_s^q$ and $\mathcal{N}$, it follows that

$$(\forall s > 0, \quad \forall A \in \mathcal{G}_s^q) \quad (\exists (B, N) \in \mathcal{G}_s \times \mathcal{N}) \qquad \text{such that} \qquad A = B + N.$$

Therefore, for all $t > s \geq 0$ and for all $A \in \mathcal{G}_s^q$ we have

$$\int_A \exp[iu'(V_t - V_s)]\, d\overline{P} = \exp -\frac{|u|^2(t-s)}{2}\, P(B). \tag{78}$$

Moreover, $P(B) = \overline{P}(A)$, and, using (78), we obtain

$$E_{\overline{P}}[\exp[iu'(V_t - V_s)]\, |\mathcal{G}_s^q] = \exp -\frac{|u|^2(t-s)}{2},$$

which gives the result. □

LEMMA A.2. *Suppose* $(\Omega, \mathcal{F}, P)$ *is a complete probability space with a complete filtration* $\{\mathcal{G}_t\}$ *and* $\{V_t\}$ *is a* $\{\mathcal{G}_t\}$ *standard Brownian motion. Then* $\{V_t\}$ *is a* $\{\mathcal{G}_{t+}\}$ *standard Brownian motion.*

Proof. For all $t > s \geq 0$ and $0 < \epsilon < t - s$ we have

$$E_P[\exp[iu'(V_t - V_{s+\epsilon})]\, |\mathcal{G}_{s+\epsilon}] = \exp -\frac{|u|^2(t-s-\epsilon)}{2}.$$

Since $\mathcal{G}_{s+} = \bigcap_{\epsilon>0} \mathcal{G}_{s+\epsilon}$, we obtain that

$$(\forall A \in \mathcal{G}_{s+}) \qquad \lim_{\substack{\epsilon \to 0 \\ \epsilon > 0}} \int_A \exp[iu'(V_t - V_{s+\epsilon})]\, dP = \exp -\frac{|u|^2(t-s)}{2}\, P(A).$$

By using the bounded convergence theorem and the fact that the $\{V_t\}$ is a continuous process, we have

$$E_P[\exp[iu'(V_t - V_s)]\, |\mathcal{G}_{s+}] = \exp -\frac{|u|^2(t-s)}{2},$$

and the result follows. □

LEMMA A.3. *Assume that* $\{\theta_t\}$ *is a* $\overline{B}_1(K)$*-valued,* $\{\mathcal{G}_t\}$ *progressively measurable process. Then* (16) *and* (17) *have a unique solution such that*

$$(\forall q \in \mathbb{N}) \qquad E_P\left[\sup_{t \in [0,\gamma]} |\xi_t|^{2q}\right] < \infty. \tag{79}$$

Moreover, there exist a $\overline{B}_1(K)$*-valued,* $\{\mathcal{G}_t\}$*-predictable process* $\{\overline{\theta}_t\}$ *such that*

$$\overline{\theta} = \theta, \quad \lambda \otimes P - a.e., \tag{80}$$

and the process $\{\overline{\xi}_t\}$ *solution of the following stochastic differential equation:*

$$\overline{\xi}_t \doteq \zeta + \int_0^t (1 - |\overline{\theta}_s|) A(\overline{\eta}_s, \overline{\xi}_s) ds + \int_0^t B(\overline{\eta}_s, \overline{\xi}_s) \overline{\theta}_s ds + \int_0^t \sqrt{1 - |\overline{\theta}_s|}\, D(\overline{\eta}_s, \overline{\xi}_s) dV_s,$$

where $\overline{\eta}_t = t - \int_0^t |\overline{\theta}_s| ds$ *is indistinguishable from* $\{\xi_t\}$.

Proof. Using (A.1) and Theorem 7, page 197 in [19], the existence and the uniqueness of the solution are straightforward. The conditions of Corollary 10, page 85 in [13] are satisfied, and the inequality (79) follows.

By hypothesis, the process θ is progressively measurable with respect to $\{\mathcal{G}_t\}$. Using Theorem 3.7 in [3], it follows that the function $\theta : \mathbb{R}_+ \times \Omega \to K$ is $\mathcal{P}^*$ measurable, where

$$\mathcal{P}^* \doteq \{A \in \mathcal{B}(\mathbb{R}_+) \otimes \mathcal{F} : A \triangle B \in \mathcal{N} \text{ for some } B \in \mathcal{P}\},$$

$\mathcal{P}$ denoting the predictable σ-field and $\mathcal{N} \doteq \{N \in \mathcal{B}(\mathbb{R}_+) \otimes \mathcal{F} : \lambda \otimes P(N) = 0\}$. Since $\overline{B}_1(K)$ is a locally compact separable metric space, we can use the lemma and its associated remark [2, pp. 59–60] to obtain the existence of a $\overline{B}_1(K)$-valued, $\{\mathcal{G}_t\}$-predictable process $\{\overline{\theta}_t\}$ satisfying (80).

Consequently,

$$\eta_t = t - \int_0^t |\overline{\theta}_s| ds. \tag{81}$$

Moreover, since $\{\overline{\eta}_t\}$ and $\{\eta_t\}$ are continuous, they are indistinguishable processes. Combining (81), (8), and (A.1), we obtain that $\int_{[0,T+M]\times\Omega} |B(.,\overline{x}_.)u_.| \, \lambda \otimes P < \infty$, so

$$(\forall t \in [0,\gamma]) \quad \int_0^t B(\overline{\eta}_s, \overline{\xi}_s)\overline{\theta}_s ds = \int_0^t B(\eta_s, \overline{\xi}_s)\theta_s ds$$

by using Fubini's theorem.

Similarly, we have that

$$(\forall t \in [0,\gamma]) \quad \int_0^t (1 - |\overline{\theta}_s|) A(\overline{\eta}_s, \overline{\xi}_s) ds = \int_0^t (1 - |\theta_s|) A(\eta_s, \overline{\xi}_s) ds$$

and

$$(\forall t \in [0,\gamma]) \quad \int_0^t (1 - |\overline{\theta}_s|) |D(\overline{\eta}_s, \overline{\xi}_s)|^2 ds = \int_0^t (1 - |\theta_s|) |D(\eta_s, \overline{\xi}_s)|^2 ds.$$

Consequently, we obtain that

$$(\forall t \in [0,\gamma]) \quad \int_0^t \sqrt{1 - |\overline{\theta}_s|} D(\overline{\eta}_s, \overline{\xi}_s) ds = \int_0^t \sqrt{1 - |\theta_s|} D(\eta_s, \overline{\xi}_s) ds,$$

which implies that $\{\overline{\xi}_t\}$ satisfies (17).

By the uniqueness of the solution of (17), $\xi_t = \overline{\xi}_t$, $P-$ a.s., for all t in $[0,\gamma]$. However, $\{\xi_t\}$ and $\{\overline{\xi}_t\}$ are continuous processes, so they are indistinguishable. ∎

Proof of Lemma 4.5. Clearly, ν and ν^n are $\{\mathcal{G}_t\}$ stopping times (for all $n \in \mathbb{N}$). Since $\Psi \in \overline{\Upsilon}^a$, we have that $E_P[G(\eta_\gamma)] < \infty$, implying that $\eta_\gamma = T$. With (15) and the definition of ν, we obtain that

$$\nu \leq \gamma \leq T + M. \tag{82}$$

Note that

$$\nu^n \leq \inf\left\{t \geq 0 : t - \int_0^t \frac{n+1}{n+2} |\theta_s| ds \geq \frac{nT}{n+1}\right\}.$$

Since the process $\{t - \int_0^t \frac{n}{n+1}|\theta_s|ds\}$ is continuous and strictly increasing, we have $\nu^n < \nu^{n+1}$. Similarly, it can be shown that

$$\nu^n < \nu. \tag{83}$$

Therefore, the sequence $\{\nu^n\}$ converges almost surely to a limit labeled $\overline{\nu}$ such that

$$\overline{\nu} \leq \nu. \tag{84}$$

By definition of ν^n, we have

$$\nu^n - \int_0^{\nu^n} \frac{n}{n+1}|\theta_s|ds = \frac{nT}{n+1}, \tag{85}$$

and letting $n \to \infty$, we obtain

$$\eta_{\overline{\nu}} = T.$$

From the definition of ν, we have $\overline{\nu} \geq \nu$. However, with (84) we obtain that $\overline{\nu} = \nu$, and so $\lim_{n\to\infty} \nu^n = \nu$.

From the definition of ν and (85), we obtain that

$$\nu - \nu^n - \frac{T}{n+1} \geq 0.$$

With (84), we have $T + M - \nu^n - \frac{T}{n+1} \geq 0$. Moreover, using (82) and (83), we obtain that $T + M - \nu^n > 0$. Finally, $0 \leq \alpha^n < 1$, which gives the result. □

Proof of Lemma 4.7. Since the process $\{\eta_t^n\}$ is strictly increasing, it follows that γ^n is the unique solution of the following equation:

$$\eta_{\gamma^n}^n = T. \tag{86}$$

However,

$$\begin{aligned} \eta_\gamma^n &= \eta_{\nu^n}^n + \gamma - \nu^n - \int_{\nu^n}^{\gamma} |\theta_s^n|ds \\ &= \frac{nT}{n+1} + \frac{T(\gamma - \nu^n)}{(n+1)(T + M - \nu^n)}. \end{aligned} \tag{87}$$

Using (15), it follows that

$$\eta_\gamma^n \leq T. \tag{88}$$

From (86), we have

$$\eta_\gamma^n + \gamma^n - \gamma = T. \tag{89}$$

Combining (87)–(89), we obtain (38).

By using the definition of $\{\theta_t^n\}$, (15), and the fact that $|\theta_t| \leq 1$, we have

$$\begin{aligned} \int_0^\gamma |\theta_s - \theta_s^n|^2 ds &\leq 2\left[\int_0^{\nu^n} |\theta_s - \theta_s^n|ds + \int_{\nu^n}^{\nu} |\theta_s - \theta_s^n|ds + \int_\nu^\gamma |\theta_s - \theta_s^n|ds\right] \\ &\leq 2\left[\frac{T+M}{n+1} + \nu - \nu^n + \frac{T}{n+1}\frac{\gamma - \nu}{T + M - \nu^n}\right]. \end{aligned}$$

From (15), (82), and (83), it follows that

$$\frac{\gamma-\nu}{T+M-\nu^n} \leq \frac{T+M-\nu}{T+M-\nu^n} \leq 1.$$

Consequently, there exists a constant C_1 such that

$$E_P\left[\left|\int_0^\gamma |\theta_s-\theta_s^n|^2 ds\right|^2\right] \leq C_1\left\{\frac{1}{(n+1)^2}+E_P[|\nu-\nu^n|^2]\right\}.$$

Similarly, it is easy to deduce the same bound for $E_P[\int_0^\gamma |\eta_s-\eta_s^n|^2 ds]$, which gives the result. □

Acknowledgments. The authors are grateful to anonymous referees for many suggestions which have greatly improved the presentation of the paper.

The authors would like to thank P. Bertrand (L2S-CNRS) and F. Lamnhabi (NCN, L2S-CNRS) for their support.

REFERENCES

[1] A. Bressan and F. Rampazzo, *Impulsive control systems with commutative vector fields*, J. Optim. Theory Appl., 71 (1991), pp. 67–83.
[2] K. L. Chung, *Lectures from Markov Processes to Brownian Motion*, Springer-Verlag, New York, 1980.
[3] K. L. Chung and R. J. Williams, *Introduction to Stochastic Integration*, Birkhäuser Boston, Boston, 1990.
[4] F. Dufour and B. Miller, *On the correspondence between singular control and generalized control*, in Proceedings of the 5th IFAC Symposium on Nonlinear Control Systems, St. Petersburg, Russia, 2001, pp. 1174–1178.
[5] N. El Karoui, H. Nguyen, and M. Jeanblanc-Picqué, *Compactification methods in the control of degenerate diffusions: Existence of an optimal control*, Stochastics Stochastics Rep., 20 (1987), pp. 169–219.
[6] R. J. Elliott, *Stochastic Calculus and Applications*, Springer-Verlag, New York, 1982.
[7] A. Friedman, *Stochastic Differential Equations and Applications*, Vol. 1, Academic Press, New York, 1975.
[8] U. G. Haussmann and J.-P. Lepeltier, *On the existence of optimal controls*, SIAM J. Control Optim., 28 (1990), pp. 851–902.
[9] U. G. Haussmann and W. Suo, *Singular optimal stochastic controls* I*: Existence*, SIAM J. Control Optim., 33 (1995), pp. 916–936.
[10] U. G. Haussmann and W. Suo, *Singular optimal stochastic controls* II*: Dynamic programming*, SIAM J. Control Optim., 33 (1995), pp. 937–959.
[11] S. He, J. Wang, and J. Yan, *Semimartingale Theory and Stochastic Calculus*, Science Press, New York, 1992.
[12] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, 2nd ed., Springer-Verlag, New York, 1991.
[13] N. V. Krylov, *Controlled Diffusion Processes*, Springer-Verlag, New York, 1980.
[14] P. A. Meyer, *Probabilités et potentiel*, Hermann, Paris, 1966.
[15] B. M. Miller, *Method of discontinuous time change in problems of control for impulse and discrete-continuous systems*, Automat. Remote Control, 54 (1993), pp. 1727–1750.
[16] B. M. Miller, *The generalized solutions of nonlinear optimization problems with impulse control*, SIAM J. Control Optim., 34 (1996), pp. 1420–1440.
[17] B. M. Miller and W. J. Runggaldier, *Optimization of observations: A stochastic control approach*, SIAM J. Control Optim., 35 (1997), pp. 1030–1052.
[18] Y. V. Orlov, *Theory of Optimal Systems with Generalized Controls*, Nauka, Moscow, 1988 (in Russian).
[19] P. Protter, *Stochastic Integration and Differential Equations*, Springer-Verlag, New York, 1990.

[20] D. Revuz and M. Yor, *Continuous Martingales and Brownian Motion*, 3rd ed., Springer-Verlag, New York, 1999.
[21] J. Warga, *Variational problems with unbounded controls*, J. Soc. Indust. Appl. Math. Ser. A Control, 3 (1965), pp. 424–438.
[22] H. Zhu, *Generalized solution in singular control: The nondegenerate problem*, Appl. Math. Optim., 25 (1992), pp. 225–245.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 6, pp. 1746–1764

SYSTEMS OF CONTROLLED FUNCTIONAL DIFFERENTIAL EQUATIONS AND ADAPTIVE TRACKING*

A. ILCHMANN†, E. P. RYAN‡, AND C. J. SANGWIN§

Abstract. An adaptive servomechanism is developed in the context of the problem of approximate or practical tracking (with prescribed asymptotic accuracy), by the system output, of any admissible reference signal (absolutely continuous and bounded with essentially bounded derivative) for every member of a class of controlled dynamical systems modelled by functional differential equations.

Key words. adaptive control, nonlinear systems, functional differential equations, practical tracking, universal servomechanism

AMS subject classifications. 93C23, 93C10, 93C40, 34K20

PII. S0363012900379704

1. Introduction.

A servomechanism problem is addressed in the context of a class of controlled dynamical systems having the interconnected structure shown in the dashed box in Figure 1. In particular, the aim is the development of an adaptive servomechanism which, for every system of the underlying class, ensures practical tracking (in the sense that prespecified asymptotic tracking accuracy, quantified by $\lambda > 0$, is assured), by the system output, of an arbitrary reference signal assumed to be locally absolutely continuous and bounded with essentially bounded derivative. (We denote by $\mathcal{R}$ the class of such functions and remark that bounded globally Lipschitz functions form an easily recognized subclass.) The system consists of the interconnection of two blocks: The dynamic block Σ_1, which can be influenced directly by the system input/control u (an $\mathbb{R}^M$-valued function), is also driven by the output w from the dynamic block Σ_2. Viewed abstractly, the block Σ_2 can be considered as a causal operator which maps the system output y (an $\mathbb{R}^M$-valued function) to w (an internal quantity, unavailable for feedback purposes).

In essence, the underlying system class $\mathcal{S}$ consists of infinite-dimensional nonlinear M-input u, M-output y systems (p, f, g, T), given by a controlled nonlinear functional differential equation of the form

(1.1)
$$\dot{y}(t) = f(p(t), (Ty)(t)) + g(p(t), (Ty)(t), u(t)), \quad y|_{[-h,0]} = y^0 \in C([-h, 0]; \mathbb{R}^M),$$

where, loosely speaking, $h \geq 0$ quantifies the "memory" of the system, p may be thought of as a (bounded) disturbance term, and T is a nonlinear causal operator. While a full description of the system class $\mathcal{S}$ is postponed to section 3, we remark here that diverse phenomena are incorporated within the class including, for example, diffusion processes, delays (both point and distributed), and hysteretic effects.

*Received by the editors October 16, 2000; accepted for publication (in revised form) June 16, 2001; published electronically February 14, 2002. This research was based on work supported in part by the UK Engineering and Physical Sciences Research Council (grant ref: GR/L78086).
http://www.siam.org/journals/sicon/40-6/37970.html

†Institute of Mathematics, Technical University Ilmenau, Weimarer Straße 25, 98693 Ilmenau, Germany (ilchmann@mathematik.tu-ilmenau.de).

‡Department of Mathematical Sciences, University of Bath, Claverton Down, Bath BA2 7AY, UK (epr@maths.bath.ac.uk).

§School of Mathematics and Statistics, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK (sangwinc@for.mat.bham.ac.uk).

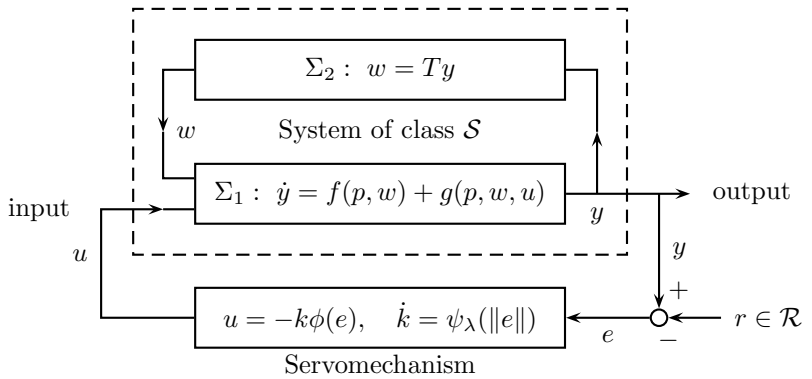


FIG. 1. $(\mathcal{R},\mathcal{S})$-*universal* λ-*servomechanism.*

Furthermore, we remark that results pertaining to adaptive control of functional differential equations are also contained in [3], wherein both the underlying class of systems and the analytic framework differs in an essential manner from those of the present paper; restricted to a problem of adaptive *stabilization*, related results are also reported in [19], with the fundamental distinction that, in [19], *discontinuous* stabilizing feedback strategies are developed within an analytic framework of differential inclusions.

The control objective is to determine an $(\mathcal{R},\mathcal{S})$-universal λ-servomechanism: specifically, to determine continuous functions $\phi : \mathbb{R}^M \to \mathbb{R}^M$ and $\psi_\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ (parameterized by $\lambda > 0$) such that, for each system of class $\mathcal{S}$ and every reference signal $r \in \mathcal{R}$, the control

$$u(t) = -k(t)\phi(y(t) - r(t)), \quad \dot{k}(t) = \psi_\lambda(\|y(t) - r(t)\|), \quad k|_{[-h,0]} = k^0 \tag{1.2}$$

applied to (1.1) ensures (i) convergence of the controller gain, and (ii) tracking of $r(\cdot)$ with asymptotic accuracy quantified by $\lambda > 0$, in the sense that $\max\{\|y(t) - r(t)\| - \lambda, 0\} \to 0$ as $t \to \infty$. See Figure 1.

Given $\lambda > 0$, $r \in \mathcal{R}$ and writing

$$F : (t,w,y,k) \mapsto (f(p(t),w) + g(p(t),w,-k\phi(y - r(t))),\ \psi_\lambda(\|y - r(t)\|)), \tag{1.3}$$

we see that analysis of the behavior of a system $(p,f,g,T) \in \mathcal{S}$ under control (1.2) constitutes a study of an initial-value problem of the form

$$\dot{x}(t) = F(t,(\widehat{T}x)(t)), \quad x|_{[-h,0]} = x^0 := (y^0,k^0) \in C([-h,0];\mathbb{R}^N), \tag{1.4}$$

where $N = M+1$, $x(t) = (y(t),k(t))$, and $\widehat{T}$ is an operator defined on $C([-h,\infty);\mathbb{R}^N)$ by

$$(\widehat{T}x)(t) = (\widehat{T}(y,k))(t) := ((Ty)(t),y(t),k(t)). \tag{1.5}$$

The contribution of this paper is threefold in theme: First, we provide an existence theory for initial-value problems of the general form (1.4) under relatively mild hypotheses on F and $\widehat{T}$; second, and within the framework of the first theme,

we develop a universal servomechanism[1] for a class of nonlinear, infinite-dimensional systems; third, we elucidate the hypotheses on the right-hand side ψ_λ of the gain adaptation equation in (1.2) under which the tracking objective is achievable. In the very specific context of the linear systems of section 2.2 below we will show that $\psi_\lambda : [0,\infty) \to [0,\infty)$ may be chosen as any continuous function with the properties $\psi_\lambda^{-1}(0) = [0,\lambda]$ and $\liminf_{s\to\infty} \psi_\lambda(s) \neq 0$. (In particular, ψ_λ may be chosen to be a bounded function; one such choice is given by $\psi_\lambda(s) = \max\{s-\lambda, 0\}/s$ for $s > 0$ with $\psi_\lambda(0) := 0$.) This ensures that the gain k can exhibit at most linear growth, a feature with attendant practical advantages.

We close this section with some remarks on notation. For $I \subset \mathbb{R}$ an interval $C(I;\mathbb{R}^N)$ (respectively, $AC_{\mathrm{loc}}(I;\mathbb{R}^N)$) denotes the set of continuous (respectively, locally absolutely continuous) functions $I \to \mathbb{R}^N$; $L^\infty_{\mathrm{loc}}(I;\mathbb{R}^N)$ denotes the space of measurable locally essentially bounded functions $I \to \mathbb{R}^N$. For $x : I \to \mathbb{R}^N$, the restriction of x to $J \subset I$ is denoted by $x|_J$. The open ball of radius $r > 0$, centered at $c \in \mathbb{R}^N$, is written as $\mathbb{B}_r(c)$. For $\lambda > 0$, d_λ denotes the Euclidean distance function for $[-\lambda,\lambda]$ given by

$$d_\lambda(\xi) := \max\{0, |\xi| - \lambda\}. \tag{1.6}$$

$\mathcal{R}$ denotes the space of bounded functions in $AC_{\mathrm{loc}}(\mathbb{R};\mathbb{R}^M)$ with essentially bounded derivative; when equipped with the norm $\|\cdot\|_{1,\infty}$ given by $\|r\|_{1,\infty} = \sup_{t\in\mathbb{R}} \|r(t)\| + \text{ess-sup}_{t\in\mathbb{R}} \|\dot{r}(t)\|$, $\mathcal{R}$ can be identified as the Sobolev space $W^{1,\infty}(\mathbb{R};\mathbb{R}^M)$. We write $\mathbb{R}_+ := [0,\infty)$. $\mathcal{K}$ denotes the class of continuous, strictly increasing functions $\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ with $\alpha(0) = 0$; the subclass of *unbounded* class $\mathcal{K}$ functions is denoted $\mathcal{K}_\infty$. $\mathcal{KL}$ is the class of functions $\gamma : \mathbb{R}_+^2 \to \mathbb{R}_+$ such that for each $t \in \mathbb{R}_+$, $\gamma(\cdot,t)$ is of class $\mathcal{K}$ and for each $s \in \mathbb{R}_+$, $\gamma(s,\cdot)$ is decreasing with $\gamma(s,t) \to 0$ as $t \to \infty$.

2. Functional differential equations. The focus of this section is the development of an existence theory, for initial-value problems of the form (1.4), of sufficient generality to accommodate the analysis of dynamic behavior of the adaptively controlled systems of later sections. While the literature is rich in existence results for functional differential equations (see, for example, [4]), we are unaware of a result directly applicable to the particular class of equations which form the focus of the present paper. For this reason, and to make the present paper self-contained, we provide an appropriate result in Theorem 2.3 below (with proof in the appendix). First, we make precise the class of admissible operators $\widehat{T}$ in (1.4).

DEFINITION 2.1 (the operator class $\mathcal{T}_h^{N,K}$). *For $h \geq 0$ and $N, K \in \mathbb{N}$, let $\mathcal{T}_h^{N,K}$ denote the space of operators $T : C([-h,\infty);\mathbb{R}^N) \to L^\infty_{\mathrm{loc}}(\mathbb{R}_+;\mathbb{R}^K)$ with the following properties.*

1. *For every $\delta > 0$ and every bounded interval $I \subset \mathbb{R}_+$, there exists $\Delta > 0$ such that, for all $x \in C([-h,\infty);\mathbb{R}^N)$,*

$$\sup_{t\in[-h,\infty)} \|x(t)\| < \delta \quad \Longrightarrow \quad \|(Tx)(t)\| < \Delta \quad \textit{for almost all (a.a.) } t \in I\,.$$

2. *For all $t \in \mathbb{R}_+$, the following hold:*
   (a) *for all $x, \xi \in C([-h,\infty);\mathbb{R}^N)$,*

$$x(\cdot) \equiv \xi(\cdot) \text{ on } [-h,t] \quad \Longrightarrow \quad (Tx)(s) = (T\xi)(s) \ \textit{ for a.a. } s \in [0,t];$$

[1] The servomechanism can also tolerate disturbances on the output measurement in a sense to be described in section 3.

(b) *for all continuous* $\zeta : [-h, t] \to \mathbb{R}^N$*, there exist* $\tau, \delta, c > 0$ *such that, for all* $x, \xi \in C([-h, \infty); \mathbb{R}^N)$ *with* $x|_{[-h,t]} = \zeta = \xi|_{[-h,t]}$ *and* $x(s), \xi(s) \in \mathbb{B}_\delta(\zeta(t))$ *for all* $s \in [t, t + \tau]$*,*

$$\operatorname*{ess\,sup}_{s\in[t,t+\tau]} \|(Tx)(s) - (T\xi)(s)\| \;\le\; c \sup_{s\in[t,t+\tau]} \|x(s) - \xi(s)\|.$$

Remark 2.2. (i) The essence of property 1 of Definition 2.1 is a "bounded-input, *locally* bounded-output" assumption.

(ii) Property 2(a) is an assumption of causality.

(iii) Property 2(b) is a technical assumption on T of a "locally Lipschitz" nature.

(iv) Let $T \in \mathcal{T}_h^{N,K}$ and $t \ge 0$. Given $x \in C([-h, t); \mathbb{R}^N)$, let x^e denote an arbitrary extension of x to $C([-h, \infty); \mathbb{R}^N)$. By virtue of property 2(a), $Tx^e|_{[0,t)}$ is uniquely determined by the function x, in the sense that the former is independent of the extension x^e chosen for the latter. Expanding on this observation, we will adopt the following notational convention: For $s \in [0, t)$, we simply write $(Tx)(s)$ in place of $(Tx^e)(s)$, where $x^e \in C([-h, \infty); \mathbb{R}^N)$ is any continuous extension of x.

(v) For $\omega \in \mathbb{R}$, let S_ω denote the shift operator on functions $\mathbb{R} \to \mathbb{R}^M$ given by $(S_\omega x)(t) := x(t + \omega)$ for all $t \in \mathbb{R}$. Then

$$T \in \mathcal{T}_h^{N,K} \quad \Longrightarrow \quad TS_{-\omega} \in \mathcal{T}_{h+\omega}^{N,K} \quad \text{for all } \omega \ge 0. \tag{2.1}$$

(vi) Let $T_1, T_2 \in \mathcal{T}_h^{N,K}$ and $\tau_1, \tau_2 \in \mathbb{R}$. Then the operator $\tau_1 T_1 + \tau_2 T_2$, defined by $(\tau_1 T_1 + \tau_2 T_2)(y)(t) := \tau_1 (T_1 y)(t) + \tau_2 (T_2 y)(t)$, is also of class $\mathcal{T}_h^{N,K}$.

(vii) The class $\mathcal{T}_h^{N,N}$ differs from class $\mathcal{T}_h^N$ of [19, Definition 4] only insofar as operators of the former class have range $C([-h, \infty); \mathbb{R}^N)$ while operators of the latter class have domain $L^\infty_{\mathrm{loc}}(\mathbb{R}; \mathbb{R}^N)$.

2.1. An existence theorem. Consider the initial-value problem

$$\dot{x}(t) = F(t, (\widehat{T}x)(t)), \quad x|_{[-h,0]} = x^0 \in C([-h, 0]; \mathbb{R}^N), \tag{2.2}$$

where $\widehat{T}$ is a causal operator of class $\mathcal{T}_h^{N,K}$ and $F : [-h, \infty) \times \mathbb{R}^K \to \mathbb{R}^N$ is a Carathéodory function. (Specifically, (i) for almost all $t \in \mathbb{R}$, $F(t, \cdot)$ is continuous; (ii) for each fixed $w \in \mathbb{R}^K$, $F(\cdot, w)$ is measurable; (iii) for each compact $C \subset \mathbb{R}^K$ there exists $\kappa \in L^1_{\mathrm{loc}}([-h, \infty); \mathbb{R}_+)$ such that

$$\|F(t, w)\| \le \kappa(t) \quad \text{for almost all } t \in [-h, \infty) \text{ and all } w \in C.)$$

By a solution of (2.2) on $[-h, \omega)$, we mean a function $x \in C([-h, \omega); \mathbb{R}^N)$, with $\omega \in (0, \infty]$ and $x|_{[-h,0]} = x^0$, such that $x|_{[0,\omega)}$ is absolutely continuous and satisfies the differential equation in (2.2) for almost all $t \in [0, \omega)$; x is maximal if it has no right extension that is also a solution.

THEOREM 2.3. *Let* $N, K \in \mathbb{N}$*,* $\widehat{T} \in \mathcal{T}_h^{N,K}$*, and* $x^0 \in C([-h, 0]; \mathbb{R}^N)$*. Assume* $F : [-h, \infty) \times \mathbb{R}^K \to \mathbb{R}^N$ *is a Carathéodory function.*

There exists a solution $x : [-h, \omega) \to \mathbb{R}^N$ *of the initial-value problem* (2.2)*, and every solution can be extended to a maximal solution; moreover, if* $F \in L^\infty_{\mathrm{loc}}([-h, \infty) \times \mathbb{R}^K; \mathbb{R}^N)$ *and* $x : [-h, \omega) \to \mathbb{R}^N$ *is a bounded maximal solution, then* $\omega = \infty$*.*

Proof. For proof, see the appendix.

Next, we show that the operators of the class $\mathcal{T}_h^{N,K}$ encompass the input-output behavior of a diverse range of subsystems Σ_2 (see Figure 1).

2.2. Linear systems. *The finite-dimensional prototype.* Consider the well-studied class $\mathcal{L}$ of finite-dimensional, real, linear, minimum-phase, M-input ($u(t)$), M-output ($y(t)$) systems having high-frequency gain $B \in \mathbb{R}^{M\times M}$ with spectrum in the open right half complex plane. Under a suitable coordinate transformation (see, for example, [5, Proposition 2.1.2]), every system in $\mathcal{L}$ can be expressed in the form of two coupled subsystems

$$\left.\begin{aligned} \dot{y}(t) &= A_1 y(t) + A_2 z(t) + Bu(t), \quad y(0) = y^0 \\ \dot{z}(t) &= A_3 y(t) + A_4 z(t), \qquad\qquad z(0) = z^0 \end{aligned}\right\} \tag{2.3}$$

with $y(t), u(t) \in \mathbb{R}^M$, $z(t) \in \mathbb{R}^{N-M}$, and where A_4 has spectrum in the open left half complex plane. Introducing the linear operator T given by

$$(Ty)(t) := A_1 y(t) + A_2 \int_0^t \exp(A_4(t-s))A_3 y(s)ds \tag{2.4}$$

and the function p given by $p(t) := A_2 \exp(A_4 t)z^0$, then, with respect to an operator theoretic viewpoint, system (2.3) can be interpreted as

$$\dot{y}(t) = p(t) + (Ty)(t) + Bu(t), \quad y(0) = y^0. \tag{2.5}$$

With reference to Figure 1, (2.4) and (2.5) correspond to components Σ_2 and Σ_1 of the interconnected system.

Regular linear systems with bounded observation operator. The following example is adapted from [19] and extends the prototype linear class $\mathcal{L}$ to an infinite-dimensional setting by replacing the second of the differential equations (2.3) by an infinite-dimensional analogue on a Hilbert space X. Let $\mathbf{G}$ denote the transfer function of a regular (in the sense of [22]) linear system with state space X, with generating operators (A, B, C, D), and with $\mathbb{R}^M$-valued input and $\mathbb{R}^Q$-valued output. This means, in particular, that (i) A generates a strongly continuous semigroup $\mathbf{S} = (\mathbf{S}_t)_{t\geq 0}$ of bounded linear operators on X, (ii) the control operator B is a bounded linear operator from $\mathbb{R}^M$ to X_{-1}, (iii) the observation operator C is a bounded linear operator from X_1 to $\mathbb{R}^Q$, and (iv) the feedthrough operator D is a linear operator from $\mathbb{R}^M$ to $\mathbb{R}^Q$. Here X_1 denotes the space $\mathrm{dom}(A)$ (the domain of A) endowed with the graph norm, and X_{-1} denotes the completion of X with respect to the norm $\|z\|_{-1} = \|(s_0 I - A)^{-1} z\|$, where s_0 is any fixed element of the resolvent set of A and $\|\cdot\|$ denotes the norm on X. As a regular linear system, the transfer function $\mathbf{G}$ is holomorphic and bounded on every half-plane $\mathbb{C}_\alpha$ with $\alpha > \omega(\mathbf{S}) := \lim_{t\to\infty} t^{-1} \ln \|\mathbf{S}_t\|$. Moreover,

$$\lim_{s\to\infty,\ s\in\mathbb{R}} \mathbf{G}(s) = D.$$

The system is said to be exponentially stable if the semigroup $\mathbf{S}$ is exponentially stable—that is, if $\omega(\mathbf{S}) < 0$. Henceforth, we assume that the system is exponentially stable and, moreover, we assume that the observation operator C can be extended to a bounded linear operator from X to $\mathbb{R}^Q$; this extended operator is again denoted by C.

In terms of the generating operators (A, B, C, D), the transfer function $\mathbf{G}$ is given by

$$\mathbf{G}(s) = C(sI - A)^{-1}B + D.$$

For any $z^0 \in X$ and input $y \in L^\infty_{\rm loc}(\mathbb{R}_+;\mathbb{R}^M)$, the state $z(\cdot)$ and the output $w(\cdot)$ of the regular system (with bounded observation operator) satisfy the equations

$$\dot z(t) = Az(t) + By(t)\,, \quad z(0) = z^0, \tag{2.6}$$

$$w(t) = Cz(t) + Dy(t) \tag{2.7}$$

for almost all $t \geq 0$. The derivative on the left-hand side of (2.6) has, of course, to be understood in X_{-1}. In other words, if we consider the initial-value problem (2.6) in the space X_{-1}, then for any $z^0 \in X$ and $y \in L^\infty_{\rm loc}(\mathbb{R}_+;\mathbb{R}^M)$, (2.6) has a unique strong solution given by the variation of parameters formula (see [16, Chapter 4, Theorem 2.9])

$$z(t) = \mathbf{S}_t z^0 + \int_0^t \mathbf{S}_{t-s} By(s)\, ds. \tag{2.8}$$

Restricting to continuous inputs, define the operator $T : C(\mathbb{R}_+;\mathbb{R}^M) \to L^\infty_{\rm loc}(\mathbb{R}_+;\mathbb{R}^Q)$ by

$$(Ty)(t) := C\int_0^t \mathbf{S}_{t-s} By(s)\ ds + Dy(t), \quad t \geq 0. \tag{2.9}$$

(We remark that the above operator is the infinite-dimensional counterpart of the operator (2.4) in the case of the finite-dimensional prototype.) By exponential stability of the semigroup $\mathbf{S}$, there then exist constants $c_1 > 0$ such that

$$\|z\|_{L^\infty(\mathbb{R}_+;X)} \leq c_1\left[\|z^0\| + \|y\|_{L^\infty(\mathbb{R}_+;\mathbb{R}^M)}\right] \quad \text{for all } \ (z^0, y) \in X \times L^\infty(\mathbb{R}_+;\mathbb{R}^M). \tag{2.10}$$

Setting $h = 0$, we see that property 2(a) of Definition 2.1 holds and property 2(b) is a consequence of the linearity of T and (2.10), in view of (2.10), and causality property 1 of Definition 2.1 also holds. Therefore, the operator T is of class $\mathcal{T}_0^{M,Q}$.

2.3. Nonlinear systems. *Input-to-state stable (ISS) systems.* Let $Z : \mathbb{R}^L \times \mathbb{R}^M \to \mathbb{R}^L$ be locally Lipschitz with $Z(0,0) = 0$. For $y \in L^\infty_{\rm loc}(\mathbb{R}_+;\mathbb{R}^M)$, let $z(\cdot, z^0, y)$ denote the maximal solution of the initial-value problem

$$\dot z(t) = Z(z(t), y(t)), \quad z(0) = z^0 \in \mathbb{R}^L. \tag{2.11}$$

Assume that the system is input-to-state stable (ISS) [20]; that is, there exist functions $\theta \in \mathcal{KL}$ and $\gamma \in \mathcal{K}$ such that, for all $(z^0, y) \in \mathbb{R}^L \times L^\infty_{\rm loc}(\mathbb{R}_+;\mathbb{R}^M)$,

$$\|z(t, z^0, y)\| \leq \theta(\|z^0\|, t) + \operatorname*{ess\,sup}_{s\in[0,t]} \gamma(\|y(s)\|) \quad \text{for all } \ t \geq 0. \tag{2.12}$$

Let $W : \mathbb{R}^L \to \mathbb{R}^Q$ be locally Lipschitz and such that there exists $c > 0$ such that $\|W(z)\| \leq c\|z\|$ for all $z \in \mathbb{R}^L$. Now consider system (2.11) with output w given by

$$w(t) = W(z(t, z^0, y)).$$

Fix $z^0 \in \mathbb{R}^L$ arbitrarily. Again, restricting to continuous inputs, define the operator $T : C(\mathbb{R}_+;\mathbb{R}^M) \to L^\infty_{\rm loc}(\mathbb{R}_+;\mathbb{R}^Q)$ by

$$(Ty)(t) := W(z(t, z^0, y)), \quad t \geq 0. \tag{2.13}$$

In view of (2.12), property 1 of Definition 2.1 evidently holds; setting $h = 0$, we see that property 2(a) also holds. Arguing as in [19, section 3.2.3], via an application of Gronwall's lemma, it can be shown that property 2(b) holds. Therefore, the operator T is of class $\mathcal{T}_0^{M,Q}$. We note that, strictly speaking, the above construction yields a family of operators T_{z^0} parameterized by the initial data z^0.

Systems in input affine form. A particular generalization of the prototype class $\mathcal{L}$ of linear, finite-dimensional, minimum-phase systems is the class of nonlinear systems in input affine form

$$\left.\begin{array}{ll} \dot{y}(t) = a(t, y(t), z(t)) + b(t, y(t), z(t))u(t), & y(0) = y^0 \\ \dot{z}(t) = c(t, y(t), z(t)), & z(0) = z^0 \end{array}\right\} \tag{2.14}$$

where $a : \mathbb{R}_+ \times \mathbb{R}^M \times \mathbb{R}^L \to \mathbb{R}^M$, $b : \mathbb{R}_+ \times \mathbb{R}^M \times \mathbb{R}^L \to \mathbb{R}^{M\times M}$, and $c : \mathbb{R}_+ \times \mathbb{R}^M \times \mathbb{R}^L \to \mathbb{R}^L$ are Carathéodory functions and (y_e, z_e, u_e) is an equilibrium ($(y_e, z_e, u_e) = (0, 0, 0)$ in the linear prototype) in the sense that

$$a(t, y_e, z_e) = 0, \quad b(t, y_e, z_e)u_e = 0, \quad c(t, y_e, z_e) = 0 \quad \text{for all } t \geq 0.$$

The problem of construction of a λ-servomechanism for such systems has been investigated in [1, 6]. There, the minimum-phase property of the linear prototype in (2.3) is replaced by the assumptions that z_e is a global, uniformly exponentially stable equilibrium of

$$\dot{\eta}(t) = c(t, y_e, \eta(t)). \tag{2.15}$$

We assume that (i) for each compact set $C \subset \mathbb{R}^M \times \mathbb{R}^L$, there exists $\kappa \in L^1_{\mathrm{loc}}(\mathbb{R}_+)$ such that $\|c(t, y, z) - c(t, \xi, \zeta)\| \leq \kappa(t)\|(y, z) - (\xi, \zeta)\|$ for almost all $t \in \mathbb{R}_+$ and all $(y, z), (\xi, \zeta) \in C$, and (ii) for some constant $c_0 > 0$,

$$\|c(t, y, z) - c(t, y_e, z)\| \leq c_0\ [1 + \|y - y_e\|] \quad \text{for all } (t, y, z) \in \mathbb{R}_+ \times \mathbb{R}^M \times \mathbb{R}^L.$$

Considering the second equations of (2.14) in isolation, for $y \in L^\infty_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^M)$ we denote by $z(\cdot, z^0, y)$ the unique solution of

$$\dot{z}(t) = c(t, y(t), z(t)) = c(t, y_e, z(t)) + [c(t, y(t), z(t)) - c(t, y_e, z(t))], \quad z(0) = z^0.$$

Invoking exponential stability of the equilibrium of (2.15) in conjunction with converse Lyapunov theory (details omitted here), we may conclude the existence of a constant $c_1 > 0$ such that, for each $(z^0, y) \in \mathbb{R}^L \times L^\infty_{\mathrm{loc}}(\mathbb{R}_+; \mathbb{R}^M)$,

$$\|z(t, z^0, y)\| \leq c_1[\|z^0\| + 1 + \operatorname*{ess\,sup}_{s\in[0,t]} \|y(s)\|] \quad \text{for all } t \geq 0. \tag{2.16}$$

Fix $z^0 \in \mathbb{R}^L$ arbitrarily. Define the operator $T : C(\mathbb{R}_+; \mathbb{R}^M) \to L^\infty_{\mathrm{loc}}(\mathbb{R}_+; \mathbb{R}^L)$ by

$$(Ty)(t) := z(t, z^0, y), \quad t \geq 0. \tag{2.17}$$

In view of (2.16), property 1 of Definition 2.1 evidently holds; setting $h = 0$, we see that property 2(a) also holds. An application of Gronwall's lemma (analogous to that adopted in [19, section 3.2.3] in the context of ISS systems) yields property 2(b). Therefore, the operator T is of class $\mathcal{T}_0^{M,L}$. As in the case of ISS systems, we remark that, strictly speaking, the above construction yields a family of operators T_{z^0} parameterized by the initial data z^0.

The general case. Elaborating on the above two cases, consider the system

$$\dot{z}(t) = Z(t, z(t), y(t)), \quad z(0) = z^0 \in \mathbb{R}^L, \tag{2.18}$$

with input $y \in L^\infty_{\mathrm{loc}}(\mathbb{R}_+; \mathbb{R}^M)$ and output

$$w(t) = W(t, z(t)) \in \mathbb{R}^Q.$$

Assume that $W : \mathbb{R}_+ \times \mathbb{R}^L \to \mathbb{R}^Q$ and $Z : \mathbb{R}_+ \times \mathbb{R}^L \times \mathbb{R}^M \to \mathbb{R}^L$ are Carathéodory functions and such that the following hold: (i) for some constant $c > 0$, $\|W(t, z)\| \le c\|z\|$ for almost all $t \ge 0$ and all $z \in \mathbb{R}^L$; (ii) for each compact set $C \subset \mathbb{R}^L \times \mathbb{R}^M$, there exists $\kappa \in L^1_{\mathrm{loc}}(\mathbb{R}_+)$ such that $\|Z(t, z, y) - Z(t, \zeta, \xi)\| \le \kappa(t)\|(z, y) - (\zeta, \xi)\|$ for almost all $t \in \mathbb{R}_+$ and all $(z, y), (\zeta, \xi) \in C$; and (iii) for each $(z^0, y) \in \mathbb{R}^L \times L^\infty_{\mathrm{loc}}(\mathbb{R}_+; \mathbb{R}^M)$, the unique maximal solution of initial-value problem (2.18) has interval of existence $\mathbb{R}_+$. (We denote the solution by $z(\cdot, z^0, y)$.) Furthermore, we assume the existence of a function $\gamma \in \mathcal{K}$ such that, for each $z^0 \in \mathbb{R}^L$, there exists a constant $c > 0$ such that, for all $y \in L^\infty_{\mathrm{loc}}(\mathbb{R}_+; \mathbb{R}^M)$,

$$\|z(t, z^0, y)\| \le c[1 + \text{ess-sup}_{s \in [0,t]} \gamma(\|y(s)\|)] \quad \text{for all } t \ge 0 \tag{2.19}$$

(a weaker condition than the ISS inequality (2.12)). Fix $z^0 \in \mathbb{R}^L$ arbitrarily. Define the operator $T : C(\mathbb{R}_+; \mathbb{R}^M) \to L^\infty_{\mathrm{loc}}(\mathbb{R}_+; \mathbb{R}^Q)$ by

$$(Ty)(t) = W(t, z(t, z^0, y)), \quad t \ge 0.$$

Then this construction yields a family (parameterized by the initial data z^0) of operators T of class $\mathcal{T}_0^{M,Q}$: This family subsumes the operators discussed in sections 2.2 and 2.3 above.

2.4. Nonlinear delay elements. Let $\mathcal{D}^{M,Q}$ denote the class of functions $\mathbb{R} \times \mathbb{R}^M \to \mathbb{R}^Q : (t, y) \mapsto \Psi(t, y)$ that are measurable in t and locally Lipschitz in y uniformly with respect to t. Precisely, (i) for each fixed y, $\Psi(\cdot, y)$ is measurable, and (ii) for every compact $C \subset \mathbb{R}^M$ there exists a constant c such that

$$\text{for a.a. } t, \quad \|\Psi(t, y) - \Psi(t, z)\| \le c\|y - z\| \quad \text{for all } y, z \in C.$$

For $i = 0, \dots, n$, let $\Psi_i \in \mathcal{D}^{M,Q}$ and $h_i \in \mathbb{R}_+$. Define $h := \max_i h_i$. For $y \in C([-h, \infty); \mathbb{R}^M)$, let

$$(Ty)(t) := \int_{-h_0}^{0} \Psi_0(s, y(t + s))\, ds + \sum_{i=1}^{n} \Psi_i(t, y(t - h_i)), \quad t \ge 0. \tag{2.20}$$

The operator T, so defined, is of class $\mathcal{T}_h^{M,Q}$; for details, see [19].

2.5. Hysteresis. A general class of nonlinear operators $C(\mathbb{R}_+; \mathbb{R}) \to C(\mathbb{R}_+; \mathbb{R})$, which includes many physically motivated hysteretic effects, is defined via assumptions (N1)–(N8) of [11, section 3]. Assumption (N1) implies that property 2(a) of Definition 2.1 holds with $h = 0$. Assumption (N5) implies that property 2(b) of Definition 2.1 holds. Finally, (N8) implies that property 1 of Definition 2.1 holds. Therefore, the nonlinear operators considered in [11] are of class $\mathcal{T}_0^{1,1}$. Examples of such operators, including relay hysteresis, backlash hysteresis, elastic-plastic hysteresis, and Preisach operators, are detailed in [11, section 5]. By way of illustration, we briefly describe the first two of these examples.

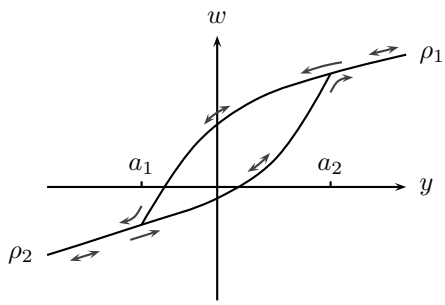


FIG. 2. *Relay hysteresis.*

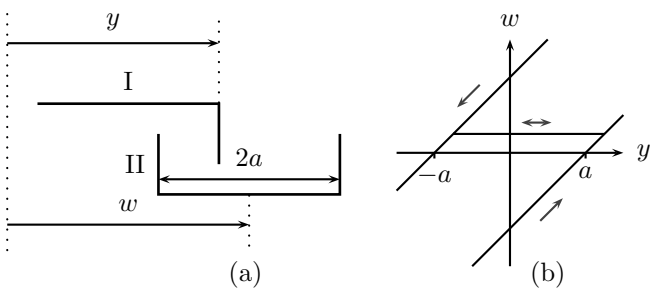


FIG. 3. *Backlash hysteresis.*

Relay hysteresis. Let $a_1 < a_2$ and let $\rho_1 : [a_1, \infty) \to \mathbb{R}$, $\rho_2 : (-\infty, a_2] \to \mathbb{R}$ be continuous, globally Lipschitz, and satisfying $\rho_1(a_1) = \rho_2(a_1)$ and $\rho_1(a_2) = \rho_2(a_2)$. For a given input $y \in C(\mathbb{R}_+; \mathbb{R})$ to the hysteresis element, the output w is such that $(y(t), w(t)) \in \operatorname{graph}(\rho_1) \cup \operatorname{graph}(\rho_2)$ for all $t \in \mathbb{R}_+$: The value $w(t)$ of the output at $t \in \mathbb{R}_+$ is either $\rho_1(y(t))$ or $\rho_2(y(t))$, depending on which of the threshold values a_2 or a_1 was "last" attained by the input y. This situation is illustrated by Figure 2.

When suitably initialized, such a hysteresis element has the property that, to each input $y \in C(\mathbb{R}_+; \mathbb{R})$ there corresponds a unique output $w = Ty \in C(\mathbb{R}_+; \mathbb{R})$; the operator T, so defined, is of class $\mathcal{T}_0^{1,1}$. Full details may be found in [11, section 5]. (See also [12, 10].)

Backlash hysteresis. Next consider a one-dimensional mechanical link consisting of the two solid parts I and II, as shown in Figure 3(a), the displacements of which (with respect to some fixed datum) at time $t \geq 0$ are given by $y(t)$ and $w(t)$ with $|y(t) - w(t)| \leq a$ for all t, and $w(0) := y(0) + \xi$ for some prespecified $-a \leq \xi \leq a$.

Within the link there is mechanical play; that is to say, the position $w(t)$ of II remains constant as long as the position $y(t)$ of I remains within the interior of II. Thus, assuming the continuity of y, we have $\dot{w}(t) = 0$ whenever $|y(t) - w(t)| < a$. Given a continuous input $y \in C(\mathbb{R}_+; \mathbb{R})$, describing the evolution of the position of I, denote the corresponding position of II by $w = Ty$. The operator T so defined (in effect we define a family T_ξ of operators parameterized by the initial offset ξ) is known as *backlash* or *play* and is of class $\mathcal{T}_0^{1,1}$. Full details may be found in [11, section 5].

3. Adaptive control. We now focus on the adaptive control problem. The following subclass $\mathcal{J}$ of $\mathcal{K}$ functions will play an important role:

$$\mathcal{J} := \{\alpha \in \mathcal{K} |\ \text{ for each } \ \delta \in \mathbb{R}_+ \text{ there exists } \Delta \in \mathbb{R}_+ :\ \alpha(\delta\tau) \le \Delta\alpha(\tau) \ \text{ for all } \ \tau \ge 0\}.$$

Furthermore, we define $\mathcal{J}_\infty := \mathcal{J} \cap \mathcal{K}_\infty$. For example, (a) for each $s > 0$, the function $\tau \mapsto \tau^s$ is of class $\mathcal{J}_\infty$, and (b) the function $\tau \mapsto \ln(1+\tau)$ is of class $\mathcal{J}_\infty$; its inverse $\tau \mapsto \exp(\tau) - 1$ is of class $\mathcal{K}_\infty$ but is *not* of class $\mathcal{J}$. In addition to their defining property, the ensuing properties of class $\mathcal{J}$ functions are readily established and will be freely invoked later in the analysis:

1. $\alpha, \beta \in \mathcal{J} \quad \Longrightarrow \quad \alpha \circ \beta \in \mathcal{J}$ and $\alpha + \beta \in \mathcal{J}$;
2. $\alpha \in \mathcal{J} \quad \Longrightarrow \quad \exists \Delta > 0 :\ \alpha(a+b) \le \Delta[\alpha(a) + \alpha(b)] \quad$ for all $\ a, b \in \mathbb{R}_+$.

We also record a property of $\mathcal{K}$ functions (and, a fortiori, a property of $\mathcal{J}$ functions):

3. Let $t > 0$, $I = [0,t]$, $\xi \in C(I;\mathbb{R}_+)$, and $\alpha \in \mathcal{K}$; then $\alpha(\max_{s\in I} \xi(s)) = \max_{s\in I} \alpha(\xi(s))$.

DEFINITION 3.1 (the system class). *Let $\alpha_f, \alpha_T \in \mathcal{J}$; then $\mathcal{S} = \mathcal{S}(\alpha_f, \alpha_T)$ denotes the class of M-input, M-output systems of the form* (1.1) *with the following properties (wherein $P, Q \in \mathbb{N}$ are arbitrary):*

1. $p \in L^\infty([-h,\infty);\mathbb{R}^P)$;
2. $f : \mathbb{R}^P \times \mathbb{R}^Q \to \mathbb{R}^M$ *is continuous and, for every compact set $C \subset \mathbb{R}^P$, there exists a constant $c_f \ge 0$ such that*

$$\|f(p,w)\| \le c_f\,[1 + \alpha_f(\|w\|)] \quad \text{for all } \ (p,w) \in C \times \mathbb{R}^Q\,;$$

3. $g : \mathbb{R}^P \times \mathbb{R}^Q \times \mathbb{R}^M \to \mathbb{R}^M$ *is continuous and, for every compact set $C \subset \mathbb{R}^P$, there exists a positive definite, symmetric $G \in \mathbb{R}^{M\times M}$ such that*

$$\langle Gu, g(p,w,u)\rangle \ge \|u\|^2 \quad \text{for all } (p,w,u) \in C \times \mathbb{R}^Q \times \mathbb{R}^M\,;$$

4. $T : C([-h,\infty);\mathbb{R}^M) \to L^\infty_{\mathrm{loc}}(\mathbb{R}_+;\mathbb{R}^Q)$ *is of class $\mathcal{T}_h^{M,Q}$, and there exist $\alpha_T \in \mathcal{J}$ and constant $c_T \ge 0$ such that, for all $y \in C([-h,\infty);\mathbb{R}^M)$,*

$$\|(Ty)(t)\| \le c_T \left[1 + \max_{s\in[0,t]} \alpha_T(\|y(s)\|)\right] \quad \text{for almost all } \ t \in \mathbb{R}_+\,. \tag{3.1}$$

For convenience, we denote a system of class $\mathcal{S}(\alpha_f, \alpha_T)$ by $(p, f, g, T) \in \mathcal{S}(\alpha_f, \alpha_T)$ and, whenever the functions α_f and α_T are contextually evident, we simply write $\mathcal{S}$ in place of $\mathcal{S}(\alpha_f, \alpha_T)$. We emphasize that, in the construction of an $(\mathcal{R}, \mathcal{S})$-universal control strategy, only the (instantaneous) tracking error $e(t) = y(t) - r(t)$ is assumed to be available for feedback, and the only a priori structural information assumed is knowledge of the functions α_f, $\alpha_T \in \mathcal{J}$. Some examples follow.

Assume f has the polynomial form given by $f(p,w) := \sum_{i=0}^{l} p_i w^i$. Then property 2 of Definition 3.1 holds with $\alpha_f : s \mapsto s^m$ for $m \ge l$; if an upper bound for the degree l of the polynomial is unknown, then the map $\alpha_f : s \mapsto \exp(s) - 1$ suffices.

If $g(p,w,u) = Bu$, as in the linear prototype (2.3), and $B \in \mathbb{R}^{M\times M}$ has spectrum in the open right half complex plane, then there exists a positive definite $G \in \mathbb{R}^{M\times M}$ satisfying $GB + B^TG = 2I$, whence property 3 of Definition 3.1.

Consider again the examples of operators in sections 2.2–2.5.

Let $T \in \mathcal{T}_h^{M,Q}$, given by (2.9), be the input-output operator of an exponentially stable regular linear system with $\mathbb{R}^M$-valued input and $\mathbb{R}^Q$-valued output. Then (2.10) and causality imply that (3.1) holds with the $\alpha_T \in \mathcal{J}$ given by $\alpha_T(s) = s$.

Let $T \in \mathcal{T}_h^{M,Q}$, given by (2.13), be the input-output operator of an ISS system with $\mathbb{R}^M$-valued input and $\mathbb{R}^Q$-valued output. If (2.12) holds for some function γ of class $\mathcal{J}$, then (3.1) holds with $\alpha_T := \gamma$.

Let $\beta \in \mathcal{J}$, $h \in \mathbb{R}_+$, and $\Psi \in \mathcal{D}^{M,Q}$ (recall section 2.4), and assume that

$$\|\Psi(t,y)\| < \mu\,[1 + \beta(\|y\|)] \quad \text{for all } (t,y) \in \mathbb{R}_+ \times \mathbb{R}^M$$

for some $\mu \in \mathbb{R}_+$. Both the point delay given by $(Ty)(t) = \Psi(t, y(t-h))$ and the distributed delay given by $(Ty)(t) = \int_{-h}^{0} \Psi(s, y(t+s))\, ds$ are of class $\mathcal{T}_h^{M,Q}$, and (3.1) holds with $\alpha_T := \beta$.

Last, for the nonlinear operators of section 2.5, assumption (N8) of [11, section 3] asserts that such operators satisfy (3.1) with the $\alpha_T \in \mathcal{J}$ given by $\alpha_T(s) = s$.

3.1. The servomechanism. The servomechanism is designed as follows. Let $\alpha_f, \alpha_T \in \mathcal{J}$. Choose $\alpha \in \mathcal{J}_\infty$ with the property

$$\liminf_{s\to\infty} \frac{\alpha(s)}{s + \alpha_f(\alpha_T(s))} \neq 0. \tag{3.2}$$

For example, the choice $\alpha : s \mapsto s + \alpha_f(\alpha_T(s))$ suffices. For $\lambda > 0$, choose $\psi_\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ to be a continuous function with the properties

$$\text{(i)} \quad \liminf_{s\to\infty} \frac{s\psi_\lambda(s)}{\alpha(s)} \neq 0 \quad \text{and} \quad \text{(ii)} \quad \psi_\lambda^{-1}(0) := \{s \mid \psi_\lambda(s) = 0\} = [0, \lambda]. \tag{3.3}$$

For example, the choice ψ_λ given by $\psi_\lambda(s) := d_{\alpha(\lambda)}(\alpha(s))/s$ for $s > 0$, with $\psi_\lambda(0) := 0$, suffices.

Define the continuous function

$$\phi : \mathbb{R}^M \to \mathbb{R}^M, \quad e \mapsto \begin{cases} \alpha(\|e\|)\|e\|^{-1}e, & e \neq 0, \\ 0, & e = 0. \end{cases} \tag{3.4}$$

Writing $\mathcal{S} = \mathcal{S}(\alpha_f, \alpha_T)$, the next objective is to show that the strategy

$$u(t) = -k(t)\phi(e(t)), \quad \dot{k}(t) = \psi_\lambda(\|e(t)\|), \quad e(t) := y(t) - r(t) \tag{3.5}$$

is an $(\mathcal{R}, \mathcal{S})$-universal λ-servomechanism.

THEOREM 3.2. *Let $\alpha_f, \alpha_T \in \mathcal{J}$. Choose $\alpha \in \mathcal{J}_\infty$ so that* (3.2) *holds and define the continuous $\phi : \mathbb{R}^M \to \mathbb{R}^M$ by* (3.4). *Let $\lambda > 0$ and let $\psi_\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ be continuous with properties* (3.3). *Then feedback strategy* (3.5) *is an $(\mathcal{R}, \mathcal{S})$-universal λ-servomechanism in the sense that for all $(p, f, g, T) \in \mathcal{S}(\alpha_f, \alpha_T)$, $r \in \mathcal{R}$, and $(y^0, k^0) \in C([-h, 0]; \mathbb{R}^{M+1})$ the feedback controlled initial-value problem*

$$\left.\begin{aligned} \dot{y}(t) &= f(p(t), (Ty)(t)) + g(p(t), (Ty)(t), -k(t)\phi(y(t) - r(t))) \\ \dot{k}(t) &= \psi_\lambda(\|y(t) - r(t)\|) \\ (y,k)|_{[-h,0]} &= (y^0, k^0) \end{aligned}\right\} \tag{3.6}$$

has a solution. Every solution can be extended to a maximal solution and every maximal solution $(y,k) : [0,\omega) \to \mathbb{R}^{M+1}$ has the following properties:

(i) *(y,k) is bounded;*

(ii) *$\omega = \infty$;*

(iii) $\lim_{t\to\infty} k(t)$ *exists and is finite;*
(iv) $\lim_{t\to\infty} d_\lambda(\|y(t)-r(t)\|)=0$, *with* d_λ *as in* (1.6).

We preface the proof of Theorem 3.2 by a proposition. (Proof of the latter is straightforward and omitted here.)

PROPOSITION 3.3. *Let* $\xi \in AC_{loc}(\mathbb{R}_+;\mathbb{R}_+)$, $k \in C(\mathbb{R}_+;\mathbb{R}_+)$, $\beta \in \mathcal{K}$, *and* $c \geq 0$. *If* k *is monotonically nondecreasing and unbounded, and* $\dot{\xi}(t) \leq c - k(t)\beta(\xi(t))$ *for almost all* $t \in \mathbb{R}_+$, *then* $\xi(t) \to 0$ *as* $t \to \infty$.

Proof of Theorem 3.2. Write $N := M+1$ and $K := Q+M+1$. Define $F : [-h,\infty)\times\mathbb{R}^K \to \mathbb{R}^N$ by (1.3) and define $\widehat{T} : C([-h,\infty);\mathbb{R}^N) \to L^\infty_{\mathrm{loc}}(\mathbb{R}_+;\mathbb{R}^K)$ by (1.5). Thus, the initial-value problem (3.6) is equivalent to (2.2). By the continuity of f, g, ϕ, ψ_λ and (essential) boundedness of p, it follows that F is a Carathéodory function with the property that, for each $w \in \mathbb{R}^K$, $F(\cdot,w) \in L^\infty_{\mathrm{loc}}([-h,\infty);\mathbb{R}^N)$. By assumption, $T \in \mathcal{T}_h^{M,Q}$ and so $\widehat{T} \in \mathcal{T}_h^{N,K}$. Therefore, by Theorem 2.3, (3.6) has a solution and every solution can be maximally extended. Moreover, every *bounded* maximal solution has interval of existence $[-h,\infty)$.

Let $(y,k) : [-h,\omega) \to \mathbb{R}^N$ be a maximal solution of (3.6). Writing $e := y - r$, we have

(3.7)
$$\left.\begin{aligned} \dot{e}(t) &= f(p(t),(T(e+r))(t)) \\ &\quad + g(p(t),(T(e+r))(t)),-k(t)\phi(e(t))) - \dot{r}(t) \\ \dot{k}(t) &= \psi_\lambda(\|e(t)\|) \end{aligned}\right\} \quad \text{for a.a. } t \in [0,\omega).$$

By (essential) boundedness of p and property 3 of Definition 3.1 of g, there exists a positive definite, symmetric G such that

(3.8)
$$\langle Ge(t), g(p(t),(T(e+r))(t)),-k(t)\phi(e(t)))\rangle \leq -k(t)\alpha(\|e(t)\|)\|e(t)\| \quad \text{for a.a. } t \in [0,\omega).$$

Define $c_0 := \sqrt{2\|G^{-1}\|}$ and $c_1 := \sqrt{2/\|G\|}$. For notational convenience, we introduce functions $V, W \in AC_{\mathrm{loc}}([0,\omega);\mathbb{R}_+)$ given by

$$V(t) := \frac{1}{2}\langle Ge(t), e(t)\rangle \quad \text{and} \quad W(t) := \sqrt{V(t)}$$

with

$$c_0^{-1}\|e(t)\| \leq W(t) \leq c_1^{-1}\|e(t)\| \quad \text{for all } t \in [0,\omega). \tag{3.9}$$

By (3.7), (3.8) and properties of f, g, and T, together with (essential) boundedness of p, r, and $\dot{r}$, there exist constants $c_f, c_T > 0$ such that

(3.10)
$$\begin{aligned} \dot{V}(t) = \langle Ge(t), \dot{e}(t)\rangle &\leq c_f\|G\|\left[1+\alpha_f\left(c_T + c_T \max_{s\in[0,t]} \alpha_T(\|e(s)+r(s)\|)\right)\right]\|e(t)\| \\ &\quad - k(t)\alpha(\|e(t)\|)\|e(t)\| + \|G\|\|r\|_{1,\infty}\|e(t)\| \quad \text{for a.a. } t \in [0,\omega). \end{aligned}$$

Invoking properties of $\mathcal{J}$ functions, we may conclude that, for some constant $c_2 > 0$,

(3.11)
$$\dot{V}(t) \leq c_2\left[1 + \max_{s\in[0,t]} \alpha_f(\alpha_T(\|e(s)\|))\right]\|e(t)\| - k(t)\alpha(\|e(t)\|)\|e(t)\| \quad \text{a.a. } t \in [0,\omega).$$

By (3.2) and the first of properties (3.3), there exist constants $\gamma > \|e(0)\|$, c_γ, $\tilde{c}_\gamma > 0$ such that

(3.12)
$$\alpha_f(\alpha_T(s)) \le c_\gamma \alpha(s) \quad \text{for all } s \ge \gamma \qquad \text{and} \qquad \psi_\lambda(s) \ge \frac{c_\gamma \alpha(s)}{\tilde{c}_\gamma s} \quad \text{for all } s \ge \gamma.$$

With a view to proving Theorem 3.2(i), we first show that e is bounded. Seeking a contradiction, suppose that e (equivalently, W) is unbounded. For each $n \in \mathbb{N}$, define

$$\tau_n := \inf\{t \in [0,\omega) \mid c_1 W(t) = n + 1 + \gamma\}, \quad \sigma_n := \sup\{t \in [0,\tau_n] \mid c_1 W(t) = n + \gamma\}.$$

Recalling that $\gamma > \|e(0)\| \ge c_1 W(0)$, this construction yields a sequence of disjoint intervals (σ_n, τ_n) such that

$$\left.\begin{aligned} &\max_{t\in[0,\tau_n]} c_1 W(t) = c_1 W(\tau_n) = n + 1 + \gamma \\ &c_1 W(\sigma_n) = n + \gamma \\ &c_1 W(t) \in (n + \gamma, n + 1 + \gamma) \ \text{ for all } \ t \in (\sigma_n, \tau_n) \end{aligned}\right\} \quad \text{for all } \ n \in \mathbb{N}.$$

Moreover, for all $n \in \mathbb{N}$,

$$\max_{s\in[0,t]} c_1 W(s) = \max_{s\in[\sigma_n,t]} c_1 W(s) \le n + 1 + \gamma < 2n + 2\gamma \le 2c_1 W(t) \quad \text{for all } \ t \in [\sigma_n, \tau_n],$$

which, together with (3.9) and properties of $\mathcal{J}$ functions, implies the existence of constants $c_3, c_4 > 0$ such that

$$\begin{aligned} (3.13)\quad \max_{s\in[0,t]} \alpha(\|e(s)\|) &\le \max_{s\in[0,t]} \alpha(c_0 W(s)) \le \alpha(2c_0 W(t)) \le \alpha(2c_0 c^{-1} \|e(t)\|) \\ &\le c_3 \alpha(\|e(t)\|) \le c_3 \alpha(c_0 W(t)) \le c_4 \alpha(c_1 W(t)) \quad \text{for all } \ t \in \cup_{n\in\mathbb{N}} [\sigma_n, \tau_n]. \end{aligned}$$

Noting that, for all $n \in \mathbb{N}$, $\alpha(\|e(t)\|) \ge \alpha(\gamma)$ for all $t \in [\sigma_n, \tau_n]$ and invoking (3.13) together with (3.9), (3.11), and (3.12), we may conclude the existence of constants $c_5, c_6 > 0$ such that

(3.14)
$$\dot{V}(t) \le [c_5 - k(t)]\alpha(\|e(t)\|)\|e(t)\| \le c_6 \alpha(c_1 W(t)) W(t) \quad \text{for all } \ t \in \cup_{n\in\mathbb{N}} [\sigma_n, \tau_n].$$

Our next task is to show that supposition of the unboundedness of e implies the unboundedness of k. Invoking (3.12), (3.14), and (3.9) yields

$$\begin{aligned} 2\ln\left(\frac{n+1+\gamma}{1+\gamma}\right) &= \ln V(\tau_n) - \ln V(\sigma_1) = \sum_{j=1}^{n} [\ln V(\tau_j) - \ln V(\sigma_j)] = \sum_{j=1}^{n} \int_{\sigma_j}^{\tau_j} \frac{\dot{V}(t)}{V(t)}\, dt \\ &\le c_6 \sum_{j=1}^{n} \int_{\sigma_j}^{\tau_j} \frac{\alpha(c_1 W(t))}{W(t)}\, dt \le c_6 c_0 \sum_{j=1}^{n} \int_{\sigma_j}^{\tau_j} \frac{\alpha(\|e(t)\|)}{\|e(t)\|}\, dt. \end{aligned} \tag{3.15}$$

By construction of (σ_n, τ_n) we have

$$\gamma < \|e(t)\| \qquad \text{if } t \in (\sigma_j, \tau_j).$$

Hence substituting the second inequality of (3.12) into (3.15) yields

$$2\ln\left(\frac{n+1+\gamma}{1+\gamma}\right) \le c_6 c_0 \frac{\tilde{c}_\gamma}{c_\gamma} \sum_{j=1}^{n} \int_{\sigma_j}^{\tau_j} \psi_\lambda(\|e(t)\|)\, dt \le c_6 c_0 \frac{\tilde{c}_\gamma}{c_\gamma} k(\tau_n) \quad \text{for all } n \in \mathbb{N},$$

and so $k(t) \to \infty$ as $t \uparrow \omega$. Let $n^* \in \mathbb{N}$ be such that $k(\sigma_{n^*}) \ge 2c_5$. By the first inequality in (3.14),

$$\dot{V}(t) \le -c_5 \alpha(\|e(t)\|)\|e(t)\| < 0 \quad \text{for a.a. } t \in [\sigma_{n^*}, \tau_{n^*}]\,,$$

which contradicts the fact that $V(\tau_{n^*}) = W^2(\tau_{n^*}) > W^2(\sigma_{n^*}) = V(\sigma_{n^*})$. Therefore, e is bounded.

By the boundedness of e and continuity of ψ_λ, it follows that $\dot{k}$ is bounded, and so k is bounded on every compact subinterval of $[0, \omega)$. Therefore $\omega = \infty$.

Next, we prove the boundedness of k. By the boundedness of e and (3.11), there exists a constant $c_9 > 0$ such that

$$\dot{V}(t) \le c_9 - k(t)\beta(V(t)) \quad \text{for a.a. } t \in [0, \infty),$$

where $\beta \in \mathcal{K}$ is given by $\beta(s) = \alpha\,(c_1\sqrt{s})\, c_1\sqrt{s}$. Seeking a contradiction, suppose k is unbounded. Then $k(t) \uparrow \infty$ as $t \to \infty$ and so, by Proposition 3.3, $V(t) \to 0$ as $t \to \infty$. Therefore, there exists $\tau \in [0, \infty)$ such that $\|e(t)\| < \lambda$ for all $t \in [\tau, \infty)$ and so $\dot{k}(t) = 0$ for all $t \in [\tau, \infty)$, which again contradicts the supposition of the unboundedness of k.

We have now established Theorem 3.2(i) and (ii). Assertion (iii) follows by the boundedness and monotonicity of k. By the boundedness of e and $\dot{e}$ (see (3.6)), it follows that $t \mapsto e(t)$ is uniformly continuous. By the continuity of $\psi_\lambda(\|\cdot\|)$, we see that $\psi_\lambda(\|e(\cdot)\|)$ is also uniformly continuous. By the boundedness of k, $\int_0^\infty \psi_\lambda(\|e(t)\|)dt < \infty$. By Barbălat's lemma [2], we conclude that $\psi_\lambda(\|e(t)\|) \to 0$ as $t \to \infty$, whence, recalling that $\psi_\lambda^{-1}(0) = [0, \lambda]$, we have assertion (iv). □

3.2. Discussion. Theorem 3.2 also holds in the situation wherein the output measurement is subject to an additive disturbance term η, in which case the control and gain adaptation become

$$u(t) = -k(t)\phi(y(t) - r(t) + \eta(t)), \quad \dot{k}(t) = \psi_\lambda(\|y(t) - r(t) + \eta(t)\|), \quad k|_{[-h,0]} = k^0.$$

If the disturbance η is of class $\mathcal{R}$, then, by Theorem 3.2, $\lim_{t\to\infty} d_\lambda(\|y(t) + \eta(t) - r(t)\|) = 0$. Thus, from a strictly analytical viewpoint, in the presence of output disturbances of class $\mathcal{R}$, the disturbance-free analysis is immediately applicable to replacing the reference signal r by the signal $r - \eta =: \hat{r} \in \mathcal{R}$. Even though the reference signal r and disturbance signal η are assumed to be of the same class $\mathcal{R}$, in practice these signals might be distinguished by their respective spectra (η typically having "high-frequency" content). Moreover, from a practical viewpoint, one might reasonably expect that the disturbance η is "small"; if an a priori bound on the magnitude of the disturbance is available, then λ should be chosen to be commensurate with such a bound.

We remark on the flexibility of choice in the controller functions $\alpha \in \mathcal{J}_\infty$ and ψ_λ (continuous), which are required only to satisfy (3.2) and (3.3). In essence, (3.2) reflects the reasonable requirement that the "strength" of the controller nonlinearity α should be capable of counteracting the potentially destabilizing effects of the

(unknown) system nonlinearities; condition (3.3)(i) translates to a requirement that the gain adaptation function ψ_λ should be commensurate (in the sense of (3.3)(i)) with the strength of the function α. Next, we illustrate by example that the latter condition is also reasonable.

Consider the scalar nonlinear system

$$\dot{y}(t) = a|y(t)|^\epsilon y(t) + u(t), \quad y(0) = y^0 \in \mathbb{R}, \tag{3.16}$$

with $a \in \mathbb{R}$ and $\epsilon > 0$. The choice $\alpha : s \mapsto s^{1+\epsilon}$ implies that (3.2) holds. For $\lambda > 0$, the choice

$$\psi_\lambda : s \mapsto s^\epsilon \min\{d_\lambda(s), 1\} \tag{3.17}$$

implies that (3.3) holds. Therefore, by Theorem 3.2, the control

$$\begin{aligned} u(t) &= -k(t)|y(t) - r(t)|^\epsilon (y(t) - r(t)), \\ \dot{k}(t) &= |y(t) - r(t)|^\epsilon \min\{d_\lambda(|y(t) - r(t)|), 1\}, \quad k(0) = k^0, \end{aligned}$$

ensures that, for every $r \in \mathcal{R}$, the tracking objective is achieved with asymptotic accuracy quantified by $\lambda > 0$.

Now assume that $\epsilon > 0$ is "small." We will investigate the consequences of replacing the above choice of ψ_λ (for which (3.3)(i) holds) by the simpler function $s \mapsto \min\{d_\lambda(s), 1\}$ (equivalent to setting $\epsilon = 0$ in (3.17) and for which (3.3)(i) fails to hold). Taking $r = 0$, $a = 1$, $y^0 > 0$, and (for simplicity) $k^0 = 0$, a straightforward calculation reveals that the control objective is not achievable by the control

$$u(t) = -k(t)|y(t)|^\epsilon y(t), \quad \dot{k}(t) = \min\{d_\lambda(s), 1\}, \quad k(0) = 0.$$

In particular, the feedback-controlled initial-value problem can exhibit finite-time "blow-up" of its solution: Specifically, for each $y^0 > (2/\epsilon)^{1/\epsilon}$, the solution of the feedback-controlled system is such that $y(t) \uparrow \infty$ as $t \uparrow T$ with $T \in (0, T^*)$, where $T^* := 1 - \sqrt{1 - (2/\epsilon)(y^0)^{-\epsilon}} < 1$.

Now consider again linear systems, such as the motivating class $\mathcal{L}$ of finite-dimensional, linear, minimum-phase systems described in section 2.2, and let $\mathcal{R}$ be the space of bounded absolutely continuous functions $\mathbb{R} \to \mathbb{R}^M$ with essentially bounded derivative. As is well known (see, for example, [7]), the following output feedback strategy (a variant of the seminal results in [23, 15, 13, 14]) is an $(\mathcal{R}, \mathcal{L})$-universal λ-servomechanism in the sense that, for each system of class $\mathcal{L}$ and reference signal $r \in \mathcal{R}$, the strategy ensures (i) boundedness of the state, (ii) convergence of the controller gain, and (iii) output tracking with prescribed accuracy λ (in the sense that $d_\lambda(\|e(t)\|) \to 0$ as $t \to \infty$, where $e(t) := y(t) - r(t)$ is the tracking error):

$$u(t) = -k(t)e(t), \qquad \dot{k}(t) = d_\lambda^2(\|e(t)\|), \quad k(0) = k^0. \tag{3.18}$$

Generalizations of this strategy to nonlinear finite-dimensional settings are reported in, for example, [7, 17, 6, 18, 24]; applications to biotechnological processes are contained in [8, 9].

Each of α_f and α_T can be taken to be the identity map $id : s \mapsto s$, and so $\mathcal{L} \subset \mathcal{S}(id, id)$. In this context, $\alpha : s \mapsto s$ and $\psi_\lambda : s \mapsto d_\lambda^2(s)$ are allowable choices, in which case we recover (3.18). Note that the latter choice for ψ_λ, being quadratic in nature, implies that the controller gain $k(\cdot)$ can exhibit rapid growth whenever the tracking error is large. Such behavior may be undesirable from a practical viewpoint.

A very simple but admissible alternative choice of a *bounded* function ψ_λ is $s \mapsto \min\{d_\lambda(s), 1\}$. This choice ensures that k exhibits at most linear growth and the overall control strategy (3.5) reduces to

(3.19)
$$u(t) = -k(t)(y(t) - r(t)), \qquad \dot{k}(t) = \min\{d_\lambda(\|y(t) - r(t)\|), 1\}, \qquad k|_{[-h,0]} = k^0.$$

Theorem 3.2 ensures that this control achieves the tracking objective, with prespecified asymptotic error bound $\lambda > 0$, not only for the motivating finite-dimensional class $\mathcal{L}$, but also for general interconnections of linear systems of the form in Figure 1, encompassing those cases where Σ_2 corresponds to linear delay elements (both pointwise and distributed) or to an exponentially stable infinite-dimensional regular linear system (such as a diffusion process), or linear combinations of these.

Appendix. Proof of Theorem 2.3. (i) By property 2(b) of Definition 2.1 there exist $\tau > 0$, $\delta > 0$, and $c > 0$ such that, for all $x, \xi \in C([-h, \infty); \mathbb{R}^N)$ with $x|_{[-h,0]} = x^0 = \xi|_{[-h,0]}$ and $x(t), \xi(t) \in \mathbb{B}_\delta(x^0(0))$ for all $t \in [0, \tau]$,

$$\operatorname*{ess\,sup}_{t\in[0,\tau]} \|(\widehat{T}x)(t) - (\widehat{T}\xi)(t)\| \le c \sup_{t\in[0,\tau]} \|x(t) - \xi(t)\|.$$

By property 1 of Definition 2.1 of $\widehat{T}$, there exists $\Delta > 0$ such that for all $x \in C([-h, \infty); \mathbb{R}^N)$,

$$\sup_{t\in[-h,\infty)} \|x(t)\| < \delta^* := \delta + \|x^0\|_\infty \quad \Longrightarrow \quad \|(\widehat{T}x)(t)\| < \Delta \quad \text{for almost all} \quad t \in [0, \tau].$$

Since F is a Carathéodory function, there exists integrable $\gamma : [0, \tau] \to \mathbb{R}$ such that

$$\|F(t, w)\| \le \gamma(t) \quad \text{for all} \quad (t, w) \in [0, \tau] \times \mathbb{B}_\Delta(0). \tag{A.1}$$

Define $\Gamma : [-h, \tau] \to \mathbb{R}_+$ by

$$\Gamma(t) := \begin{cases} 0, & t \in [-h, 0), \\ \int_0^t \gamma(s)ds, & t \in [0, \tau], \end{cases}$$

and let $0 < \beta < \tau$ be such that $\Gamma(\beta) < \delta$.

Next, we construct a sequence $\{x_n\}_{n\in\mathbb{N}}$ of continuous functions $[-h, \beta] \to \mathbb{R}^N$ as follows. Let $n \in \mathbb{N}$. For $i = 1, \dots, n$, define $x_n^i : [-h, i\beta/n] \to \mathbb{R}^N$ by the recursive formula:

$$i = 1: \quad x_n^1(t) := \begin{cases} x^0(t), & t \in [-h, 0], \\ x^0(0), & t \in (0, \beta/n], \end{cases}$$

$$i > 1: \quad x_n^i(t) := \begin{cases} x_n^{i-1}(t), & t \in [-h, (i-1)\beta/n], \\ x^0(0) + \int_0^{t-(\beta/n)} F(s, (\widehat{T}x_n^{i-1})(s))ds, & t \in ((i-1)\beta/n, i\beta/n]. \end{cases}$$

Observe that if $i \in \{1, \dots, n-1\}$ and $\|x_n^i(t)\| < \delta^*$ for all $t \in [-h, (i\beta)/n]$, then (a) $\|x_n^{i+1}(t)\| < \delta^*$ for all $t \in [-h, (i\beta)/n]$, and (b) $\|(\widehat{T}x_n^i)(t)\| < \Delta$ for all $t \in [0, (i\beta)/n]$, which, in turn, implies for all $t \in (i\beta/n, (i+1)\beta/n]$

$$\|x_n^{i+1}(t) - x^0(0)\| \le \int_0^{t-\beta/n} \|F(s, (\widehat{T}x_n^i)(s))\|ds \le \int_0^{t-\beta/n} \gamma(s)ds = \Gamma(t - \beta/n) < \delta.$$

Noting that $\|x_n^1(t)\| \le \|x^0\|_\infty < \delta^*$ for all $t \in [-h, \beta/n]$, we may now infer (by induction on i) that

$$\|x_n^i(t)\| < \delta^* \quad \text{for all } \ i \in \{1, \dots, n\}, \quad t \in [-h, i\beta/n].$$

For notational convenience, we write $x_n := x_n^n$. By causality of $\widehat{T}$, the sequence $\{x_n\}_{n\in\mathbb{N}}$ so constructed has the property that, for each $n \in \mathbb{N}$,

$$x_n(t) = \begin{cases} x^0(t), & t \in [-h, 0], \\ x^0(0), & t \in (0, \beta/n], \\ x^0(0) + \int_0^{t-(\beta/n)} F(s, (\widehat{T}x_n)(s))ds, & t \in (\beta/n, \beta]. \end{cases} \tag{A.2}$$

Moreover, for all $n \in \mathbb{N}$, $\|x_n(t)\| < \delta^*$ for all $t \in [-h, \beta]$, and so the sequence $\{x_n\}_{n\in\mathbb{N}}$ is uniformly bounded.

Next we prove that the sequence $\{x_n\}_{n\in\mathbb{N}}$ is equicontinuous. Let $\epsilon > 0$. On the closed interval $[0, \beta]$, Γ is uniformly continuous, and so there exists some $\bar{\delta} > 0$ such that

$$t, s \in [0, \beta] \quad \text{with} \quad |t - s| < \bar{\delta} \implies |\Gamma(t) - \Gamma(s)| < \epsilon. \tag{A.3}$$

Let $n \in \mathbb{N}$, $s, t \in [0, \beta]$ with $|t - s| < \bar{\delta}$. Without loss of generality, we assume that $s \le t$. We consider three exhaustive cases.

First, if $0 \le s \le t \le \beta/n$, then $\|x_n(t) - x_n(s)\| = 0$. Second, if $0 < s \le \beta/n \le t \le \beta$, then $t - \beta/n < \bar{\delta}$, and so

$$\|x_n(t) - x_n(s)\| = \|x_n(t) - x^0(0)\| \le \Gamma(t - \beta/n) < \epsilon.$$

Third, if $\beta/n \le s \le t \le \beta$, then

$$\|x_n(t) - x_n(s)\| \le |\Gamma(t - \beta/n) - \Gamma(s - \beta/n)| < \epsilon.$$

Recalling that $x_n|_{[-h,0]} = x^0$ for all n, we conclude that the sequence $\{x_n\}_{n\in\mathbb{N}}$ is equicontinuous. By the Arzelà–Ascoli theorem and extracting a subsequence if necessary, we may assume that the sequence $\{x_n\}_{n\in\mathbb{N}}$ converges uniformly on $[-h, \beta]$ to a continuous limit which we denote by x. Clearly $x|_{[-h,0]} = x^0$.

By property 2(b) of Definition 2.1, $\lim_{n\to\infty}(\widehat{T}x_n)(t) = (\widehat{T}x)(t)$ for almost all $t \in [0, \beta]$ and so, by the continuity of the function $F(t, \cdot)$,

$$\lim_{n\to\infty} F(t, (\widehat{T}x_n)(t)) = F(t, (\widehat{T}x)(t)) \quad \text{for a.a.} \quad t \in [0, \beta].$$

Noting that $\|(\widehat{T}x_n)(s)\| < \Delta$ for all $s \in [0, \beta]$, and also invoking (A.1), we next have $\|F(s, (\widehat{T}x)(s))\| \le \gamma(s)$ for all $s \in [0, \beta]$ and all $n \in \mathbb{N}$. Therefore,

$$\lim_{n\to\infty} \int_{t-\beta/n}^{t} F(s, (\widehat{T}x_n)(s))ds = 0 \quad \text{for all } t \in (0, \beta] \tag{A.4}$$

and, by the Lebesgue dominated convergence theorem,

$$\lim_{n\to\infty} \int_0^t F(s, (\widehat{T}x_n)(s))ds = \int_0^t F(s, (\widehat{T}x)(s))ds \quad \text{for all } t \in [0, \beta]. \tag{A.5}$$

By (A.2), (A.4), and (A.5), it follows that

$$x(t) = \begin{cases} x^0(t), & t \in [-h, 0], \\ x^0(0) + \int_0^t F(s, (\widehat{T}x)(s))ds, & t \in (0, \beta], \end{cases}$$

and so x is a solution of the initial-value problem.

(ii) Let $x : [-h, \omega) \to \mathbb{R}^N$ be a solution of (2.2). Define

$$\mathcal{A} := \left\{(\rho, \xi) |\ \omega \le \rho \le \infty,\ \xi : [-h, \rho) \to \mathbb{R}^N \text{ is a solution of (2.2) with } \xi|_{[-h,\omega)} = x\right\}.$$

On this nonempty set define a partial order $\preceq$ by

$$(\rho_1, \xi_1) \preceq (\rho_2, \xi_2) \quad \Longleftrightarrow \quad \rho_1 \le \rho_2 \text{ and } \xi_1(t) = \xi_2(t) \text{ for all } t \in [-h, \rho_1).$$

Let $\mathcal{O}$ be a totally ordered subset of $\mathcal{A}$. Let $P := \sup\{\rho | (\rho, \xi) \in \mathcal{O}\}$ and let $\Xi : [-h, P) \to \mathbb{R}^M$ be defined by the property that, for every $(\rho, \xi) \in \mathcal{O}$, $\Xi|_{[0,\rho)} = \xi$. Then (P, Ξ) is in $\mathcal{A}$ and is an upper bound for $\mathcal{O}$. By Zorn's lemma, it follows that $\mathcal{A}$ contains at least one maximal element.

(iii) Assume that $x \in C([-h, \omega); \mathbb{R}^N)$ is a bounded maximal solution of (2.2) and that $F \in L^\infty_{\text{loc}}([-h, \infty) \times \mathbb{R}^K; \mathbb{R}^N)$. Seeking a contradiction, suppose $\omega < \infty$. By the boundedness of x, together with property 1 of Definition 2.1 of $\widehat{T}$, it follows that $\dot{x}(\cdot)$ is essentially bounded. Therefore, x is uniformly continuous and so extends to a continuous function $x : [-h, \omega] \to \mathbb{R}^N$. Now consider the initial-value problem

$$\dot{v}(t) = S_\omega F\big(t, (\widehat{T} S_{-\omega} v)(t)\big), \qquad v|_{[-(h+\omega),0]} = S_\omega x. \tag{A.6}$$

By (2.1) and the above existence result, the initial-value problem (A.6) has a solution $\tilde{v} : [-(h + \omega), \tau) \to \mathbb{R}^N$, $\tau > 0$. It follows that $\tilde{x} = S_{-\omega}\tilde{v} : [-h, \omega + \tau) \to \mathbb{R}^N$ is a solution of the original initial-value problem (2.2) and is a proper right extension of the solution x. This contradicts the maximality of x. Therefore, $\omega = \infty$. □

REFERENCES

[1] F. Allgöwer, J. Ashman, and A. Ilchmann, *High-gain adaptive λ-tracking for nonlinear systems*, Automatica J. IFAC, 33 (1997), pp. 881–888.

[2] I. Barbălat, *Systèmes d'équations différentielles d'oscillations non linéaires,* Rev. Math. Pures Appl., 4 (1959), pp. 267–270.

[3] F. Blanchini and E.P. Ryan, *A Razumikhin-type result for functional differential equations with application to adaptive control*, Automatica J. IFAC, 36 (1999), pp. 809–818.

[4] G. Gripenberg, S.-O. Londen, and O. Staffans, *Volterra Integral and Functional Equations*, Cambridge University Press, Cambridge, UK, 1990.

[5] A. Ilchmann, *Non-Identifier-Based High-Gain Adaptive Control,* Springer-Verlag, London, 1993.

[6] A. Ilchmann, *Adaptive λ-tracking for polynomial minimum phase systems*, Dyn. Stab. Syst., 13 (1998), pp. 341–371.

[7] A. Ilchmann and E.P. Ryan, *Universal λ-tracking for nonlinearly perturbed systems in the presence of noise*, Automatica J. IFAC, 30 (1994), pp. 337–346.

[8] A. Ilchmann, M.F. Weirig and I.M.Y. Mareels, *Modelling of biochemical processes and adaptive control*, in Proceedings of 4th IFAC Nonlinear Control Systems Design Symposium, NOLCOS '98, H. Huijberts, H. Nijmeijer, A.J. van der Schaft, and J.M.A. Scherpen, eds., Pergamon, Amsterdam, 1998, pp. 465-470.

[9] A. Ilchmann and M.F. Weirig, *Modelling of general biotechnological processes*, Math. Comput. Modelling Dyn. Syst., 5 (1999), pp. 152–178.

[10] M.A. Krasnosel'skiĭ and A.V. Pokrovskiĭ, *Systems with Hysteresis,* Springer-Verlag, Berlin, 1989.

[11] H. Logemann and A.D. Mawby, *Low-gain integral control of infinite dimensional regular linear systems subject to input hysteresis*, in Advances in Mathematical Systems Theory, F. Colonius, U. Helmke, D. Prätzel-Wolters, and F. Wirth, eds., Birkhäuser Verlag, Basel, Switzerland, 2000, pp. 255–293.

[12] J.W. Macki, P. Nistri, and P. Zecca, *Mathematical models for hysteresis*, SIAM Rev., 35 (1993), pp. 94–123.

[13] I.M.Y. Mareels, *A simple selftuning controller for stably invertible systems*, Systems Control Lett., 4 (1984), pp. 5–16.

[14] B. Mårtensson, *The order of any stabilizing regulator is sufficient a priori information for adaptive stabilization*, Systems Control Lett., 6 (1985), pp. 87–91.

[15] A.S. Morse, *Recent problems in parameter adaptive control*, in Outils et Modèles Mathématiques pour l'Automatique, l'Analyse de Systèmes et le Traitment du Signal, Vol. 3, I.D. Landau, ed., Editions du CNRS, Paris, France, 1983, pp. 733–740.

[16] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations,* Springer-Verlag, New York, 1983.

[17] E.P. Ryan, *A nonlinear universal servomechanism*, IEEE Trans. Automat. Control, 39 (1994), pp. 753–761.

[18] E.P. Ryan, *An integral invariance principle for differential inclusions with applications in adaptive control*, SIAM J. Control Optim., 36 (1998), pp. 960–980.

[19] E.P. Ryan and C.J. Sangwin, *Controlled functional differential equations and adaptive stabilization*, Internat. J. Control, 74 (2001), pp. 77–90.

[20] E.D. Sontag, *Smooth stabilization implies coprime factorization*, IEEE Trans. Automat. Control, 34 (1989), pp. 435–443.

[21] C. Sparrow, *The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors*, Springer-Verlag, New York, 1982.

[22] G. Weiss, *Transfer functions of regular linear systems, Part* 1*: Characterization of regularity*, Trans. Amer. Math. Soc., 342 (1994), pp. 827–854.

[23] J.C. Willems and C.I. Byrnes, *Global adaptive stabilization in the absence of information on the sign of the high frequency gain*, in Analysis and Optimization of Systems, Part I, Lecture Notes in Control and Inform. Sci. 62, Springer-Verlag, Berlin, 1984, pp. 49–57.

[24] X. Ye, *Universal λ-tracking for nonlinearly perturbed systems without restrictions on the relative degree*, Automatica J. IFAC, 35 (1999), pp. 109–119.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 6, pp. 1765–1790

OPTIMAL CONSUMPTION AND PORTFOLIO WITH BOTH FIXED AND PROPORTIONAL TRANSACTION COSTS*

BERNT ØKSENDAL† AND AGNÈS SULEM‡

Abstract. We consider a market model with one risk-free and one risky asset, in which the dynamics of the risky asset are governed by a geometric Brownian motion. In this market we consider an investor who consumes from the bank account and who has the opportunity at any time to transfer funds between the two assets. We suppose that these transfers involve a fixed transaction cost $k > 0$, independent of the size of the transaction, plus a cost proportional to the size of the transaction.

The objective is to maximize the cumulative expected utility of consumption over a planning horizon. We formulate this problem as a combined stochastic control/impulse control problem, which in turn leads to a (nonlinear) quasi-variational Hamilton–Jacobi–Bellman inequality (QVHJBI). We prove that the value function is the unique viscosity solution of this QVHJBI. Finally, numerical results are presented.

Key words. portfolio selection, transaction cost, impulse control, quasi-variational inequalities, viscosity solutions

AMS subject classifications. Primary, 93E20, 91B28; Secondary, 60H30, 49L25, 35R45

PII. S0363012900376013

1. Introduction.

Let $(\Omega, \mathcal{F}, P)$ be a probability space with a given filtration $\{\mathcal{F}_t\}_{t\geq 0}$. We denote by $X(t)$ the amount of money the investor has in the bank at time t and by $Y(t)$ the amount of money invested in the risky asset at time t. We assume that in the absence of consumption and transactions the process $X(t)$ grows deterministically at exponential rate r, while $Y(t)$ is a geometric Brownian motion; i.e.,

$$dX(t) = rX(t)dt, \qquad X(0) = x, \tag{1.1}$$

$$dY(t) = \alpha Y(t)dt + \sigma Y(t)dW(t), \qquad Y(0) = y, \tag{1.2}$$

where $W(t)$ is one-dimensional $\mathcal{F}_t$-Brownian motion and $\alpha > r > 0$ and $\sigma \neq 0$ are constants.

Suppose that at any time t the investor is free to choose a *consumption rate* $c(t) \geq 0$. This consumption is automatically drawn from the bank account holding with no extra costs. Moreover, at any time the investor can decide to transfer money from the bank account to the stock and conversely. Assume that a purchase of size ℓ of stocks incurs a transaction cost consisting of a sum of a fixed cost $k > 0$ (independent of the size of the transaction) plus a cost $\lambda\ell$ proportional to the transaction ($\lambda \geq 0$). These costs are drawn from the bank account. Similarly a sale of size m of stocks incurs the fixed cost $K > 0$ plus the proportional cost μm ($\mu \geq 0$). For simplicity we will assume that $K = k$ and $\mu = \lambda$. In this context the investor will only change his

*Received by the editors July 28, 2000; accepted for publication (in revised form) July 3, 2001; published electronically February 14, 2002. This work was partially supported by the French-Norwegian cooperation project Aur 99–050.

http://www.siam.org/journals/sicon/40-6/37601.html

†Department of Mathematics, University of Oslo, P. O. Box 1053 Blindern, N–0316 Oslo, Norway (oksendal@math.uio.no) and Norwegian School of Economics and Business Administration, Helleveien 30, N–5045 Bergen, Norway.

‡INRIA, Domaine de Voluceau-Rocquencourt B.P. 105, F–78153 Le Chesnay Cedex, France (Agnes.Sulem@inria.fr).

portfolio finitely many times in any finite time interval. The control of the investor will consist of a combination of a regular *stochastic control* $c(t)$ and an *impulse control* $v = (\tau_1, \tau_2, \ldots; \xi_1, \xi_2, \ldots)$. Here $0 \leq \tau_1 < \tau_2 < \cdots$ are $\mathcal{F}_t$-stopping times giving the times when the investor decides to change his portfolio, and $\{\xi_j \in \mathbf{R};\ j = 1, 2, \ldots\}$ are $\mathcal{F}_{\tau_j}$-measurable random variables giving the sizes of the transactions at these times. We assume that

$$c(t) \text{ is } \mathcal{F}_t\text{-adapted}, \quad c(t, \omega) \geq 0, \quad \text{and } \lim_{j \to \infty} \tau_j = \infty \quad \text{a.s.} \tag{1.3}$$

(possibly $\tau_n = \infty$ a.s. for some $n < \infty$).

If such a control $w = (c, v)$ is applied to the system $(X(t), Y(t))$, it gets the form

$$dX(t) = (rX(t) - c(t))dt, \qquad \tau_i \leq t < \tau_{i+1}, \tag{1.4}$$

$$dY(t) = \alpha Y(t)dt + \sigma Y(t)dW(t), \qquad \tau_i \leq t < \tau_{i+1}, \tag{1.5}$$

$$X(\tau_{i+1}) = X(\tau_{i+1}^-) - k - \xi_{i+1} - \lambda|\xi_{i+1}|, \tag{1.6}$$

$$Y(\tau_{i+1}) = Y(\tau_{i+1}^-) + \xi_{i+1}. \tag{1.7}$$

Thus a positive value of ξ_{i+1} indicates that money is being taken from the bank account at time τ_{i+1} to buy stocks, and conversely if ξ_{i+1} is negative.

If our agent has the amounts x in the bank account and y in stocks, his *net wealth* is given by

$$H(x, y) = \max\{x + y - \lambda|y| - k, \min\{x, y\}\}. \tag{1.8}$$

Therefore it is natural to define the *solvency region* $\mathcal{S}$ by

$$\mathcal{S} = \{(x, y) \in \mathbf{R}^2; H(x, y) \geq 0\}, \tag{1.9}$$

and we set

$$\widetilde{\mathcal{S}} = \mathbf{R}^+ \times \mathcal{S}. \tag{1.10}$$

Define the line segments ℓ_1, ℓ_2 by

$$\ell_1 = \{(x, y);\ x + (1 - \lambda)y = k,\ x < 0\}, \tag{1.11}$$

$$\ell_2 = \{(x, y);\ x + (1 + \lambda)y = k,\ y < 0\}, \tag{1.12}$$

and let the points P, Q be the end points of these segments, i.e.,

$$P = \left(0, \frac{k}{1 - \lambda}\right), \qquad Q = (k, 0). \tag{1.13}$$

(See Figure 1.1 and also Remark 2.4.) The investor's objective is to maximize over all combined controls $w = (c, v)$ the expression

$$J^w(s, x, y) = E^{s,x,y}\left[\int_0^\infty e^{-\delta(s+t)} \frac{c^\gamma(t)}{\gamma} dt\right] = e^{-\delta s} E^{x,y}\left[\int_0^\infty e^{-\delta t} \frac{c^\gamma(t)}{\gamma} dt\right], \tag{1.14}$$

where $\delta > 0$, $0 < \gamma < 1$ are constants ($1 - \gamma$ is the relative risk aversion coefficient) and $E^{s,x,y}$ denotes the expectation with respect to the probability law $P^{s,x,y}$ of

$$Z(t) = Z^w(t) := (s + t, X(t), Y(t)), \qquad t \geq 0, \tag{1.15}$$

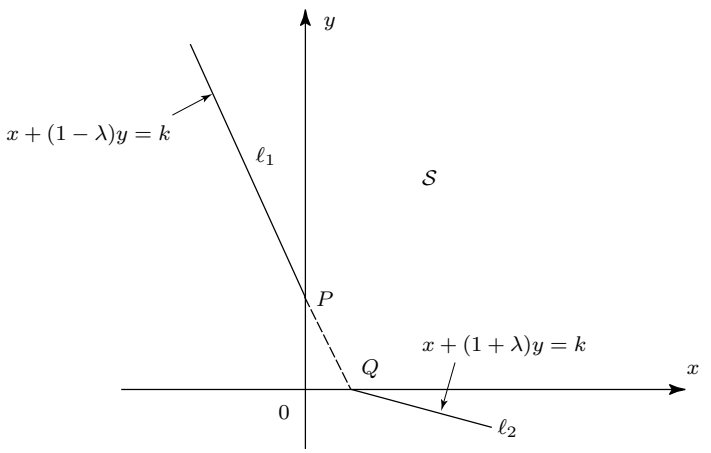


FIG. 1.1. *The solvency region.*

starting at $z = (s, x, y)$.

We seek the value function(s)

$$\Phi(s,x,y) = \sup_{w \in \mathcal{W}} J^w(s,x,y), \qquad \Psi(x,y) = \Phi(0,x,y), \tag{1.16}$$

where $\mathcal{W} = \mathcal{W}(x,y)$ is the set of all *admissible* controls, i.e., all combined controls which satisfy (1.3) and which do not cause the process $Z(t)$ to exit from $\mathcal{S}$. Note that

$$J^w(s,x,y) = e^{-\delta s} J^w(0,x,y) \quad \text{and} \quad \Phi(s,x,y) = e^{-\delta s}\Phi(0,x,y) = e^{-\delta s}\Psi(x,y), \tag{1.17}$$

so the introduction of the s-variable is not really necessary. However, it turns out to be convenient in order to simplify the notation and the arguments in some of the later proofs.

We also seek a corresponding *optimal* control, i.e., a combined control w^* such that

$$\Phi(s,x,y) = J^{w^*}(s,x,y) = e^{-\delta s}\Psi(x,y). \tag{1.18}$$

This problem may be regarded as a generalization of the optimal consumption and portfolio problems studied by Merton [M] and Davis and Norman [DN]. See also Shreve and Soner [SS]. [M] considers the case with no transaction costs ($\lambda = k = 0$), in which case the problem is no longer a combined control problem but a pure stochastic control problem. In this case it is proved in [M] that it is optimal to choose the portfolio such that

$$\frac{Y(t)}{X(t)} = \frac{\pi^*}{1-\pi^*} \qquad \text{for all } t \tag{1.19}$$

(the Merton line), where

$$\pi^* = \frac{\alpha - r}{(1-\gamma)\sigma^2}\ . \tag{1.20}$$

Moreover, the corresponding value function in the Merton case $\lambda = k = 0$ is given by

$$\Psi_0(x,y) = C_1(x+y)^\gamma, \tag{1.21}$$

where

$$C_1 = \frac{1}{\gamma} C_0^{\gamma-1} \qquad \text{with} \qquad C_0 = \frac{1}{1-\gamma}\left[\delta - \gamma r - \frac{\gamma(\alpha - r)^2}{2\sigma^2(1-\gamma)}\right], \tag{1.22}$$

provided that

$$\delta > \gamma\left[r + \frac{(\alpha - r)^2}{2\sigma^2(1-\gamma)}\right]. \tag{1.23}$$

See, e.g., [DN, section 2].

From now on we assume that (1.23) holds.

It is easy to see that

$$\Psi(x, y) \leq \Psi_0(x, y). \tag{1.24}$$

This is also pointed out in Corollary 2.2, to be proved later.

[DN] and [SS] consider the case with proportional transaction costs only ($k = 0$), in which case the problem can be formulated as a singular stochastic control problem. It is proved in [DN] and [SS] that under some conditions there exist two straight lines Γ_1, Γ_2 through the origin, bounding a cone NT, such that it is optimal to make no transactions if $(X(t), Y(t)) \in NT$ and make transactions corresponding to local time at $\partial(NT)$, resulting in reflections back to NT every time $(X(t), Y(t)) \in \partial(NT)$. Depending on the parameters, the Merton line may or may not go between the lines Γ_1, Γ_2 (see Figure 1.2 and the discussion in [AMS, section 7.2]). For an extension of these results to markets with jumps, see [FØS1] and [FØS2].

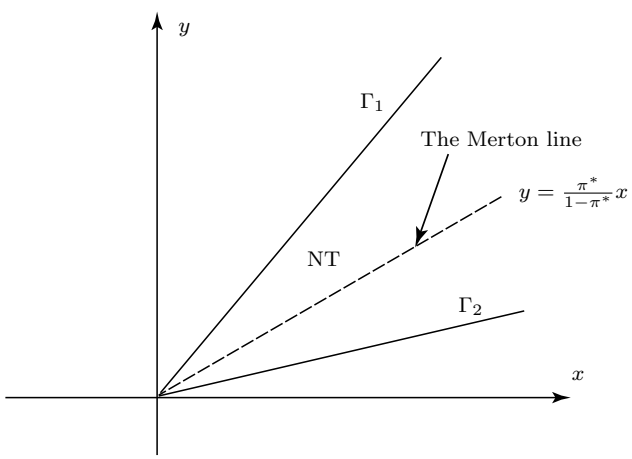


FIG. 1.2. *The no-transaction cone when* $k = 0$.

The first paper to model markets with fixed transaction costs $k > 0$ by impulse control theory seems to be [EH], but they do not consider optimal consumption.

Perhaps the paper which is closest to ours is [K]. Here optimal consumption in markets with fixed transaction costs is considered, but consumption is allowed only at the discrete times of the transactions. This makes it possible to put the problem within the framework of impulse control and quasi-variational inequalities.

In our paper we allow consumption to take place at any time, independent of the (discrete) times chosen for the transactions. As explained above, we model this

as a combined stochastic control and impulse control problem, or a *combined control* problem, for short.

In section 2 we introduce quasi-variational Hamilton–Jacobi–Bellman inequalities (QVHJBI) associated with this combined control problem. We point out that if a function $\psi(x,y)$ satisfies these QVHJBI (and some additional smoothness conditions), then ψ coincides with the value function Ψ, defined by (1.16). (See Theorem 2.1).

In section 3 we prove that the value function Ψ is the unique viscosity solution of the QVHJBI formulated in section 2.

Finally in section 4 we present some numerical estimates for Ψ and the optimal consumption-investment policy $w^* = (c^*, v^*)$.

For other recent papers on impulse control and combined control see, e.g., [BØ], [MØ], [CZ1], [CZ2], and [BP] and the references therein. We refer to [BL] and [ØS] for a comprehensive treatment of the general theory of impulse control and its quasi-variational inequalities.

Remark 1.1. Another natural choice of solvency region would be the set

$$\mathcal{S}_+ := [0,\infty) \times [0,\infty). \tag{1.25}$$

This choice models a situation in which no borrowing or short-selling is allowed. We will mostly use the choice $\mathcal{S}$ given by (1.9) in this paper, but we point out that the proofs carry over to the $\mathcal{S}_+$ case with only minor modifications. (Usually the $\mathcal{S}_+$ case is simpler.)

2. Quasi-variational Hamilton–Jacobi–Bellman inequalities (QVHJBI). Let A^c be the generator of the process $Z^c(t) = (s+t, X^c(t), Y^c(t))$ when there are no transactions; i.e., A^c is the partial differential operator given by

$$(A^c f)(s,x,y) = \frac{\partial f}{\partial s} + (rx - c)\frac{\partial f}{\partial x} + \alpha y \frac{\partial f}{\partial y} + \tfrac{1}{2}\sigma^2 y^2 \frac{\partial^2 f}{\partial y^2} \tag{2.1}$$

for any $f : \mathbf{R}^3 \to \mathbf{R}$ and (s,x,y) such that the derivatives exist. In particular, if $f(s,x,y) = e^{-\delta s} g(x,y)$, then

$$(A^c f)(s,x,y) = e^{-\delta s} L^c g(x,y),$$

where

$$L^c g(x,y) = -\delta g + (rx - c)\frac{\partial g}{\partial x} + \alpha y \frac{\partial g}{\partial y} + \tfrac{1}{2}\sigma^2 y^2 \frac{\partial^2 g}{\partial y^2}. \tag{2.2}$$

For $(x,y) \in \mathcal{S}$ and $\xi \neq 0$ set

$$x' = x'(\xi) = x - k - \xi - \lambda|\xi|, \qquad y' = y'(\xi) = y + \xi. \tag{2.3}$$

We define the *intervention operator* $\mathcal{M}$ by

$$\mathcal{M}h(x,y) = \sup\{h(x',y'); \xi \in \mathbf{R} \setminus \{0\}, (x',y') \in \mathcal{S}\} \tag{2.4}$$

for all locally bounded $h : \mathcal{S} \to \mathbf{R}^+$, $(x,y) \in \mathcal{S}$.

If $(x',y') \notin \mathcal{S}$ for all $\xi \in \mathbf{R} \setminus \{0\}$, we set $\mathcal{M}h(x,y) = 0$. If for all $(x,y) \in \mathcal{S}$ there exists $(x',y') = (x'(\xi), y'(\xi)) \in \mathcal{S}$ such that

$$\mathcal{M}h(x,y) = h(x',y'),$$

then we set

$$\widehat{\xi}(x,y) = \widehat{\xi}_h(x,y) = (x',y'). \tag{2.5}$$

(More precisely, we let $\widehat{\xi}(x,y)$ denote a measurable selection of the map $(x,y) \to (x',y')$.)

If Φ is the value function for our problem, then for each s we can interpret $\mathcal{M}\Phi(s,x,y)$ as the maximal value we can obtain by making an admissible transaction at (s,x,y).

Following [BØ] we call a locally bounded function $h : \widetilde{\mathcal{S}} \to \mathbf{R}^+$ *stochastically* C^2 with respect to Z^c if $(A^c h)(z)$ exists for almost all $z = (s,x,y)$ with respect to *the Green measure* (expected occupation time measure) $G(z_0,\cdot)$, and the generalized Dynkin formula holds for h, i.e.,

$$E^{z_0}[h(Z^c(\tau'))] = E^{z_0}[h(Z^c(\tau))] + E^{z_0}\left[\int_\tau^{\tau'} (A^c h)(Z^c(t))dt\right]$$

for all stopping times τ, τ' such that

$$\tau \le \tau' \le T_R := \inf\{t > 0, |Z^c(t)| \ge R\} \wedge R \qquad \text{for some } R < \infty. \tag{2.6}$$

Recall (see, e.g., [Ø]) that for each $z_0 \in \widetilde{\mathcal{S}}$ the *Green measure* $G(z_0,\cdot)$ of the process Z^c in $\widetilde{\mathcal{S}}$ is defined by

$$G(z_0,H) = E^{z_0}\left[\int_0^\tau \mathcal{X}_H(Z^c(t))dt\right] \qquad \text{for all Borel sets } H \subset \widetilde{\mathcal{S}},$$

where $\tau = \inf\{t > 0; Z^c(t) \notin \widetilde{\mathcal{S}}\}$ and $\mathcal{X}_H(z) = 1$ if $z \in H$, $\mathcal{X}_H(z) = 0$ if $z \notin H$.

If h is a function on $\mathcal{S}$, we define

$$\mathcal{L}h(x,y) = \sup_{c \ge 0}\left\{L^c h(x,y) + \frac{c^\gamma}{\gamma}\right\}, \qquad (x,y) \in \mathcal{S}, \tag{2.7}$$

and

$$\mathcal{L}_0 h(x,y) = L^0 h(0,y) = -\delta h + \alpha y\frac{\partial h}{\partial y} + \tfrac{1}{2}\sigma^2 y^2 \frac{\partial^2 h}{\partial y^2} \tag{2.8}$$

for all points (x,y) where the partial derivatives of h involved in $L^c h$ exist.

We then set (see (1.11)–(1.13) for definitions of ℓ_1, ℓ_2, and P)

$$\mathcal{L}_1 h(x,y) = \begin{cases} \mathcal{L}h(x,y) & \text{for} \quad (x,y) \in \mathcal{S} \setminus (\ell_1 \cup \ell_2) \setminus [0,P], \\ \mathcal{L}_0 h(x,y) & \text{for} \quad (x,y) \in [0,P]. \end{cases} \tag{2.9}$$

Note that at $[0,P]$ the only admissible consumption is $c = 0$.

By adapting Theorem 3.1 in [BØ] to our situation, we get the following sufficient QVHJBI.

THEOREM 2.1. *Let $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ be as defined in* (1.9) *and put $U = \mathcal{S} \setminus (\ell_1 \cup \ell_2)$, $\widetilde{U} = [0,\infty) \times U$.*

(i) *Suppose we can find a locally bounded function* $\psi : \mathcal{S} \to \mathbf{R}^+$ *such that* $\psi \in C^1(U)$ *and*

$$\phi(s,x,y) := e^{-\delta s}\psi(x,y) \quad \text{is stochastically } C^2 \text{ with respect to } Z^c(t) \tag{2.10}$$
$$\text{for all Markov controls } c = c(x,y);$$

$$\mathcal{L}_1\psi \le 0 \qquad \text{a.e. with respect to } G(z_0,\cdot) \text{ on } \widetilde{U} \text{ for all } z_0 \in \widetilde{U}; \tag{2.11}$$

$$\psi(x,y) \ge \mathcal{M}\psi(x,y) \qquad \text{for all } (x,y) \in U. \tag{2.12}$$

Then

$$\psi(x,y) \ge \Psi(x,y) \qquad \text{for all } (s,x,y) \in \widetilde{U}.$$

(ii) *Define the continuation region*

$$D = \{(x,y) \in U; \psi(x,y) > \mathcal{M}\psi(x,y)\}.$$

Suppose

$$\mathcal{L}_1\psi(x,y) = 0 \qquad \text{on } D \tag{2.13}$$

and that $\widehat{\xi}(x,y) = \widehat{\xi}_\psi(x,y)$ *(defined in* (2.5)*) exists for all* $(x,y) \in \mathcal{S}$. *Define*

$$c^*(x,y) = \begin{cases} (\frac{\partial\psi}{\partial x})^{\frac{1}{\gamma-1}} & \text{for} \quad (x,y) \in U \setminus [0,P], \\ 0 & \text{for} \quad (x,y) \in [0,P], \end{cases}$$

and define the impulse control

$$v^* := (\tau_1^*, \tau_2^*, \ldots; \xi_1^*, \xi_2^*, \ldots)$$

as follows.

Put $\tau_0^* = 0$ *and inductively*

$$\tau_{k+1}^* = \inf\{t > \tau_k^*; (X^{(k)}(t), Y^{(k)}(t)) \notin D\}, \tag{2.14}$$

$$\xi_{k+1}^* = \widehat{\xi}(X^{(k)}(\tau_{k+1}^{*-}), Y^{(k)}(\tau_{k+1}^{*-})), \tag{2.15}$$

where $\widehat{\xi}$ *is as defined in* (2.5) *and* $(X^{(k)}, Y^{(k)})$ *is the process obtained by applying the combined control*

$$w^{(k)} := (c^*, (\tau_1^*, \ldots, \tau_k^*; \xi_1^*, \ldots, \xi_k^*)), \qquad k = 1, 2, \ldots.$$

Suppose $w^* := (c^*, v^*) \in \mathcal{W}$ *and that*

$$e^{-\delta t}\psi(X^{(w^*)}(t), Y^{(w^*)}(t)) \to 0 \qquad \text{as } t \to \infty \text{ a.s.} \tag{2.16}$$

and that the family

$$\{e^{-\delta\tau}\psi(X^{(w^*)}(\tau), Y^{(w^*)}(\tau));\ \tau \ \text{ stopping time}\} \tag{2.17}$$

is uniformly integrable. Then

$$\psi(x,y) = \Psi(x,y) \tag{2.18}$$

and w^* is optimal.

Proof. This follows by the proof of Theorem 3.1 in [BØ] with only minor modifications. Note that the Hamilton–Jacobi–Bellman inequality (HJBI) (3.7) in [BØ] has the following form in our case, if $(x,y) \in U \setminus [0,P]$:

$$\mathcal{L}\psi(x,y) = \sup_{c\geq 0}\left\{ -\delta\psi + (rx-c)\frac{\partial\psi}{\partial x} + \alpha y\frac{\partial\psi}{\partial y} + \frac{1}{2}\sigma^2 y^2\frac{\partial^2\psi}{\partial y^2} + \frac{c^\gamma}{\gamma}\right\} \leq 0.$$

This can only hold if $\frac{\partial\psi}{\partial x} > 0$, and then the supremum of this expression is obtained when

$$c = c^* = \left(\frac{\partial\psi}{\partial x}\right)^{\frac{1}{\gamma-1}}. \tag{2.19}$$

If $(x,y) \in [0,P]$, then only the zero consumption $c = c^* = 0$ is admissible, so again by the HJBI we get $L^0\psi(0,y) = 0$. □

We can use this to prove the claim (1.24), as follows.

Corollary 2.2.

(i) *As in* (1.21)–(1.22) *let*

$$\Psi_0(x,y) = C_1(x+y)^\gamma \tag{2.20}$$

be the value function for the Merton problem ($k = \lambda = 0$). *Then*

$$\Psi(x,y) \leq \Psi_0(x,y) \qquad \text{for all } (x,y) \in \mathcal{S}. \tag{2.21}$$

(ii) *Let b be a constant such that*

$$1-\lambda \leq b \leq 1+\lambda. \tag{2.22}$$

Suppose

$$\delta > \gamma\alpha. \tag{2.23}$$

Then there exists $K < \infty$ such that

$$\Psi(x,y) \leq K(x+by)^\gamma \qquad \text{for all } (x,y) \in \mathcal{S}. \tag{2.24}$$

Proof. (i) We verify that $\psi := \Psi_0$ satisfies the conditions of Theorem 2.1(i). First, $\phi(s,x,y) = e^{-\delta s}\psi(x,y)$ is C^2 and therefore trivially stochastically C^2. Hence (2.10) holds. Second, ψ satisfies (2.11), because Ψ_0 satisfies the Hamilton–Jacobi–Bellman (HJB) equation. Third, if we put, as in (2.3),

$$x' = x'(\xi) = x - \xi - k - \lambda|\xi| \qquad \text{and} \qquad y' = y'(\xi) = y + \xi,$$

then $x' + y' \leq x + y$ for all x, y, ξ and therefore

$$\begin{aligned}\mathcal{M}\Psi_0(x,y) &= \sup_{\xi\neq 0}\Psi_0(x',y') = \sup_{\xi\neq 0}\{C_1(x'+y')^\gamma\}\\ &\leq C_1(x+y)^\gamma = \Psi_0(x,y),\end{aligned} \tag{2.25}$$

where C_1 is defined in (1.22). Therefore (2.12) also holds. Hence (2.21) follows.

(ii) We proceed as in (i), except that now we choose $K < \infty$ and define

$$u(x,y) = K(x+by)^\gamma. \tag{2.26}$$

Then we get

$$x' + by' = \begin{cases} x + by - k - \xi(1+\lambda-b) & \text{for } \xi > 0, \\ x + by - k - \xi(1-\lambda-b) & \text{for } \xi < 0. \end{cases}$$

Thus in any case we have, by (2.22),

$$x' + by' \le x + by,$$

and this proves that

$$u(x,y) \ge \mathcal{M}u(x,y).$$

It remains to verify that $\psi := u$ satisfies (2.11). By (2.26) we get

$$\mathcal{L}u(x,y) = (x+by)^{\gamma-2}\left[\left(\frac{1-\gamma}{\gamma}(K\gamma)^{\frac{\gamma}{\gamma-1}} - \delta K\right)(x+by)^2 + K\gamma(rx+\alpha by)(x+by) - \tfrac{1}{2}\sigma^2 K\gamma(1-\gamma)b^2y^2\right].$$

Hence $\mathcal{L}u(x,y) \le 0$ for all $(x,y) \in \mathcal{S}$ if and only if

$$\left[\frac{1-\gamma}{\gamma}(K\gamma)^{\frac{\gamma}{\gamma-1}} - \delta K + K\gamma\alpha\right](x+by)^2 \le \tfrac{1}{2}\sigma^2 K\gamma(1-\gamma)b^2y^2$$

for all $(x,y) \in \mathcal{S}$. This holds if and only if

$$\delta > \gamma\alpha + (1-\gamma)(K\gamma)^{\frac{1}{\gamma-1}}. \tag{2.27}$$

If (2.23) holds, then (2.27) holds for K large enough. Thus (2.24) follows from Theorem 2.1(i). □

Remark 2.3. Corollary 2.2 proves in particular that the value function Ψ is *finite*. Moreover, $\Psi(x,y)$ is *bounded* on every straight line in $\mathcal{S}$ of the form

$$x + by = \text{constant}$$

for every constant $b \in [1-\lambda, 1+\lambda]$.

Remark 2.4 (Some comments on the boundary behavior). Suppose the current position of the investor is a point $(x,y) \in \mathcal{S}$. If we make a transaction of size ξ at that instant, then after the transaction the new position is given by

$$\begin{cases} x' = x - \xi - \lambda|\xi| - k, \\ y' = y + \xi. \end{cases} \tag{2.28}$$

Hence

$$x' + (1-\lambda)y' = x + (1-\lambda)y - k - \lambda(|\xi| + \xi) \tag{2.29}$$

and

$$x' + (1+\lambda)y' = x + (1+\lambda)y - k - \lambda(|\xi| - \xi). \tag{2.30}$$

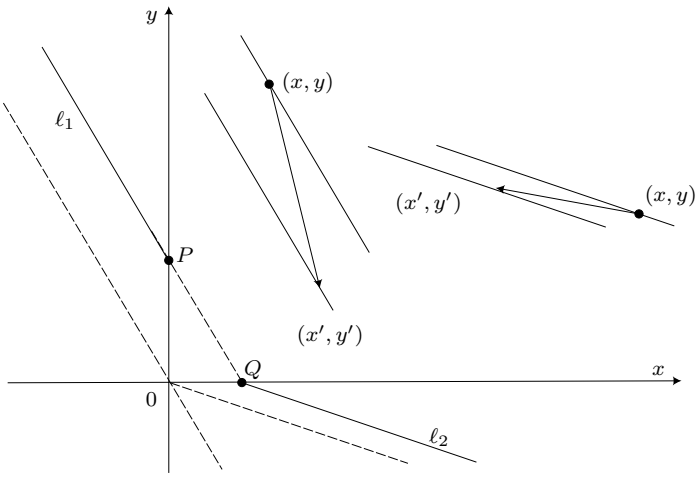


FIG. 2.1. *Examples of transactions (buying and selling).*

In particular, if we *sell* stocks ($\xi < 0$), then $x' + (1-\lambda)y' = x + (1-\lambda)y - k$, so (x, y) will move to a point (x', y') on the line parallel to ℓ_1 lying $\frac{k}{1-\lambda}$ units below the parallel of ℓ_1 through (x, y). See Figure 2.1. Similarly, if we *buy* stocks ($\xi > 0$), then $x' + (1+\lambda)y' = x + (1+\lambda)y - k$, so (x, y) will move to a point (x', y') on the line parallel to ℓ_2 lying $\frac{k}{1+\lambda}$ units below the parallel of ℓ_2 through (x, y).

We now use this to deduce the boundary behavior of the value function Ψ on $\partial\mathcal{S}$.

(a) If $(x, y) \in \ell_1$, then we have to make an immediate transaction to avoid the diffusion $Y(t)$ taking us out of $\mathcal{S}$. By the above we see that the only possibility is to *sell* stocks of such an amount that $(x', y') = (0, 0)$. We conclude that

$$\Psi(x, y) = \mathcal{M}\Psi(x, y) = 0 \qquad \text{for } (x, y) \in \ell_1. \tag{2.31}$$

(b) If $(x, y) \in \ell_2$, we argue similarly: The only admissible action is to *buy* stocks immediately of such an amount that $(x', y') = (0, 0)$. Hence

$$\Psi(x, y) = \mathcal{M}\Psi(x, y) = 0 \qquad \text{for } (x, y) \in \ell_2. \tag{2.32}$$

(c) On the segment $0 < x < k$, $y = 0$, we are not allowed to make any transaction. There is no diffusion and all we can do is consume optimally. Hence the HJB equation indicates that, with $\mathcal{L}$ as in (2.7), we should have

$$\mathcal{L}\Psi = -\delta\Psi + rx\frac{\partial\Psi}{\partial x} + \frac{1-\gamma}{\gamma}\left(\frac{\partial\Psi}{\partial x}\right)^{\frac{\gamma}{\gamma-1}} = 0 \qquad \text{for } x \in (0, k), \tag{2.33}$$

provided that Ψ is smooth enough (see section 4).

(d) On the segment $x = 0$, $0 < y < \frac{k}{1-\lambda}$, we cannot consume because this would bring us outside $\mathcal{S}$. Hence the HJB equation indicates that

$$\mathcal{L}\Psi(0, y) = c^* = 0 \qquad \text{for } 0 < y < \frac{k}{1-\lambda}, \tag{2.34}$$

and hence that

$$\mathcal{L}_0\Psi := -\delta\Psi + \alpha y\frac{\partial\Psi}{\partial y} + \tfrac{1}{2}\sigma^2 y^2\frac{\partial^2\Psi}{\partial y^2} = 0 \qquad \text{for } 0 < y < \frac{k}{1-\lambda}, \tag{2.35}$$

provided that Ψ is smooth enough (see section 4).

Summarizing, we see that the boundary behavior of Ψ on $\partial\mathcal{S}$ can be described by

$$(2.36)\quad \begin{cases} \Psi(x,y) = \mathcal{M}\Psi(x,y) = 0 & \text{for } (x,y) \in \ell_1 \cup \ell_2, \\ \mathcal{L}\Psi(x,y) = 0 & \text{for } 0 \le x \le k,\ y = 0, \text{ i.e., } (x,y) \in [0,Q], \\ \mathcal{L}_0\Psi(x,y) = 0 & \text{for } x = 0,\ 0 \le y \le y_1 = \frac{k}{1-\lambda}, \text{ i.e., } (x,y) \in [0,P]. \end{cases}$$

Note that Ψ is not continuous on ∂S: The points $(0, \frac{k}{1-\lambda})$ and $(k,0)$ are points of discontinuity. However, Ψ is upper semicontinuous.

3. Viscosity solutions. Theorem 2.1 is a *verification theorem*, stating that if we can find a smooth enough function satisfying the required (quasi-) variational inequalities, then we have also found the value function of the problem. It is natural to ask if the converse is also true: Is the value function always a solution of the corresponding (quasi-) variational inequalities? The problem is that the value function need not be smooth enough for these inequalities to be well defined in the strong sense. In fact, the value function is not even continuous at the points P and Q (see (2.36) and below). However, we shall see that the inequalities are satisfied in a weak sense: The value function is a *viscosity solution* of the (quasi-) variational inequalities.

We first recall the following concepts, which will be useful for us.

DEFINITION 3.1. *If C is a topological space and $u\colon C \to \mathbf{R}$ is a function, then the* upper semi-continuous *(usc)* envelope $\overline{u}\colon C \to \mathbf{R}$ *and the* lower semi-continuous *(lsc)* envelope $\underline{u}\colon C \to \mathbf{R}$ *of u are defined by*

$$\overline{u}(x) = \limsup_{\substack{y \to x \\ y \in C}} u(y), \qquad \underline{u}(x) = \liminf_{\substack{y \to x \\ y \in C}} u(y), \quad \textit{respectively.}$$

We let $\mathrm{USC}(C)$ and $\mathrm{LSC}(C)$ denote the set of usc functions and lsc functions on C, respectively.

Note that in general we have

$$\underline{u} \le u \le \overline{u},$$

and that u is usc if and only if $u = \overline{u}$, u is lsc if and only if $u = \underline{u}$. In particular, u is continuous if and only if

$$\underline{u} = u = \overline{u}.$$

We establish some auxiliary results about the operator $\mathcal{M}$, as follows.

LEMMA 3.2.

(i) *If $u : \mathcal{S} \to \mathbf{R}$ is* usc*, then $\mathcal{M}u$ is* usc*.*

(ii) *If $u : \mathcal{S} \to \mathbf{R}$ is continuous, then $\mathcal{M}u$ is continuous.*

Proof. (i) Suppose that $u : \mathcal{S} \to \mathbf{R}$ is usc. For $\zeta = (x,y) \in \mathcal{S}$ define

$$\ell(\zeta) = \ell(x,y) = \{(x'(\xi), y'(\xi)) \in \mathcal{S};\quad \xi \in \mathbf{R} \setminus \{0\}\},$$

where x', y' are as in (2.3). Then $\ell(\zeta)$ is a union of two *closed finite* line segments, and since u is usc there exists $\zeta^* \in \ell(\zeta)$ such that

$$\mathcal{M}u(\zeta) = \sup\{u(\zeta'); \zeta' \in \ell(\zeta)\} = u(\zeta^*).$$

Fix $\zeta_0 \in \mathcal{S}$ and let $\{\zeta_n\}_{n=1}^{\infty}$ be a sequence in $\mathcal{S}$ such that $\zeta_n \to \zeta_0$ as $n \to \infty$. We must show that

$$\mathcal{M}u(\zeta_0) \ge \limsup_{n \to \infty} \mathcal{M}u(\zeta_n).$$

Let $\hat{\zeta}$ be a cluster point of $\{\zeta_n^*\}_{n=1}^{\infty}$, i.e., $\hat{\zeta}$ is the limit of some convergent subsequence $\{\zeta_{n_k}^*\}_{k=1}^{\infty}$ of $\{\zeta_n^*\}_{n=1}^{\infty}$. Since $\zeta_n \to \zeta_0$, we see that $\ell(\zeta_n) \to \ell(\zeta_0)$, in the natural sense. Hence, since $\zeta_{n_k}^* \in \ell(\zeta_{n_k})$ for all k, we conclude that $\hat{\zeta} = \lim_{k\to\infty} \zeta_{n_k}^* \in \ell(\zeta_0)$. Therefore

$$\mathcal{M}u(\zeta_0) \geq u(\hat{\zeta}) \geq \limsup_{n\to\infty} u(\zeta_n^*) = \limsup_{n\to\infty} \mathcal{M}u(\zeta_n).$$

(ii) Suppose that $u : \mathcal{S} \to \mathbf{R}$ is continuous. Fix $\zeta_0 \in \mathcal{S}$ and let $\zeta_n \to \zeta_0$ as in (i). By (i) it suffices to show that

$$(*) \qquad \mathcal{M}u(\zeta_0) \leq \liminf_{n\to\infty} \mathcal{M}u(\zeta_n).$$

Suppose not. Then $u(\zeta_0^*) = \mathcal{M}u(\zeta_0) > \liminf_{n\to\infty} \mathcal{M}u(\zeta_n) + \varepsilon = \liminf_{n\to\infty} u(\zeta_n^*) + \varepsilon$ for some $\varepsilon > 0$.

Since u is continuous, there is a neighborhood G of ζ_0^* such that

$$u(\zeta') \geq \liminf u(\zeta_n^*) + \varepsilon \qquad \text{for all } \zeta' \in G.$$

But if n is big enough we have $\ell(\zeta_n) \cap G \neq \emptyset$, so since ζ_n^* is a maximum point of u on $\ell(\zeta_n)$ we have

$$u(\zeta_n^*) \geq u(\zeta') \qquad \text{for } n \text{ big enough.}$$

This contradiction shows that $(*)$ holds, and the proof is complete. □

LEMMA 3.3.

(i) *Let $u : \mathcal{S} \to \mathbf{R}$. Then $\overline{\mathcal{M}u} \leq \mathcal{M}\bar{u}$.*

(ii) *Let $\psi : \mathcal{S} \to \mathbf{R}$ be such that $\psi \geq \mathcal{M}\psi$. Then $\underline{\psi} \geq \mathcal{M}\underline{\psi}$.*

Proof. (i) Choose $\zeta_0, \zeta_n \in \mathcal{S}$, $n = 1, 2, \ldots$, such that $\zeta_n \to \zeta_0$ and $\mathcal{M}u(\zeta_n) \to \overline{\mathcal{M}u}(\zeta_0)$ as $n \to \infty$. Then by Lemma 3.2(i) applied to the usc function $\bar{u}$,

$$\overline{\mathcal{M}u}(\zeta_0) = \lim_{n\to\infty} \mathcal{M}u(\zeta_n) \leq \limsup_{n\to\infty} \mathcal{M}\bar{u}(\zeta_n) \leq \mathcal{M}\bar{u}(\zeta_0).$$

(ii) Choose $\zeta_0, \zeta_n \in \mathcal{S}$, $n = 1, 2, \ldots$, such that $\zeta_n \to \zeta_0$ and $\mathcal{M}\psi(\zeta_n) \to \underline{\mathcal{M}\psi}(\zeta_0)$ as $n \to \infty$. Then

$$\underline{\psi}(\zeta_0) \geq \underline{\mathcal{M}\psi}(\zeta_0) = \lim_{n\to\infty} \mathcal{M}\psi(\zeta_n) \geq \liminf_{n\to\infty} \mathcal{M}\underline{\psi}(\zeta_n) \geq \mathcal{M}\underline{\psi}(\zeta_0). \qquad \square$$

COROLLARY 3.4. *Suppose $u : \mathcal{S} \to \mathbf{R}$ is* usc *and $u(\zeta_0) > \mathcal{M}u(\zeta_0) + \eta$ for some $\zeta_0 \in \mathcal{S}$, $\eta > 0$. Then $u(\zeta_0) > \overline{\mathcal{M}u}(\zeta_0) + \eta$.*

Proof. $u(\zeta_0) > \mathcal{M}u(\zeta_0) + \eta = \mathcal{M}\bar{u}(\zeta_0) \geq \overline{\mathcal{M}u}(\zeta_0) + \eta$ by Lemma 3.3(i). □

As in (2.7) we let $\mathcal{L}$ be the differential operator

$$(3.1) \qquad \mathcal{L}h(x,y) = \sup_{c\geq 0} \left\{ -\delta h + (rx - c)\frac{\partial h}{\partial x} + \alpha y \frac{\partial h}{\partial y} + \frac{1}{2}\sigma^2 y^2 \frac{\partial^2 h}{\partial y^2} + \frac{c^\gamma}{\gamma} \right\},$$

and as in (2.2) we set

$$(3.2) \qquad \mathcal{M}h(x,y) = \sup_{\xi\neq 0} \{h(x',y'); (x',y') \in \mathcal{S}\} \qquad \text{for } (x,y) \in \mathcal{S},$$

where

$$(3.3) \qquad x' = x - k - \xi - \lambda|\xi|, \qquad y' = y + \xi.$$

The inequalities (2.11), (2.12), and (2.13) of Theorem 2.1 together with the boundary behavior (2.36) can be combined into one equation as follows:

$$F(D^2\Psi(\zeta), D\Psi(\zeta), \Psi, \zeta) = 0 \qquad \text{for all } \zeta = (x, y) \in \mathcal{S}, \tag{3.4}$$

where

$$F\colon \mathbf{R}^{2\times 2} \times \mathbf{R}^2 \times \mathbf{R}^{\mathcal{S}} \times \mathbf{R}^2 \to \mathbf{R}$$

is defined by

$$F(A, p, g, \zeta) = \begin{cases} \max\{\Lambda(A, p, g, \zeta), (\mathcal{M}g - g)(\zeta)\}, & \zeta \in \mathcal{S}^0, \\ \Lambda(A, p, g, \zeta), & \zeta \in [0, Q], \\ \Lambda_0(A, p, g, \zeta), & \zeta \in [0, P], \\ (\mathcal{M}g - g)(\zeta), & \zeta \in \ell_1 \cup \ell_2, \end{cases} \tag{3.5}$$

where

$$\Lambda(A, p, g, \zeta) = -\delta g + r\zeta_1 p_1 + \alpha\zeta_2 p_2 + \tfrac{1}{2}\sigma^2\zeta_2^2 A_{22} + \max_{c\geq 0}\left(-cp_1 + \frac{c^\gamma}{\gamma}\right) \tag{3.6}$$

and

$$\Lambda_0(A, p, g, \zeta) = -\delta g + \alpha\zeta_2 p_2 + \tfrac{1}{2}\sigma^2\zeta_2^2 A_{22}, \qquad A = \left[A_{ij}\right]_{1\leq i,j\leq 2}. \tag{3.7}$$

Note that F is not a local operator: The value of F at (A, p, g, ζ) depends on the value of g on the whole space $\mathcal{S}$.

Also note that

$$\overline{F}(A, p, g, \zeta) = \max\{\Lambda(A, p, g, \zeta), (\mathcal{M}g - g)(\zeta)\} \text{ for } \textit{all } \zeta \in \mathcal{S} \tag{3.8}$$

and that

$$\underline{F}(A, p, g, \zeta) = F(A, p, g, \zeta) \qquad \text{(i.e., } F \text{ is lsc)}. \tag{3.9}$$

Following Barles [B], we now give the definition of the viscosity solution of elliptic equations of type (3.4).

DEFINITION 3.5.

(i) *A function* $u \in \mathrm{USC}(\mathcal{S})$ *is a* viscosity subsolution *of*

$$F(D^2u(\zeta), Du(\zeta), u, \zeta) = 0 \qquad \textit{for all } \zeta = (x, y) \in \mathcal{S} \tag{3.10}$$

if for every function f *which is* C^2 *in a neighbourhood of* $\mathcal{S}$ *and for every point* $\zeta_0 \in \mathcal{S}$ *such that* $f \geq u$ *on* $\mathcal{S}$ *and* $f(\zeta_0) = u(\zeta_0)$ *we have*

$$\overline{F}(D^2 f(\zeta_0), Df(\zeta_0), u, \zeta_0) \geq 0. \tag{3.11}$$

(ii) *A function* $u \in \mathrm{LSC}(\mathcal{S})$ *is a* viscosity supersolution *of* (3.10) *if for every function* f *which is* C^2 *in a neighbourhood of* $\mathcal{S}$ *and for every point* $\zeta_0 \in \mathcal{S}$ *such that* $f \leq u$ *on* $\mathcal{S}$ *and* $f(\zeta_0) = u(\zeta_0)$ *we have*

$$\underline{F}(D^2 f(\zeta_0), Df(\zeta_0), u, \zeta_0) \leq 0. \tag{3.12}$$

(iii) *We say that a function* $u: \mathcal{S} \to \mathbf{R}$ *is a* viscosity solution *of* (3.10) *if* u *is locally bounded and* $\overline{u}$ *is a viscosity subsolution and* $\underline{u}$ *is a viscosity supersolution of* (3.10).

An equivalent definition of viscosity solutions which is useful for proving uniqueness results is the following (see [CIL, section 2]).

DEFINITION 3.6.

(i) *A function* $u \in \mathrm{USC}(\mathcal{S})$ *is a* viscosity subsolution *of* (3.4) *if*

$$\overline{F}(A,p,u,\zeta) \geq 0 \qquad \text{for all } (p,A) \in \bar{J}^{2,+}_{\mathcal{S}} u(\zeta),\ \zeta \in \mathcal{S}. \tag{3.13}$$

(ii) *A function* $u \in \mathrm{LSC}(\mathcal{S})$ *is a* viscosity supersolution *of* (3.4) *if*

$$\underline{F}(A,p,u,\zeta) \leq 0 \qquad \text{for all } (p,A) \in \bar{J}^{2,-}_{\mathcal{S}} u(\zeta),\ \zeta \in \mathcal{S}.$$

Here the second order "superjets" $J^{2,+}_{\mathcal{S}}$, $J^{2,-}_{\mathcal{S}}$ *and their "closures"* $\bar{J}^{2,+}_{\mathcal{S}}$, $\bar{J}^{2,-}_{\mathcal{S}}$ *are defined by*

$$J^{2,+}_{\mathcal{S}} u(\zeta) = \Bigg\{ (p,A) \in \mathbf{R}^2 \times \mathbf{R}^{2\times 2};$$

$$\limsup_{\substack{\eta \to \zeta \\ \eta \in \mathcal{S}}} \big\{ [u(\eta) - u(\zeta) - p \cdot (\eta - \zeta) - \tfrac{1}{2}(\eta-\zeta)^T A (\eta - \zeta)] |\eta - \zeta|^{-2} \big\} \leq 0 \Bigg\} \tag{3.14}$$

(where $(\ )^T$ *denotes matrix transposed),*

$$\begin{aligned} \bar{J}^{2,+}_{\mathcal{S}} u(\zeta) = \{ (p,A) \in \mathbf{R}^2 \times \mathbf{R}^{2\times 2}\ ;\ & \exists (\zeta_n, p_n, A_n) \in \mathcal{S} \times \mathbf{R}^2 \times \mathbf{R}^{2\times 2}, \\ & \text{with } (p_n, A_n) \in J^{2,+}_{\mathcal{S}} u(\zeta_n) \text{ and } (\zeta_n, u(\zeta_n), p_n, A_n) \\ & \to (\zeta, u(\zeta), p, A), \quad \text{when } n \to \infty \}, \end{aligned} \tag{3.15}$$

and

$$J^{2,-}_{\mathcal{S}} u = -J^{2,+}_{\mathcal{S}}(-u), \qquad \bar{J}^{2,-}_{\mathcal{S}} u = -\bar{J}^{2,+}_{\mathcal{S}}(-u). \tag{3.16}$$

We are now ready for the first main result of this section.

THEOREM 3.7. *Suppose that* (2.23) *holds. Then the value function* Ψ *is a viscosity solution of* (3.4).

Proof. We first make some useful observations. Suppose $w = (c, v) \in \mathcal{W}$ is an admissible control with $v = (\tau_1, \tau_2, \ldots; \xi_1, \xi_2, \ldots)$, where $\tau_1 > 0$ a.s. Then by the *Markov property* we have, with J^w as in (1.14),

$$J^w(z) = E^z \Big[\int_0^\tau e^{-\delta(s+t)} \frac{c^\gamma(t)}{\gamma} dt + J^w(Z^{(w)}(\tau)) \Big] \tag{3.17}$$

for all stopping times $\tau \leq \tau_1$.

Note that

$$\Psi(\zeta) \geq \mathcal{M}\Psi(\zeta) \qquad \text{for all } \zeta \in \mathcal{S}. \tag{3.18}$$

To see this, suppose on the contrary that there exists ζ_1 such that

$$\Psi(\zeta_1) < \mathcal{M}\Psi(\zeta_1).$$

This would mean that we could improve the performance at ζ_1 by making a transaction immediately. This contradicts that $\Psi(\zeta_1)$ is the optimal performance value at ζ_1.

Also note that since τ_1 is a stopping time, we know that $\{\omega; \tau_1(\omega) = 0\}$ is $\mathcal{F}_0$-measurable and hence this event has probability either 1 or 0. So we either have

$$\tau_1(\omega) = 0 \quad \text{a.s.} \quad \text{or} \qquad \tau_1(\omega) > 0 \quad \text{a.s.}$$

(A) We prove that $\bar{\Psi}$ is a viscosity subsolution. To this end, let f be a C^2 function in a neighborhood of $\mathcal{S}$ and let $\zeta_0 \in \mathcal{S}$ be such that $f \geq \bar{\Psi}$ on $\mathcal{S}$ and $f(\zeta_0) = \bar{\Psi}(\zeta_0)$. We consider the following two cases separately.

Case 1. $\bar{\Psi}(\zeta_0) \leq \mathcal{M}\bar{\Psi}(\zeta_0)$.
Then by (3.8) $\bar{F}(D^2 f(\zeta_0), Df(\zeta_0), f(\zeta_0), \bar{\Psi}, \zeta_0) \geq (\mathcal{M}\bar{\Psi} - \bar{\Psi})(\zeta_0) = 0$, and hence (3.11) holds at ζ_0 for $u = \bar{\Psi}$.

Case 2. $\bar{\Psi}(\zeta_0) > \mathcal{M}\bar{\Psi}(\zeta_0)$.
It suffices to prove that $\mathcal{L}f(\zeta_0) \geq 0$. We argue by contradiction: *Suppose* $\zeta_0 = (x_0, y_0) \in \mathcal{S}$ *and* $\mathcal{L}f(\zeta_0) < 0$. Then from the definition (3.1) of $\mathcal{L}$ we deduce that $\frac{\partial f}{\partial x}(\zeta_0) > 0$. Hence by continuity, $\frac{\partial f}{\partial x}(\zeta) > 0$ in a neighborhood G of ζ_0. But then, with $\zeta = (x, y)$,

$$\mathcal{L}f(\zeta) = -\delta f(\zeta) + (rx - \widehat{c})\frac{\partial f}{\partial x} + \alpha y \frac{\partial f}{\partial y} + \tfrac{1}{2}\sigma^2 y^2 \frac{\partial^2 f}{\partial y^2} + \frac{\widehat{c}^{\gamma}}{\gamma}$$

with $\widehat{c} = \widehat{c}(\zeta) = (\frac{\partial f}{\partial x})^{\frac{1}{\gamma - 1}}$ for all $\zeta \in G \cap \mathcal{S}$.

Hence $\mathcal{L}f(\zeta)$ is continuous on $G \cap \mathcal{S}$ and so there exists a (bounded) neighborhood G_ρ of ζ_0 such that $G_\rho = \{(x, y); |x - x_0| < \rho \text{ and } |y - y_0| < \rho\}$ for some $\rho > 0$ and

$$\mathcal{L}f(\zeta) < \tfrac{1}{2}\mathcal{L}f(\zeta_0) < 0 \qquad \text{for all } \zeta \in G_\rho \cap \mathcal{S}. \tag{3.19}$$

Now let η be any number such that

$$0 < \eta < (\bar{\Psi} - \mathcal{M}\bar{\Psi})(\zeta_0). \tag{3.20}$$

Since $\bar{\Psi}(\zeta_0) > \mathcal{M}\bar{\Psi}(\zeta_0) + \eta$, we can by Corollary 3.4 find a sequence $\{\zeta_n\}_{n=1}^{\infty} \subset G_\rho \cap \mathcal{S}$ such that $\zeta_n \to \zeta_0$ and $\Psi(\zeta_n) \to \bar{\Psi}(\zeta_0)$ as $n \to \infty$ and

$$\mathcal{M}\Psi(\zeta_n) < \Psi(\zeta_n) - \eta \qquad \text{for all } n \geq 1. \tag{3.21}$$

Choose $\varepsilon \in (0, \eta)$. Since $\bar{\Psi}(\zeta_0) = f(\zeta_0)$, we can choose n_0 such that

$$|\Psi(\zeta_n) - f(\zeta_n)| < \varepsilon \qquad \text{for all } n \geq n_0. \tag{3.22}$$

In the following we fix $n \geq n_0$ and put $\zeta = \zeta_n$.

Let $\widetilde{w} = (\widetilde{c}, \widetilde{v})$ with $\widetilde{v} = (\widetilde{\tau}_1, \widetilde{\tau}_2, \ldots; \widetilde{\xi}_1, \widetilde{\xi}_2, \ldots)$ be an ε-optimal control for ζ, in the sense that

$$\Psi(\zeta) \leq J^{\widetilde{w}}(0, \zeta) + \varepsilon.$$

If $\widetilde{\tau}_1 = 0$ a.s., then $(X^{(\widetilde{w})}, Y^{(\widetilde{w})})$ makes an immediate jump from ζ to some point $\zeta' \in \mathcal{S}$, and hence by (3.17)

$$J^{\widetilde{w}}(0, \zeta) = E^{0,\zeta_0}[J^{\widetilde{w}}(0, \zeta')].$$

But then

$$\Psi(\zeta) \le J^{\tilde{w}}(0,\zeta) + \varepsilon = E^{0,\zeta}[J^{\tilde{w}}(0,\zeta')] + \varepsilon \le E^{0,\zeta}[\Psi(\zeta')] + \varepsilon \le \mathcal{M}\Psi(\zeta) + \varepsilon,$$

which contradicts (3.21). We conclude that $\widetilde{\tau}_1 > 0$ a.s.

Fix $R < \infty$ and define τ to be the stopping time

$$\tau = \tau(\varepsilon) = \widetilde{\tau}_1 \wedge R \wedge \inf\{t > 0; (X^{(\tilde{w})}(t), Y^{(\tilde{w})}(t)) \notin G_\rho\}.$$

Then by the Dynkin formula we have

$$E^{0,\zeta}[e^{-\delta\tau} f(X^{(\tilde{w})}(\tau), Y^{(\tilde{w})}(\tau))] = f(\zeta) + E^{0,\zeta}\bigg[\int_0^\tau e^{-\delta t} L^{\tilde{c}} f(X^{(\tilde{w})}(t), Y^{(\tilde{w})}(t))dt\bigg]$$
$$+ E^{0,\zeta}[e^{-\delta\tau}[f(X^{(\tilde{w})}(\tau), Y^{(\tilde{w})}(\tau)) - f(X^{(\tilde{w})}(\tau^-), Y^{(\tilde{w})}(\tau^-))]]$$

or

$$E^{0,\zeta}[e^{-\delta\tau} f(X^{(\tilde{w})}(\tau^-), Y^{(\tilde{w})}(\tau^-))] = f(\zeta) + E^{0,\zeta}\bigg[\int_0^\tau e^{-\delta t} L^{\tilde{c}} f(X^{(\tilde{w})}(t), Y^{(\tilde{w})}(t))dt\bigg].$$

Combining this with (3.17) we get, since $\Psi \ge \mathcal{M}\Psi$,

$$\begin{aligned}
\Psi(\zeta) &\le J^{(\tilde{w})}(0,\zeta) + \varepsilon \\
&= E^{0,\zeta}\bigg[\int_0^\tau e^{-\delta t}\frac{\tilde{c}^\gamma(t)}{\gamma}dt + J^{\tilde{w}}(Z^{(\tilde{w})}(\tau))\bigg] + \varepsilon \\
&\le E^{0,\zeta}\bigg[\int_0^\tau e^{-\delta t}\frac{\tilde{c}^\gamma(t)}{\gamma}dt + e^{-\delta\tau}\Psi(X^{(\tilde{w})}(\tau), Y^{(\tilde{w})}(\tau))\bigg] + \varepsilon \\
&\le E^{0,\zeta}\bigg[\int_0^\tau e^{-\delta t}\frac{\tilde{c}^\gamma(t)}{\gamma}dt + e^{-\delta\tau}\{\Psi(X^{(\tilde{w})}(\tau^-), Y^{(\tilde{w})}(\tau^-)) \cdot \chi_{\tau<\tilde{\tau}_1} \\
&\qquad + \mathcal{M}\Psi(X^{(\tilde{w})}(\tau^-), Y^{(\tilde{w})}(\tau^-)) \cdot \chi_{\tau=\tilde{\tau}_1}\}\bigg] + \varepsilon \\
&\le E^{0,\zeta}\bigg[\int_0^\tau e^{-\delta t}\frac{\tilde{c}^\gamma(t)}{\gamma}dt + e^{-\delta\tau}\Psi(X^{(\tilde{w})}(\tau^-), Y^{(\tilde{w})}(\tau^-))\bigg] + \varepsilon \\
&\le E^{0,\zeta}\bigg[\int_0^\tau e^{-\delta t}\frac{\tilde{c}^\gamma(t)}{\gamma}dt + e^{-\delta\tau} f(X^{(\tilde{w})}(\tau^-), Y^{(\tilde{w})}(\tau^-))\bigg] + \varepsilon \\
&= f(\zeta) + E^{0,\zeta}\bigg[\int_0^\tau e^{-\delta t}\bigg(L^{\tilde{c}} f(X^{(\tilde{w})}(t), Y^{(\tilde{w})}(t)) + \frac{\tilde{c}^\gamma(t)}{\gamma}\bigg)dt\bigg] + \varepsilon \\
&\le \Psi(\zeta) + E^{0,\zeta}\bigg[\int_0^\tau e^{-\delta t}\mathcal{L} f(X^{(\tilde{w})}(t), Y^{(\tilde{w})}(t))dt\bigg] + 2\varepsilon.
\end{aligned}$$

We conclude from this that

$$E^{0,\zeta}\Big[\int_0^\tau e^{-\delta t}\mathcal{L}f(X^{(\tilde{w})}(t),Y^{(\tilde{w})}(t))dt\Big] \geq -2\varepsilon. \tag{3.23}$$

On the other hand, from (3.19) we deduce that

$$E^{0,\zeta}\Big[\int_0^\tau e^{-\delta t}\mathcal{L}f(X^{(\tilde{w})}(t),Y^{(\tilde{w})}(t))dt\Big] \leq \frac{1}{2\delta}\mathcal{L}f(\zeta)(1-E^{0,\zeta}[e^{-\delta\tau}]). \tag{3.24}$$

We claim that

$$E^{0,\zeta_n}[e^{-\delta\tau(\varepsilon)}] \quad \text{is bounded away from 1} \tag{3.25}$$

when $n\to\infty$ and $\varepsilon\to 0$.

If this claim is proved, then we see that (3.23) contradicts (3.24) if ε is small enough. This contradiction proves that $\mathcal{L}f(\zeta_0)\leq 0$ and hence (3.11) holds. Therefore, to complete the proof we must verify the claim (3.25).

First note that for $t<\tau$ we have by (1.4)

$$X^{(\tilde{w})}(t) = X(0)e^{rt} - e^{rt}\int_0^t e^{-rs}\,\tilde{c}(s)ds \geq X(0)-\rho,$$

and hence, with some constant $C_2<\infty$,

$$\begin{aligned}
&\int_0^\tau e^{-\delta t}\,\frac{\tilde{c}^\gamma(t)}{\gamma}dt \leq \frac{1}{\gamma}\Big[\int_0^\tau e^{-rt}\tilde{c}(t)dt\Big]^\gamma\Big[\int_0^\tau e^{\frac{r\gamma-\delta}{1-\gamma}t}dt\Big]^{1-\gamma}\\
&\leq C_2\left(X(0)(1-e^{-r\tau})+\rho e^{-r\tau}\right)^\gamma, \qquad \text{since } r\gamma-\delta<0 \text{ by (1.2) and (2.23).}
\end{aligned}$$

Combining this with (3.17), we get

$$\begin{aligned}
&\Psi(\zeta)-\varepsilon \leq J^{(\tilde{w})}(0,\zeta)\\
&\leq E^{0,\zeta}\Big[\int_0^\tau e^{-\delta t}\,\frac{\tilde{c}^\gamma(t)}{\gamma}dt + e^{-\delta\tau}\Psi(X^{(\tilde{w})}(\tau),Y^{(\tilde{w})}(\tau))\Big]\\
&\leq E^{0,\zeta}[C_2(x-(x-\rho)e^{-r\tau})^\gamma] + E^{0,\zeta}[e^{-\delta\tau}\Psi(X^{(\tilde{w})}(\tau^-),Y^{(\tilde{w})}(\tau^-))\cdot\mathcal{X}_{\tilde{\tau}_1>\tau}]\\
&\quad + E^{0,\zeta}[e^{-\delta\tau}\{\Psi(X^{(\tilde{w})}(\tau),Y^{(\tilde{w})}(\tau)) - \Psi(X^{(\tilde{w})}(\tau^-),Y^{(\tilde{w})}(\tau^-))\}\cdot\mathcal{X}_{\tilde{\tau}_1\leq\tau}]\\
&\leq E^{0,\zeta}[C_2(x-(x-\rho)e^{-r\tau})^\gamma] + E^{0,\zeta}[e^{-\delta\tau}\Psi(X^{(\tilde{w})}(\tau^-),Y^{(\tilde{w})}(\tau^-))\cdot\mathcal{X}_{\tilde{\tau}_1>\tau}]\\
&\quad + E^{0,\zeta}[e^{-\delta\tau}\mathcal{M}\Psi(X^{(\tilde{w})}(\tau^-),Y^{(\tilde{w})}(\tau^-))\cdot\mathcal{X}_{\tilde{\tau}_1\leq\tau}]\\
&\leq E^{0,\zeta}[C_2(x-(x-\rho)e^{-r\tau})^\gamma] + E^{0,\zeta}[e^{-\delta\tau}\mathcal{X}_{\tilde{\tau}_1>\tau}]\cdot\sup\{\Psi(\tilde{\zeta});\tilde{\zeta}\in G_\rho\}\\
&\quad + E^{0,\zeta}[e^{-\delta\tau}\mathcal{X}_{\tilde{\tau}_1\leq\tau}]\cdot\sup\{\mathcal{M}\Psi(\tilde{\zeta});\tilde{\zeta}\in G_\rho\}.
\end{aligned} \tag{3.26}$$

Now if there exists a sequence $\varepsilon_k\to 0$ and a subsequence $\{\zeta_{n_k}\}$ of $\{\zeta_n\}$ such that

$$E^{0,\zeta_{n_k}}[e^{-\delta\tau(\varepsilon_k)}]\to 1 \qquad \text{when } k\to\infty,$$

then

$$E^{0,\zeta_{n_k}}[e^{-\delta\tau(\varepsilon_k)}\mathcal{X}_{\tilde{\tau}_1>\tau}] \to 0 \qquad \text{when } \ k \to \infty,$$

so by choosing $\zeta = \zeta_{n_k}$, $\tau = \tau(\varepsilon_k)$ in (3.26) and letting $k \to \infty$, we obtain

$$\bar{\Psi}(\zeta_0) \le C_2\rho^\gamma + \sup\{\mathcal{M}\Psi(\tilde{\zeta}); \tilde{\zeta} \in G_\rho\}.$$

Hence by Lemma 3.3 and (3.20)

$$\begin{aligned}\bar{\Psi}(\zeta_0) &\le \lim_{\rho\to 0}(C_2\rho^\gamma + \sup\{\mathcal{M}\Psi(\tilde{\zeta}); \tilde{\zeta} \in G_\rho\}) \\ &= \overline{\mathcal{M}\Psi}(\zeta_0) \le \mathcal{M}\bar{\Psi}(\zeta_0) < \Psi(\zeta_0) - \eta.\end{aligned}$$

This contradiction proves claim (3.25) and completes the proof that $\bar{\Psi}$ is a viscosity subsolution.

(B) Next we prove that $\underline{\Psi}$ is a viscosity supersolution. So we let f be a C^2 function in a neighborhood of $\mathcal{S}$ and we let $\zeta_0 \in \mathcal{S}$ be such that $f \le \underline{\Psi}$ on $\mathcal{S}$ and $f(\zeta_0) = \underline{\Psi}(\zeta_0)$. We want to show that

$$F(D^2 f(\zeta_0), Df(\zeta_0), f(\zeta_0), \underline{\Psi}, \zeta_0) \le 0.$$

Since by Lemma 3.3(ii) $\mathcal{M}\underline{\Psi} - \underline{\Psi} \le 0$ everywhere, we see from (3.5) that this holds for $\zeta_0 \in \ell_1 \cup \ell_2$ and it suffices to show that

$$\mathcal{L}f(\zeta_0) \le 0 \qquad \text{for } \ \zeta_0 \in \mathcal{S}^0 \cup [0, Q]$$

and

$$\mathcal{L}_0 f(\zeta_0) \le 0 \qquad \text{for } \ \zeta_0 \in [0, P].$$

For $\varepsilon > 0$ let $\widehat{w} = \widehat{w}_{\varepsilon,c}$ be an admissible control beginning with a constant consumption rate $c \ge 0$ and no transactions up to the first time τ_ε at which the process $Z^c(t)$ exits from

$$K_\varepsilon = \{(s,x,y); |(s,x,y) - (0,x_0,y_0)| < \varepsilon\} \cap \widetilde{\mathcal{S}},$$

where $\zeta_0 = (x_0, y_0)$. Choose $\zeta_n \in K_\varepsilon$ such that $\zeta_n \to \zeta_0$ and $\Psi(\zeta_n) \to \underline{\Psi}(\zeta_0)$ as $n \to \infty$.

Then by combining Dynkin's formula with the *dynamic programming principle* ([Kr, Theorem 6, p. 150]) we get for all n

$$\begin{aligned}\Psi(\zeta_n) &\ge E^{0,\zeta_n}\left[\int_0^\tau e^{-\delta t}\frac{c^\gamma}{\gamma}dt + e^{-\delta\tau}\Psi(X^{(\hat{w})}(\tau), Y^{(\hat{w})}(\tau))\right] \\ &\ge E^{0,\zeta_n}\left[\int_0^\tau e^{-\delta t}\frac{c^\gamma}{\gamma}dt + e^{-\delta\tau} f(X^{(\hat{w})}(\tau), Y^{(\hat{w})}(\tau))\right] \\ &= f(\zeta_n) + E^{0,\zeta_n}\left[\int_0^\tau e^{-\delta t}\left(L^c f(X^{(\hat{w})}(t), Y^{(\hat{w})}(t)) + \frac{c^\gamma}{\gamma}\right)dt\right].\end{aligned}$$

We conclude that

$$E^{0,\zeta_n}\left[\int_0^{\tau_\varepsilon} e^{-\delta t}\left\{L^c f(X^{(\hat{w})}(t), Y^{(\hat{w})}(t)) + \frac{c^\gamma}{\gamma}\right\} dt\right] \le \Psi(\zeta_n) - f(\zeta_n) \qquad \text{for all } n.$$

Taking the limit as $n \to \infty$, we obtain

$$E^{0,\zeta_0}\left[\int_0^{\tau_\varepsilon} h(t)dt\right] \le 0,$$

where

$$h(t) = e^{-\delta t}\left(L^c f(X^{(\hat{w})}(t), Y^{(\hat{w})}(t)) + \frac{c^\gamma}{\gamma}\right).$$

By dividing the left-hand side by $E^{0,\zeta_0}[\tau_\varepsilon]$ we get

$$\frac{E^{0,\zeta_0}\left[\int_0^{\tau_\varepsilon} h(t)dt\right]}{E^{0,\zeta_0}[\tau_\varepsilon]} = \frac{E^{0,\zeta_0}\left[\int_0^{\tau_\varepsilon}(h(t)-h(0))dt\right] + h(0)E^{0,\zeta_0}[\tau_\varepsilon]}{E^{0,\zeta_0}[\tau_\varepsilon]}$$
$$\longrightarrow h(0) \qquad \text{as } \varepsilon \to 0, \text{ since } h(t) \text{ is continuous at } t = 0.$$

We conclude that

$$h(0) = L^c f(\zeta_0) + \frac{c^\gamma}{\gamma} \le 0 \tag{3.27}$$

for all $c \ge 0$ such that $\widehat{w}_{\varepsilon,c}$ is admissible for ε small enough. If $\zeta_0 \in \mathcal{S}^0 \cup [0, Q]$, then this is clearly the case for all $c \ge 0$, and therefore (3.27) implies that $\mathcal{L}f(\zeta_0) \le 0$. If $\zeta_0 \in [0, P]$, then the only such admissible c is $c = 0$. Therefore we get $\mathcal{L}_0 f(\zeta_0) \le 0$ in this case, as required. □

Next we turn to the question of uniqueness. Our second main result in the section is the following theorem.

THEOREM 3.8 (Comparison theorem).

(i) *Suppose that u is a viscosity subsolution and v is a viscosity supersolution of* (3.4) *and that u and v satisfy the estimates*

$$-C|x+y|^\gamma \le u(x,y) \qquad \textit{for all} \quad (x,y) \in \mathcal{S}, \tag{3.28}$$

$$v(x,y) \le C|x+y|^\gamma \qquad \textit{for all} \quad (x,y) \in \mathcal{S}, \tag{3.29}$$

for some constant $C < \infty$. Then

$$u \le v \qquad \textit{in } \mathcal{S}^0.$$

(ii) *Moreover, if in addition*

$$v(x,y) = \liminf_{\substack{(\xi,\eta)\to(x,y)\\(\xi,\eta)\in\mathcal{S}^0}} v(\xi,\eta) \qquad \textit{for all } (x,y) \in \partial\mathcal{S}, \tag{3.30}$$

then

$$u \le v \qquad \textit{in } \mathcal{S}.$$

COROLLARY 3.9. *Suppose that u and v are two viscosity solutions of* (3.4) *satisfying* (3.28) *and* (3.29). *Then*

$$u = v \qquad \text{in } \mathcal{S}^0,$$

and u is continuous in $\mathcal{S}^0$. In particular, if (2.23) *holds, then the value function Ψ is continuous on $\mathcal{S}^0$.*

Proof of Corollary 3.9. Since u is a viscosity solution, it follows that $\overline{u}$ is a viscosity subsolution and $\underline{u}$ is a viscosity supersolution, and similarly for v. Hence, by Theorem 3.8,

$$\overline{u} \le \underline{v} \le \overline{v} \le \underline{u} \le \overline{u} \qquad \text{in } \mathcal{S}^0.$$

This implies that

$$\underline{u} = \overline{u} = \underline{v} = \overline{v} \qquad \text{in } \mathcal{S}^0.$$

The last statement of Corollary 3.9 now follows from Theorem 3.7 and Corollary 2.2. □

Proof of Theorem 3.8. The proof is based on the technique of Ishii (see [B], [CIL], and [IL]) and on the proofs of Lemma 3.12 in [AMS] and Theorem 5.7 in [AST]. Consequently we shall not give a detailed proof here but rather point out the special treatment required to handle the nonlocal intervention operator $\mathcal{M}$.

Let u and v be as in Theorem 3.8. We first construct a strict supersolution of (3.4) by making a perturbation of v. Choose $\gamma' \in (\gamma, 1)$ such that (see (1.23))

$$\delta > \gamma' \left[r + \frac{(\alpha - r)^2}{2\sigma^2(1-\gamma')} \right]. \tag{3.31}$$

Set

$$g(x,y) = (x+y)^{\gamma'} \tag{3.32}$$

and choose $\varepsilon > 0$. Then

$$\mathcal{M}(v + \varepsilon g) \le \mathcal{M}v + \varepsilon \mathcal{M}g \tag{3.33}$$

and hence

$$\mathcal{M}(v+\varepsilon g) - (v + \varepsilon g) \le (\mathcal{M}v - v) + \varepsilon(\mathcal{M}g - g). \tag{3.34}$$

Since v is a supersolution, we have

$$\mathcal{M}v - v \le 0. \tag{3.35}$$

Moreover, with $\zeta = (x,y)$,

$$\begin{aligned}(\mathcal{M}g - g)(\zeta) &= \sup_{\xi \ne 0}\{g(x - k - \xi - \lambda|\xi|, y + \xi)\} - g(x,y) \\ &= \sup_{\xi \ne 0}\{(x + y - k - \lambda|\xi|)^{\gamma'}\} - (x+y)^{\gamma'} \\ &\le (x+y)^{\gamma'} \left[\left(1 - \frac{k}{x+y}\right)^{\gamma'} - 1 \right]. \end{aligned} \tag{3.36}$$

Therefore, for each compact subset C of $\mathcal{S} \setminus \{0\}$ there exists $\rho_1 > 0$ such that $(\mathcal{M}g - g)(\zeta) \le -\rho_1$ for all $\zeta \in C$. So from (3.34) and (3.35) we get

$$\mathcal{M}(v + \varepsilon g) - (v + \varepsilon g) \le -\varepsilon\rho_1 \qquad \text{in } C. \tag{3.37}$$

Now if we define the operator L^0 by (see (2.2))

$$L^0 = -\delta I + rx\frac{\partial}{\partial x} + \alpha y\frac{\partial}{\partial y} + \tfrac{1}{2}\sigma^2 y^2\frac{\partial^2}{\partial y^2}, \tag{3.38}$$

where I is the identity operator, then

$$\begin{aligned} L^0 g(x,y) &= -\delta(x+y)^{\gamma'} + (rx + \alpha y)\gamma'(x+y)^{\gamma'-1} \\ &\quad + \tfrac{1}{2}\sigma^2 y^2\gamma'(\gamma'-1)(x+y)^{\gamma'-2} \\ &= (x+y)^{\gamma'}\left[-\delta + \gamma'\frac{rx+\alpha y}{x+y} + \tfrac{1}{2}\sigma^2\gamma'(\gamma'-1)\frac{y^2}{(x+y)^2}\right]. \end{aligned} \tag{3.39}$$

If we put

$$\eta = \frac{y}{x+y} \qquad \text{so that} \qquad \frac{x}{x+y} = 1 - \eta,$$

then we get

$$L^0 g(x,y) = (x+y)^{\gamma'}[-\delta + \gamma' r + \gamma'(\alpha - r)\eta + \tfrac{1}{2}\sigma^2\gamma'(\gamma'-1)\eta^2].$$

By (3.31) it follows that

$$L^0 g(x,y) < 0 \qquad \text{for all } (x,y) \ne (0,0).$$

Consequently, on every compact C of $\mathcal{S} \setminus \{0\}$ there exists $\rho_2 > 0$ such that

$$L^0 g(x,y) + \max_{c \ge 0}\left(-c\frac{\partial g}{\partial x}\right) \le -\rho_2 \qquad \text{on } C.$$

Therefore, since v is a supersolution of (3.4), we conclude that on every compact C of $\mathcal{S} \setminus \{0\}$ there exists $\rho > 0$ such that

$$v_\varepsilon := v + \varepsilon g$$

is a viscosity supersolution of

$$F(D^2 v_\varepsilon(\zeta), Dv_\varepsilon(\zeta), v_\varepsilon(\zeta), v_\varepsilon, \zeta) = -\varepsilon\rho \qquad \text{for } \zeta \in C.$$

Let us now prove the theorem by contradiction. Assume that

$$\sup_{\zeta \in \mathcal{S}}\{u(\zeta) - v(\zeta)\} > 0. \tag{3.40}$$

Choose $\varepsilon > 0$ such that

$$\sup_{\zeta \in \mathcal{S}}\{u(\zeta) - v_\varepsilon(\zeta)\} > 0. \tag{3.41}$$

Define

$$h(\zeta) := u(\zeta) - v_\varepsilon(\zeta). \tag{3.42}$$

Since h is usc and tends to $-\infty$ when $|\zeta| \to \infty$, the set

$$\text{Argmax } h := \{\overline{\zeta}; h(\overline{\zeta}) = \sup\{h(\zeta); \zeta \in \mathcal{S}\}\}$$

is nonempty and compact in $\mathcal{S} \setminus \{0\}$. Choose an open set $G \subset \mathcal{S} \setminus \{0\}$ containing this compact (G open relative to $\mathcal{S}$) and with $\overline{G}$ compact. In order to get a contradition it suffices to prove that

$$u \leq v_\varepsilon \qquad \text{in } G.$$

Thus we have reduced the problem to proving a comparison theorem for a strict supersolution v_ε and a subsolution u of (3.4) in an open subset G of $\mathcal{S} \setminus \{0\}$ with compact closure $\overline{G}$, when the supremum of $u - v_\varepsilon$ is attained in G only. This is proved by using Ishii's technique, adapted as in [B, Theorem 4.6] and in [AMS, Theorem 5.7] for the boundary conditions. We now explain this in more detail, as follows.

For $j = 1, 2, \ldots$, define, for $(\zeta, \eta) \in \mathcal{S} \times \mathcal{S}$,

$$H_j(\zeta, \eta) = u(\zeta) - v_\varepsilon(\eta) - \frac{j}{2}|\zeta - \eta|^2 \tag{3.43}$$

and set

$$m_j = \sup\{H_j(\zeta, \eta); (\zeta, \eta) \in \mathcal{S} \times \mathcal{S}\} \tag{3.44}$$

and

$$m = \sup\{h(\zeta); \zeta \in \mathcal{S}\}. \tag{3.45}$$

Proceeding exactly as in [AMS], we obtain that there exist ζ_j, η_j in $\mathcal{S}$ such that

$$m_j = H_j(\zeta_j, \eta_j) < \infty. \tag{3.46}$$

Moreover,

$$j|\zeta_j - \eta_j|^2 \to 0 \qquad \text{as } j \to \infty \tag{3.47}$$

and

$$m_j \to m \qquad \text{as } j \to \infty. \tag{3.48}$$

Suppose

$$\text{Argmax } h \text{ is contained in } \mathcal{S}^0 \tag{3.49}$$

and choose $\hat{\zeta} \in \mathcal{S}^0$ such that

$$m = h(\hat{\zeta}).$$

Then we get that

$$(\zeta_j, \eta_j) \in \mathcal{S}^0 \times \mathcal{S}^0 \qquad \text{for } j \text{ large enough.} \tag{3.50}$$

By applying [CIL, Theorem 3.2] we now obtain that there exist 2×2 matrices P_j, Q_j such that

$$(p_j, P_j) \in \bar{J}^{2,+} u(\zeta_j) \qquad \text{and} \qquad (q_j, Q_j) \in \bar{J}^{2,-} v_\varepsilon(\eta_j)$$

and

$$\begin{bmatrix} P_j & 0 \\ 0 & -Q_j \end{bmatrix} \leq 3j \begin{bmatrix} I & -I \\ -I & I \end{bmatrix}, \tag{3.51}$$

where

$$p_j = j(\zeta_j - \eta_j) \qquad \text{and} \qquad q_j = p_j.$$

Since u is a subsolution and v_ε is a strict supersolution, we obtain

$$\overline{F}(P_j, p_j, u, \zeta_j) \geq 0 \tag{3.52}$$

and

$$\underline{F}(Q_j, q_j, v_\varepsilon, \eta_j) \leq -\varepsilon\rho. \tag{3.53}$$

From (3.53) it follows that

$$\Lambda(Q_j, q_j, v_\varepsilon, \eta_j) < 0,$$

and proceeding as in [AMS], we obtain

$$\Lambda(P_j, p_j, u, \zeta_j) - \Lambda(Q_j, q_j, v_\varepsilon, \eta_j) < 0. \tag{3.54}$$

Consequently $\Lambda(P_j, p_j, u, \zeta_j) < 0$, and from (3.52) we obtain that

$$(\mathcal{M}u - u)(\zeta_j) \geq 0. \tag{3.55}$$

From (3.53) we have

$$(\mathcal{M}v_\varepsilon - v_\varepsilon)(\eta_j) \leq -\varepsilon\rho < 0. \tag{3.56}$$

Therefore, combining (3.55) and (3.56), we get

$$m_j < u(\zeta_j) - v_\varepsilon(\eta_j) < \mathcal{M}u(\zeta_j) - \mathcal{M}v_\varepsilon(\eta_j) - \varepsilon\rho. \tag{3.57}$$

Since $\zeta_j, \eta_j \to \hat{\zeta}$ and u is usc, we obtain

$$m < \liminf_{j\to\infty} [\mathcal{M}u(\zeta_j) - \mathcal{M}v_\varepsilon(\eta_j)]. \tag{3.58}$$

Since u is usc and v_ε is lsc we see, after some reflections, that

$$\limsup_{j\to\infty} \mathcal{M}u(\zeta_j) \leq \mathcal{M}u(\hat{\zeta}) \tag{3.59}$$

and

$$\limsup_{j\to\infty} (-\mathcal{M}v_\varepsilon(\eta_j)) \leq -\mathcal{M}v_\varepsilon(\hat{\zeta}). \tag{3.60}$$

Hence we get the desired contradiction:

$$\begin{aligned} m &< \mathcal{M}u(\hat{\zeta}) - \mathcal{M}v_\varepsilon(\hat{\zeta}) \\ &= \sup_{\xi_1 \neq 0} \{u(\hat{x}'(\xi_1), \hat{y}'(\xi_1))\} - \sup_{\xi_2 \neq 0} \{v_\varepsilon(\hat{x}'(\xi_2), \hat{y}'(\xi_2))\} \\ &\leq \sup_{\xi \neq 0} \{(u - v_\varepsilon)(\hat{x}'(\xi), \hat{y}'(\xi))\} \\ &\leq \sup\{(u - v_\varepsilon)(\zeta); \zeta \in \mathcal{S}\} = m, \end{aligned}$$

where $\hat{\zeta} = (\hat{x}, \hat{y})$ and (see (2.3))

$$\hat{x}'(\xi) = \hat{x} - \xi - \lambda|\xi| - k, \qquad \hat{y}'(\xi) = \hat{y} + \xi.$$

This contradiction shows that assumption (3.49) cannot hold. Therefore we must have

$$\hat{\zeta} \in \text{Argmax } h \cap \partial\mathcal{S}. \tag{3.61}$$

To treat the boundary points we proceed exactly as in the proof of Theorem 5.7 of [AST], which itself is based on the proof of Theorem 4.6 in [B] (see the appendix there, p. 166).

From (3.30) there exists a sequence $\{\zeta_j\} \subset \mathcal{S}^0 \cap G$ converging to $\hat{\zeta}$ such that $v_\varepsilon(\zeta_j) \to v_\varepsilon(\hat{\zeta})$ when $j \to \infty$. Let $\varepsilon_j = |\zeta_j - \hat{\zeta}|$. The idea is now to introduce the test function

$$w_j(\zeta, \eta) = u(\zeta) - v_\varepsilon(\eta) - \theta_j(\zeta, \eta), \qquad (\zeta, \eta) \in \mathcal{S} \times \mathcal{S},$$

where

$$\theta_j(\zeta, \eta) = \frac{|\zeta - \eta|^2}{2\varepsilon_j} + \frac{1}{4}\left(\frac{d(\eta) - d(\zeta)}{d(\zeta_j)} - 1\right)^4 + \frac{1}{4}|\zeta - \hat{\zeta}|^4.$$

Here $d(\eta)$ denotes the distance from η to $\partial\mathcal{S}$, and similarly for $d(\zeta), d(\zeta_j)$.

Following exactly the same steps as in [AST] and treating the term $\mathcal{M}u - u$ as we did before, we obtain a contradiction also in the case (3.61). This shows that (3.40) cannot hold and this completes the proof of Theorem 3.8. □

We summarize the results of this section in the following.

THEOREM 3.10. *Let $\Psi(x, y)$ be the value function given by* (1.17). *Suppose* (2.23) *holds, i.e.,*

$$\delta > \gamma\alpha. \tag{3.62}$$

Then Ψ is continuous on $\mathcal{S}^0$, and Ψ is the unique viscosity solution of (3.4) *with the property that there exists $C < \infty$ such that*

$$|\Psi(x, y)| \le C|x + y|^\gamma \qquad \textit{for all } (x, y) \in \mathcal{S}. \tag{3.63}$$

4. Numerical results. In this section we present the result of a numerical method used to approximate the viscosity solution of (3.4) in the case when the solvency region is $\mathcal{S}_+ = [0, \infty) \times [0, \infty)$. (See Remark 1.1.) This method is detailed in [CØS].

The problem is first localized on $D := (0, L) \times (0, L)$, assuming zero Neumann boundary conditions on the localized boundary. The localized problem is then solved by using an iterative method, which permits us to obtain the QVHJBI as a limit of variational HJBIs. Each variational inequality is approximated by a finite difference scheme and then solved by a Howard algorithm.

Figure 4.1 represents the optimal transaction policy for the following values of the parameters: $k = 0.05, \lambda = 0.1, \sigma = 0.3, r = 0.07, \delta = 0.1, \gamma = 0.3, \alpha = 0.11, L = 100$.

The results are relevant only in a smaller domain, for example $[0, 50] \times [0, 50]$, because of the side effects of the truncature and the artificial boundary conditions set for $x = 100$ and $y = 100$.

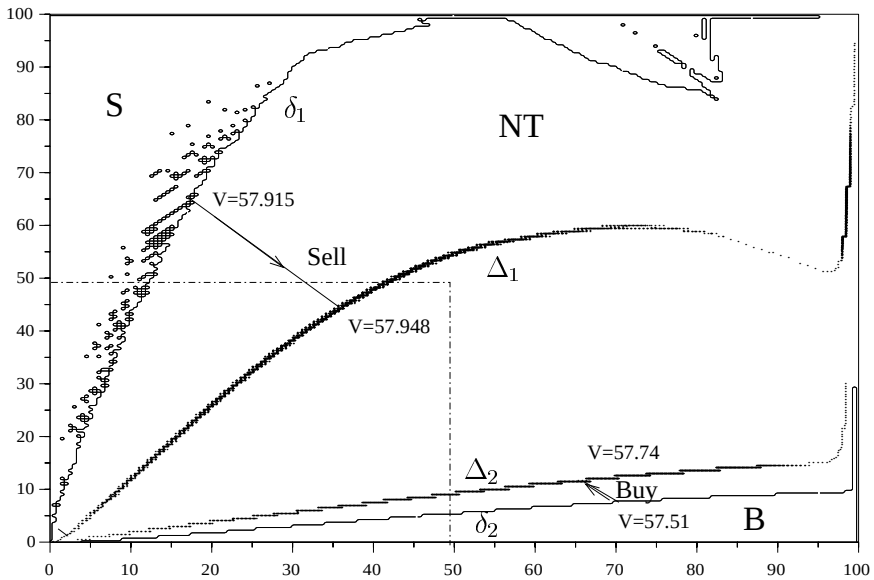


FIG. 4.1. *Optimal transaction policy (see text for parameter values).*

The domain consists of 3 regions: buy (B), sell (S), and no transaction (NT).

The set of states Δ_1 and Δ_2 reached after a purchase or a sale of stock are plotted. They are situated inside the continuation set NT. Unlike the case of no fixed costs, these lines do not coincide with the boundaries δ_1 and δ_2 of NT.

After a transaction, the position of the investor evolves as a pure diffusion process inside NT until it reaches the boundary. Then, a jump occurs back to the closest of the two lines Δ_1, Δ_2 in the transaction directions.

It is natural to ask whether it is possible to obtain more information about the shape of the no-transaction region NT, both for this choice $\mathcal{S}_+$ of solvency region and for the choice $\mathcal{S}$ given by (1.9). As mentioned in the introduction, we know that if $k = 0$, then NT is bounded by two straight lines from the origin [DN] (see Figure 1.2). If $k > 0$, is NT still bounded by two curves? If so, what can be said about the form of these curves? Can they be given an explicit description?

Remark 4.1. For results on the viscosity solutions of QVIs corresponding to impulse control problems (which, however, do not apply to our situation), see [I], [P], and [TY].

Acknowledgments. We wish to thank Marianne Akian, Guy Barles, Jean-Philippe Chancelier, Nils Christian Framstad, and Kristin Reikvam for very helpful comments and fruitful discussions.

REFERENCES

[AMS] M. AKIAN, J. L. MENALDI, AND A. SULEM, *On an investment-consumption model with transaction costs*, SIAM J. Control Optim., 34 (1996), pp. 329–364.

[AST] M. AKIAN, A. SULEM, AND M. I. TAKSAR, *Dynamic optimization of long term growth rate for a portfolio with transaction costs and logarithmic utility*, Math. Finance, 11 (2001), pp. 153–188.

[B] G. Barles, *Solutions de Viscosité des Équations de Hamilton-Jacobi*, Math. Appl. 17, Springer-Verlag, Paris, 1994.

[BL] A. Bensoussan and J.-L. Lions, *Impulse Control and Quasi-Variational Inequalities*, Gauthier–Villars, Paris, 1984.

[BØ] K. A. Brekke and B. Øksendal, *A verification theorem for combined stochastic control and impulse control*, in Stochastic Analysis and Related Topics, Vol. 6, Progr. Probab. 42, Birkhäuser Boston, Cambridge, MA, 1998, pp. 211–220.

[BP] T. R. Bielecki and S. R. Pliska, *Risk sensitive asset management with transaction costs*, Finance Stoch., 4 (2000), pp. 1–33.

[CIL] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., 27 (1992), pp. 1–67.

[CZ1] A. Cadenillas and F. Zapatero, *Optimal central bank intervention in the foreign exchange market*, J. Econom. Theory, 87 (1999), pp. 218–242.

[CZ2] A. Cadenillas and F. Zapatero, *Classical and impulse stochastic control of the exchange rate using interest rates and reserves*, Math. Finance, 10 (2000), pp. 141–156.

[CØS] J.-P. Chancelier, B. Øksendal, and A. Sulem, *Combined stochastic control and optimal stopping, with application to portfolio optimization under fixed transaction costs,* Proceedings of the Steklov Institute of Mathematics, to appear.

[DN] M. H. A. Davis and A. Norman, *Portfolio selection with transaction costs*, Math. Oper. Res., 15 (1990), pp. 676–713.

[EH] J. E. Eastham and K. J. Hastings, *Optimal impulse control of portfolios*, Math. Oper. Res., 13 (1988), pp. 588–605.

[FØS1] N. C. Framstad, B. Øksendal, and A. Sulen, *Optimal consumption and portfolio in a jump diffusion market*, in A. Shiryaev and A. Sulem, eds., Workshop on Mathematical Finance, INRIA, Paris, 1998.

[FØS2] N. C. Framstad, B. Øksendal, and A. Sulen, *Optimal consumption and portfolio in a jump diffusion market with proportional transaction costs*, J. Math. Econom., 35 (2001), pp. 233–257.

[I] K. Ishii, *Viscosity solutions of nonlinear second order elliptic PDEs associated with impulse control problems*, Funkcial. Ekvac., 36 (1993), pp. 123–141.

[IL] H. Ishii and P. L. Lions, *Viscosity solutions of fully nonlinear second-order elliptic partial differential equations*, J. Differential Equations, 83 (1990), pp. 26–78.

[IW] N. Ikeda and S. Watanabe, *Stochastic Differential Equations and Diffusion Processes*, 2nd ed., North–Holland/Kodansha, Tokyo, 1989.

[K] R. Korn, *Portfolio optimization with strictly positive transaction costs and impulse control*, Finance Stoch., 2 (1998), pp. 85–114.

[Kr] N. V. Krylov, *Controlled Diffusion Processes*, Springer-Verlag, New York, Berlin, 1980.

[M] R. C. Merton, *Optimum consumption and portfolio rules in a continuous time model*, J. Econom. Theory, 3 (1971), pp. 373–413.

[MØ] G. Mundaca and B. Øksendal, *Optimal stochastic intervention control with application to the exchange rate*, J. Math. Econom., 29 (1998), pp. 225–243.

[Ø] B. Øksendal, *Stochastic Differential Equations*, 5th ed., Springer-Verlag, Berlin, New York, 2000.

[ØS] B. Øksendal and A. Sulem, *Introduction to Impulse Control Theory and Applications to Economics*, lecture notes, INRIA/University of Oslo, Oslo, Norway, 2001.

[P] B. Perthame, *Recent results on the quasi-variational inequality of the impulse control. Contributions to Nonlinear Partial Differential Equations, Vol.* II, Pitman Res. Notes Math. Ser., 155 (Paris, 1985), pp. 210–219.

[RØ] K. Reikvam and B. Øksendal, *Viscosity solutions of optimal stopping problems*, Stochastics Stochastics Rep., 62 (1998), pp. 285–301.

[SS] S. E. Shreve and H. M. Soner, *Optimal investment and consumption with transaction costs*, Ann. Appl. Probab., 4 (1994), pp. 609–692.

[TY] S. Tang and J. Yong, *Finite horizon stochastic optimal switching and impulse controls with a viscosity solution approach*, Stochastics Stochastics Rep., 45 (1993), pp. 145–176.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 6, pp. 1791–1820

ROBUST CONTROL VIA SEQUENTIAL SEMIDEFINITE PROGRAMMING*

B. FARES†, D. NOLL†, AND P. APKARIAN‡

Abstract. This paper discusses nonlinear optimization techniques in robust control synthesis, with special emphasis on design problems which may be cast as minimizing a linear objective function under linear matrix inequality (LMI) constraints in tandem with nonlinear matrix equality constraints. The latter type of constraints renders the design numerically and algorithmically difficult. We solve the optimization problem via *sequential semidefinite programming* (SSDP), a technique which expands on sequential quadratic programming (SQP) known in nonlinear optimization. Global and fast local convergence properties of SSDP are similar to those of SQP, and SSDP is conveniently implemented with available semidefinite programming (SDP) solvers. Using two test examples, we compare SSDP to the augmented Lagrangian method, another classical scheme in nonlinear optimization, and to an approach using concave optimization.

Key words. nonlinear programming, sequential semidefinite programming, robust gain-scheduling control design, linear matrix inequalities, nonlinear matrix equalities

AMS subject classifications. 93B51, 90C55

PII. S0363012900373483

1. Introduction.

A variety of problems in robust control design can be cast as minimizing a linear objective subject to linear matrix inequality (LMI) constraints and additional nonlinear matrix equality constraints:

$$(D)\quad \begin{array}{ll} \text{minimize} & d^T x \\ \text{subject to} & \mathcal{A}(x) \le 0, \\ & \mathcal{B}(x) = 0, \end{array}$$

where d is a given vector, x denotes the vector of decision variables, $\mathcal{A}(x)$ is an affine symmetric matrix function, ≤ 0 means negative semidefinite, and $\mathcal{B}(x)$ is a nonlinear matrix valued function, which in many cases is bilinear in x. In the present paper, we are primarily interested in robust gain-scheduling control design, but a variety of other design problems may be cast in the form (D). Without aiming at completeness, let us just mention examples like fixed or reduced-order $\mathcal{H}_2$ and $\mathcal{H}_\infty$ synthesis, robust control synthesis with different classes of scalings, robust control design with parameter-dependent Lyapunov functions, robust control of nonlinear systems with integral-quadratic-constraints (IQC)-defined components, and, more generally, minimization or feasibility problems with bilinear matrix inequality (BMI) constraints. We discuss some of these applications of (D) at more detail.

Example 1. Observe that the reduced-order $\mathcal{H}_\infty$ synthesis problem may be cast

*Received by the editors June 6, 2000; accepted for publication (in revised form) June 5, 2001; published electronically March 5, 2002.
http://www.siam.org/journals/sicon/40-6/37348.html
†Université Paul Sabatier, Mathématiques pour l'Industrie et la Physique, 118, route de Narbonne, 31062 Toulouse, France (fares@cict.fr, noll@mip.ups-tlse.fr).
‡ONERA-CERT, Control Systems Department, 2 av. Edouard Belin, 31055 Toulouse, France (apkarian@cert.fr).

as

$$(H_\infty)\quad \begin{array}{ll} \text{minimize} & d^T x \\ \text{subject to} & \mathcal{A}(x) \le 0, \\ & \text{rank } \mathcal{Q}(x) \le r, \end{array}$$

where $\mathcal{A}(x)$ and $\mathcal{Q}(x)$ are symmetric affine. One way to transform (H_∞) into the form (D) is to introduce a slack matrix variable W of size $q \times r$, q the dimension of $\mathcal{Q}(x)$, let $\tilde{x} = (x, W)$ be the new decision vector, and introduce the quadratic equality constraint

$$\mathcal{B}(\tilde{x}) = \mathcal{Q}(x) - W^T W = 0.$$

In special situations, there may be better suited ways to obtain the form (D).

Example 2. The BMI-feasibility problem is a near at hand application of our method. If the BMI appears in standard form

$$\mathcal{B}(x) = \mathcal{A}(x) + \sum_{1 \le i < j \le n} B_{ij} x_i x_j$$

for an affine symmetric matrix valued function $\mathcal{A}$ and symmetric matrices B_{ij}, we are readily led to introduce a slack variable $z_{ij} = x_i x_j$, and replace the BMI with a new LMI in tandem with the nonlinear constraints $z_{ij} - x_i x_j = 0$. In practice, we are more likely to encounter BMIs or even multilinear matrix inequalities, featuring terms of the form $X_i A X_j$ with X_i, X_j parts of the decision vector. In this event, introducing an auxiliary decision matrix variable $Z_{ij} = X_i A X_j$ will have the same effect and transform the constraint set into the form of LMIs plus algebraic equalities.

Example 3. As a special case of a BMI problem, consider static output feedback control design, where we have to find a Lyapunov matrix variable $X > 0$ and a controller K such that for given matrices A, B, C the BMI

$$(A + BKC)X + X(A + BKC)^T < 0$$

is satisfied. Introducing a new variable $W = KCX$, we could readily transform this into an LMI plus a nonlinear matrix equality, $KCX - W = 0$, to obtain the program (D).

An alternative way to obtain the form (D) is to open the BMI via the projection lemma [18]. This leads to two LMIs,

$$\mathcal{N}_{B^T}^T (AX + XA^T)\mathcal{N}_{B^T} < 0, \qquad \mathcal{N}_C^T (YA + A^T Y)\mathcal{N}_C < 0,$$

in tandem with $X = Y^{-1}$. Here $\mathcal{N}_{B^T}$, $\mathcal{N}_C$ are bases for the null spaces of B^T, C. With the nonlinear equality constraint rearranged as $XY - I = 0$, we obtain a second version of (D).

It seems appealing to include the LMI

$$\begin{pmatrix} X & I \\ I & Y \end{pmatrix} \ge 0$$

among the above; as with $Y > 0$, and via Schur complement, it is equivalent to $X - Y^{-1} \ge 0$. While becoming redundant near the optimum, the new LMI will

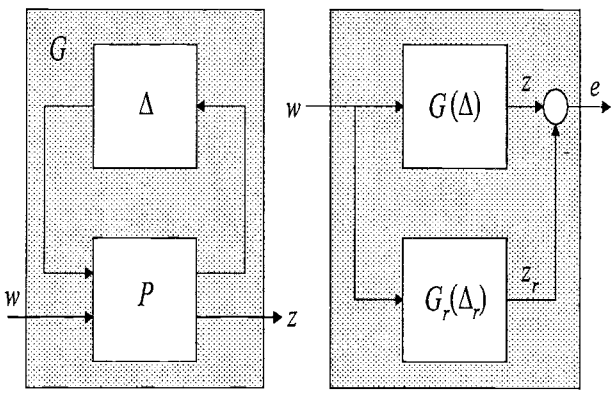


FIG. 1. *Designing reduced LFT model.*

help to stabilize the problem. Notice, however, that this idea, which has even been used to relax the static output feedback problem into an LMI problem, is no longer applicable in the more complicated robust design problem we shall present in more detail in section 2.

Example 4. Yet another important case is robust control design via generalized Popov multipliers (cf. [33, 28]), also known as k_m or μ synthesis. Here we encounter a BMI of the form

$$(P + UKV)^T S^T + S(P + UKV) \leq 0$$

to be solved for S and K for given P, U, V. By introducing a slack matrix variable $G = SUKV + (SUKV)^T$, the design problem may be cast in the form (D) as

$$\begin{aligned} \text{minimize} \quad & d^T x, x = (S, C, G) \\ \text{subject to} \quad & P^T S^T + SP + G \leq 0, \\ & SUKV + (SUKV)^T - G = 0. \end{aligned}$$

A similar situation occurs in mixed $\mathcal{H}_2/\mathcal{H}_\infty$-control design, where a BMI of the more general form

$$\Psi + (P + UKV)^T S^T \Phi + \Phi S(P + UKV) \leq 0$$

with fixed Ψ, Φ, P, U, V, to be solved for S, K, arises. This could now be handled using $G = \Phi SUKV + (\Phi SUKV)^T$.

Example 5. For many robust control problems, linear fractional transformations (LFT) are used to model plants with uncertain components or to represent nonlinear systems as uncertain linear systems. The corresponding LFTs are often highly complex and difficult to handle numerically, and techniques for reducing the order of LFT representations are required. One way to compute a reduced-order LFT approximation of the nominal LFT is by minimizing the worst-case energy discrepancy between outputs of the nominal and the reduced plant in response to arbitrary finite-energy input signals (see Figure 1). This approach admits a formulation of the form (D). See, e.g., [21] for more details.

The nonlinear constraint $\mathcal{B}(x) = 0$ renders problem (D) highly complex and difficult to solve in practice (cf. [13]). Nonetheless, due to its importance, various

heuristics and ad hoc methods have been developed over recent years to obtain suboptimal solutions to (D). Methods currently employed are usually *coordinate descent schemes*, which alternatively and iteratively fix parts of the coordinates of the decision vector, x, trying to optimize the remaining indices. The D-K (scaling-controller) iteration procedure is an example of this type [6, 37], whose popularity may be attributed to the fact that it is conceptually simple and easily implemented as long as the intermediate steps are convex LMI programs. The latter may often be guaranteed through an appropriate choice of the decision variables held fixed at each step. However, a major drawback of coordinate descent schemes is that they almost always fail to converge, even for starting points close to a local solution (see [22]). As a result, controllers obtained via such methods are highly questionable and bear the risk of unnecessary conservatism.

A new optimization approach to robust control design was initiated in [5], where the authors showed that reduced-order $\mathcal{H}_\infty$ control could be cast as a concave minimization problem. It was observed, however, that in a number of cases local concave minimization, which is known to be numerically difficult, produced unsatisfactory results. This occurs, in particular, when iterations get stalled, which is probably due to the lack of second-order information.

In [16], we therefore proposed a different approach to (D), again based on nonlinear optimization techniques. The *augmented Lagrangian method* from nonlinear optimization was successfully extended to program (D). The difficult nonlinear constraints were incorporated into an augmented Lagrangian function, while the LMI constraints, due to their linear structure, were kept explicitly during optimization. A Newton-type method including a line search, or, alternatively, a trust-region strategy, was shown to work if the penalty parameters were appropriately increased at each step, and if the so-called first-order update rule for the Lagrange multiplier estimates (cf. [9]) was used.

The disadvantage of the augmented Lagrangian method is that its convergence is at best linear if the penalty parameter c is held fixed. Superlinear convergence is guaranteed if $c \to \infty$, but the use of large c, due to the inevitable ill-conditioning, is prohibitive in practice. The present investigation therefore aims at adapting methods with better convergence properties, like sequential quadratic programming (SQP), to the case of LMI constrained problems. Minimizing at each step the second-order Taylor expansion of the Lagrangian of (D) about the current iterate defines the *tangent subproblem*, (T), whose solution will provide the next iterate. Due to the constraints $\mathcal{A}(x) \leq 0$, (T) is not a quadratic program, as in the case of SQP, but requires minimizing a quadratic objective function under LMI constraints. After convexification of the objective, (T) may be turned into a semidefinite program, conveniently solved with current LMI tools (cf., for instance, [20, 36]). We refer to this approach as *sequential semidefinite programming* (SSDP). It will be discussed in section 4, and a local convergence analysis will be presented in section 5. Although more complex than most coordinate descent schemes, the advantages of the new approach are at hand:

- The entire vector x of decision variables is updated at each step, so, for instance, we do not have to separate Lyapunov and scaling variables from controller variables.
- Like SQP, SSDP is guaranteed to converge globally, which means, for an arbitrary and possibly remote initial guess, if an appropriate line search or trust region strategy is applied.
- Being of second-order type, the rate of convergence of SSDP is superlinear in

a neighborhood of attraction of a local optimum.

The present paper discusses and compares three nonlinear optimization techniques suited for the design problem (D) with special emphasis on SSDP since it performed best. The reader might be missing an approach via interior-point techniques—perhaps more in the spirit of the age. In fact, in a different context, Jarre [24] proposes such a method based on the log-barrier function known from the interior-point approach to the semidefinite programming (SDP) problem but does not present any numerical evidence as to the practicality of the approach. Theoretical *and* practical results are presented by Leibfritz and Mostafa [25, 26], who consider static output feedback control and mixed $\mathcal{H}_2/\mathcal{H}_\infty$-control. Our own numerical experiments [3] with interior-point methods for robust control design seem to indicate that those are generally less robust and that the different parameters may be difficult to tune. We emphasize that the method proposed for robust control design is modulable in the sense that the optimization procedure featuring SSDP may be replaced by any other tool based on the user's favorite optimizer. Future investigations will show which methods work best in a given situation, and the present contribution does not claim to present the ultimate tool.

The paper is organized as follows. Section 2 presents and develops the setting of the robust gain-scheduling control, a particularly important application of (D). Even though the full robust gain-scheduling case has never been presented, let alone attacked algorithmically, we keep this part rather cursory, as the individual steps of the method are essentially known. We rely on a recent excellent exposition of the material by Scherer [35] and related texts [29, 1, 21]. We have chosen this problem as our main motivating case study, as it seems to be among the most difficult and numerically demanding cases of the scheme (D).

Section 3 aims at practical aspects. We offer more specific choices of parameter uncertainties and scaling variables which help to reduce the algorithmic complexity of the problem and, as far as our own experiments go, work well in practice.

Section 4 gives a description of the SSDP method as it naturally emerges from the classical SQP method. Local superlinear and quadratic convergence of SSDP is shown in section 5. While several convergence proofs for the SQP method are known in the literature, (cf. [11, 12]), they all seem to depend heavily on the polyhedrality of the classical-order cone, and no extension addressing the semidefinite cone seems available. The proof we present here is fairly general and includes nonlinear programming with more general order cones.

Numerical aspects of the SSDP method are discussed in section 7. Using two typical test examples, we compare it to the augmented Lagrangian method and to concave programming. While apparently of moderate size, these examples represent cases where classical approaches like the D-K iteration perform poorly or are even at complete loss.

2. Robust gain-scheduling control design. We wish to design a robust gain-scheduling controller for a plant which depends rationally on the uncertain and scheduled parameters. Consider an LFT plant in standard form described by the state-space equations

$$\left(\begin{array}{c} \dot{x} \\ \hline z_\theta \\ z \\ y \end{array}\right) = \left(\begin{array}{c|ccc} A & B_\theta & B_1 & B_2 \\ \hline C_\theta & D_{\theta\theta} & D_{\theta 1} & D_{\theta 2} \\ C_1 & D_{1\theta} & D_{11} & D_{12} \\ C_2 & D_{2\theta} & D_{21} & 0 \end{array}\right) \left(\begin{array}{c} x \\ \hline w_\theta \\ w \\ u \end{array}\right), \qquad w_\theta = \Theta z_\theta, \tag{1}$$

where $\Theta(t)$ is a time-varying matrix valued parameter assumed to have a two-block diagonal structure

$$\Theta = \begin{pmatrix} \Theta_m & \\ & \Theta_u \end{pmatrix}. \tag{2}$$

Here $\Theta_m(t)$ represents the scheduled parameters, measured on-line, and $\Theta_u(t)$ represents the time-varying parametric uncertainties, which we allow to vary in a known compact set $\mathcal{K}$ of matrices. We call parameters Θ of this form *admissible*, and the set of admissible (scheduled and bounded uncertain) parameters is denoted Θ.

We recall that the limiting case, no Θ_u (all parameters measured), is called the linear parameter-varying (LPV) or gain-scheduling control problem, while the case no Θ_m (all parameters uncertain) is referred to as the robust control problem.

The state-space entries of the plant (1) with inputs w, u and outputs z, y are rational functions of the parameters Θ_m and Θ_u. The meaning of the signals is as follows: u is the control input, y is the measurement signal, w stands for the vector of exogenous signals, and z stands for regulated variables.

The robust gain-scheduling control design requires finding a linear controller K of the form

$$\left(\begin{array}{c} \dot{\bar{x}} \\ \hline \bar{z}_\theta \\ u \end{array}\right) = \left(\begin{array}{c|cc} A_K & B_{K\theta} & B_{K1} \\ \hline C_{K\theta} & D_{K\theta\theta} & D_{K\theta 1} \\ C_{K2} & D_{K2\theta} & D_{K21} \end{array}\right) \left(\begin{array}{c} \bar{x} \\ \hline \bar{w}_\theta \\ y \end{array}\right), \quad \bar{w}_\theta = \phi(\Theta_m)\bar{z}_\theta, \tag{3}$$

where ϕ is called the *scheduling function*, to be determined as part of the design, such that (3) fulfills the following requirements:

- The closed-loop system, obtained by substituting (3) into (1), is internally stable.
- The L_2-gain of the closed-loop operator mapping w to z is bounded by γ.
- The above specifications hold for *all* admissible parameter trajectories $\Theta \in \Theta$.

In order to continue our analysis, we apply a convenient procedure first used in [29, 1]. We gather all parameter-dependent components into a single block, which leads to an augmented plant $\bar{P}(s)$ described in the frequency domain as

$$\left(\begin{array}{c} \bar{z}_\theta \\ \hline z_\theta \\ z \\ y \\ \hline \bar{w}_\theta \end{array}\right) = \overbrace{\left(\begin{array}{c|c|c} 0 & 0 & I \\ \hline 0 & P(s) & 0 \\ \hline I & 0 & 0 \end{array}\right)}^{\bar{P}(s)} \left(\begin{array}{c} \bar{w}_\theta \\ \hline w_\theta \\ w \\ u \\ \hline \bar{z}_\theta \end{array}\right). \tag{4}$$

It is easy to verify pictorially that the original scheme shown on the left-hand side of Figure 2 is equivalent to the one shown on the right-hand side using the augmented system $\bar{P}(s)$. After closing the loop, i.e., substituting (3) into (1), respectively, (4), the closed-loop systems mapping exogenous inputs w to regulated outputs z are the same on both sides.

By inspecting the left-hand diagram in Figure 2, we see that the original robust gain-scheduling control problem can now be viewed as a standard robust control problem for the time-invariant plant $\bar{P}$ facing the augmented uncertain parameter matrix $\tilde{\Theta}$, where

$$\tilde{\Theta} = \begin{pmatrix} \phi(\Theta_m) & \\ & \Theta \end{pmatrix}, \quad \Theta = \begin{pmatrix} \Theta_m & \\ & \Theta_u \end{pmatrix}.$$

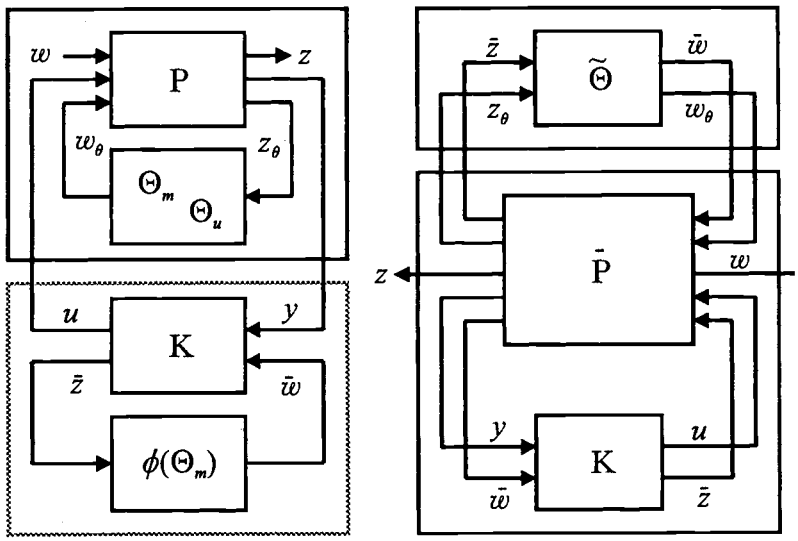


FIG. 2. *Gain-scheduling robust control.*

Based on this reformulation, sufficient conditions for the existence of a robust gain-scheduling linear time-invariant (LTI) controller (3) are consequently obtained by a suitable extension of the usual procedure in robust control design: Apply the bounded real lemma and the generalized S-procedure with a suitable choice of scalings to obtain sufficient conditions for robust stability of the closed-loop system (Lemma 1). Then use the projection theorem [18, 35] to eliminate the state-space variables of the controller K. The sufficient conditions for solvability are now again formulated in terms of the state-space entries (1) in conjunction with the Lyapunov and scaling variables (Theorem 2). They form, as we shall see, a mix of LMIs and nonlinear algebraic equalities. Now, at the core of the procedure, calculate the optimal gain using the proposed optimization techniques. As a final step, extract the robust controller K from the decision parameters used during optimization using, e.g., the method in [1].

We proceed to present the details of this scheme for the robust gain-scheduling control case. At the present stage, we aim at a fairly general approach, but the next section will focus on the practical aspects, where some of the theoretically possible steps will have to be reconsidered regarding their numerical performance. This concerns, in particular, the choice of the Lyapunov test matrix used in the bounded real lemma, the S-procedure, and the choice of the scalings (structured or general).

In this section, we allow for a fairly general class of scalings $\mathcal{Q}$ of the form

$$\mathcal{Q} = \begin{pmatrix} \mathcal{Q}_m & \\ & Q_u \end{pmatrix}, \quad \mathcal{Q}_m = \begin{pmatrix} Q_1 & Q_2 \\ Q_2^T & Q_m \end{pmatrix} \tag{5}$$

compatible with the block structure of $\tilde{\Theta}$. Later on we shall, at the cost of some conservatism, consider more special classes of scalings in order to reduce the numerical burden in the design.

Remark. Let us address the question of choosing the Lyapunov test function. Although parameter-dependent Lyapunov functions can be used (see [7, 17] for discussions), in the present paper, we shall restrict our attention to the more traditional

single quadratic Lyapunov approach based on a parameter-independent Lyapunov matrix $\mathcal{P}_0$. This choice is at the cost of some conservatism but keeps the theoretical descriptions simple and practically useful.

For the notation, observe that we use script matrix symbols $\mathcal{A}$, $\mathcal{B}_1$, $\mathcal{B}_\theta$, etc., for the state-space data of the closed-loop system obtained by substituting (3) into (4); see Figure 2. We have the following lemma.

LEMMA 1. *Suppose there exists a Lyapunov matrix $\mathcal{P}_0 > 0$ and scalings $\mathcal{Q}$, $\mathcal{R}$, and $\mathcal{S}$ of the above form* (5) *such that the nonlinear matrix inequality*

$$\star \left(\begin{array}{cc|cc|cc} 0 & I & & & & \\ I & 0 & & & & \\ \hline & & \mathcal{Q} & \mathcal{S} & & \\ & & \mathcal{S}^T & \mathcal{R} & & \\ \hline & & & & -\gamma & 0 \\ & & & & 0 & \frac{1}{\gamma} \end{array}\right) \left(\begin{array}{ccc} \mathcal{P}_0 & 0 & 0 \\ \mathcal{A} & \mathcal{B}_\theta & \mathcal{B}_1 \\ \hline 0 & I & 0 \\ \mathcal{C}_\theta & \mathcal{D}_{\theta\theta} & \mathcal{D}_{\theta 1} \\ \hline 0 & 0 & I \\ \mathcal{C}_1 & \mathcal{D}_{1\theta} & \mathcal{D}_{11} \end{array}\right) < 0 \tag{6}$$

is satisfied. Further suppose that the scalings satisfy the condition

$$\left(\begin{array}{c} \tilde{\Theta} \\ I \end{array}\right)^T \left(\begin{array}{cc} \mathcal{Q} & \mathcal{S} \\ \mathcal{S}^T & \mathcal{R} \end{array}\right) \left(\begin{array}{c} \tilde{\Theta} \\ I \end{array}\right) \geq 0 \tag{7}$$

for each admissible $\tilde{\Theta}$. Then the closed-loop system is robustly stable over the uncertain set Θ. Moreover, for every admissible $\tilde{\Theta} \in \Theta$, the operator mapping the exogenous signal w into the regulated variables z has an L_2-gain bounded above by γ.

Proof. The result is essentially the same as that of Theorem 10.4 in [35]. It consists of applying the bounded real lemma in tandem with the full block S-procedure. ∎

Remark. The derived sufficient conditions for robust gain-scheduling control are not suited for practice as they stand. This is mainly due to the infinite constraint (7), which involves an infinity of test matrices $\tilde{\Theta}$. In the following section, we shall indicate in which way (7) may, at the cost of some conservatism, be turned into a finite condition.

A second aspect of the derived criteria is that (6) is not jointly convex in the decision variables $\mathcal{P}_0$, $\mathcal{Q}$, $\mathcal{R}$, $\mathcal{S}$, and K. As a consequence, using these variables in the design is a difficult problem not suited for the usual convexity techniques in control.

As we shall see in our next step, the nonconvexity of the design problem may to some extent be reduced through the projection lemma [18]. As a result, the solvability conditions are stated back in terms of the original state-space entries in tandem with the Lyapunov and scaling variables, whereas the controller variable K has been eliminated. The mild inconvenience of this is that the actual controller has to be obtained in an extra step using the decision variables in Theorem 2 below. This step may itself be numerically demanding if the scheduling function ϕ has some undesirable properties.

THEOREM 2. *Consider the LFT plant* (1) *with scheduled and uncertain parameters $\Theta \in \Theta$ as in* (2). *Let $\mathcal{N}_X$ and $\mathcal{N}_Y$ be bases of the null spaces of $(C_2, D_{\theta 2}, D_{12}, 0, 0)$ and $(B_2^T, D_{2\theta}^T, D_{21}^T, 0, 0)$, respectively. Suppose there exist scalings Q, R, S, $\tilde{Q}$, $\tilde{R}$, $\tilde{S}$ of the form*

$$Q = \left(\begin{array}{cc} Q_m & \\ & Q_u \end{array}\right), \quad R = \left(\begin{array}{cc} R_m & \\ & R_u \end{array}\right), \text{etc.} \tag{8}$$

compatible with the block structure of Θ *in* (2) *and a pair of symmetric matrices* (X, Y) *satisfying the matrix completion conditions*

$$\begin{pmatrix} X & I \\ I & Y \end{pmatrix} > 0, \tag{9}$$

such that the linear matrix inequalities (10)–(12)

$$\mathcal{N}_X^T \begin{pmatrix} A^TX + XA & XB_\theta + C_\theta^T S^T & XB_1 & C_\theta^T R & C_1^T \\ B_\theta^T X + SC_\theta & Q + SD_{\theta\theta} + D_{\theta\theta}^T S^T & SD_{\theta 1} & D_{\theta\theta}^T R & D_{1\theta}^T \\ B_1^T X & D_{\theta 1}^T S^T & -\gamma I & D_{\theta 1}^T R & D_{11}^T \\ RC_\theta & RD_{\theta\theta} & RD_{\theta 1} & -R & 0 \\ C_1 & D_{1\theta} & D_{11} & 0 & -\gamma I \end{pmatrix} \mathcal{N}_X < 0, \tag{10}$$

$$\mathcal{N}_Y^T \begin{pmatrix} AY + YA^T & YC_\theta^T + B_\theta \tilde{S} & YC_1^T & B_\theta \tilde{Q} & B_1 \\ C_\theta Y + \tilde{S}^T B_\theta^T & D_{\theta\theta}\tilde{S} + \tilde{S}^T D_{\theta\theta} - \tilde{R} & \tilde{S}^T D_{1\theta}^T & D_{\theta\theta}\tilde{Q} & D_{\theta 1} \\ C_1 Y & D_{1\theta}\tilde{S} & -\gamma I & D_{1\theta}\tilde{Q} & D_{11} \\ \tilde{Q}B_\theta^T & \tilde{Q}D_{\theta\theta}^T & \tilde{Q}D_{1\theta}^T & \tilde{Q} & 0 \\ B_1^T & D_{\theta 1}^T & D_{11}^T & 0 & -\gamma I \end{pmatrix} \mathcal{N}_Y < 0, \tag{11}$$

$$\begin{pmatrix} \Theta \\ I \end{pmatrix}^T \begin{pmatrix} Q & S \\ S^T & R \end{pmatrix} \begin{pmatrix} \Theta \\ I \end{pmatrix} \geq 0 \text{ for every } \Theta \in \boldsymbol{\Theta} \tag{12}$$

in tandem with the nonlinear algebraic equality

$$\begin{pmatrix} Q_u & S_u \\ S_u^T & R_u \end{pmatrix}^{-1} = \begin{pmatrix} \tilde{Q}_u & \tilde{S}_u \\ \tilde{S}_u^T & \tilde{R}_u \end{pmatrix} \tag{13}$$

are satisfied. Then there exists an nth order gain-scheduling controller K *(*n *the order of the plant* (1)*), and a choice of the scheduling function* ϕ *such that the closed-loop system is internally and robustly stable, and the operator mapping* w *into* z *has* L_2*-gain bounded by* γ *for all admissible parameter trajectories* $\Theta \in \boldsymbol{\Theta}$*.*

Proof. The argument is based on a solvability test for quadratic inequalities developed in [34, 1, 2]. The most recent reference is Lemma 10.2 in [35]. This result is used to eliminate the controller variable K from the solvability conditions (6) in Lemma 1.

When applying the solvability test, due to the special structure of $\bar{P}(s)$, the solvability conditions obtained simplify to (9)–(12), (13). The question which remains is how the scheduled part of the coupling condition (13) is avoided. Following Theorem 10.11 of [35], one may show that the variables Q_1, Q_2, R_1, R_2, S_1, S_2 in the scheduled part of the multipliers are not called for by the matrix inequalities (9)–(12) and are therefore free to be chosen to satisfy the scheduled part of (13). This also requires a special choice of the scheduling function ϕ given in [35]. □

As already mentioned, condition (12) needs to be worked on in order to become numerically tractable. This aspect is treated in the next section.

3. Choices suited for practice. In this section, we address the practical aspects of the control design part and indicate that, at the cost of some conservatism, the difficulty of the design may be greatly reduced by accepting some restrictions in the general outline.

To begin with, let us assume that the uncertain matrix function $\Theta_u(t)$ varies in a polyhedral convex and compact set $\mathcal{K}$ of matrices, i.e., $\Theta_u(t) \in \mathcal{K} = \mathrm{co}\{\Theta_{u1}, \ldots, \Theta_{uN}\}$ at all times t. We refer to the Θ_{ui} as the vertices of the value set. Let us examine the consequence of this choice. Observe that due to the block structure of $\tilde{\Theta}$, the infinite dimensional scaling condition (12) already decouples into a scheduled part and an uncertain part. Concerning the uncertain part, we have the following lemma.

LEMMA 3. *Suppose the value set of $\Theta_u(t)$ is polyhedral and the scaling satisfies $Q_u < 0$. Then the uncertain part of condition* (12) *is equivalent to the finite condition*

$$\begin{pmatrix} \Theta_{ui} \\ I \end{pmatrix}^T \begin{pmatrix} Q_u & S_u \\ S_u^T & R_u \end{pmatrix} \begin{pmatrix} \Theta_{ui} \\ I \end{pmatrix} \geq 0 \qquad \text{for every } i = 1, \ldots, N. \tag{14}$$

The proof is in fact a straightforward convexity argument based on $Q_u < 0$ and may be found, e.g., in [19, 35]. This settles the question of finiteness for the uncertain part of (12) at the slight cost of conservatism introduced by assuming $Q_u < 0$.

Remark. We mention that in practice it is sufficient to let Θ_u have a block diagonal structure of the form

$$\Theta_u(t) = \mathrm{diag}(\theta_{u1}(t) I_{p_1}, \ldots, \theta_{ur}(t) I_{p_r}), \tag{15}$$

where we may without loss assume that $|\theta_{uj}| \leq 1$, so the set $\mathcal{K}$ will be a cube with the 2^r vertices $\theta_{uj}(t) = \pm 1$.

The conservatism introduced to obtain the finite condition (14) is minor and acceptable in practice. Notice that the number N may become inconveniently large if the number of parameters θ_{ui} grows. We therefore mention another strategy to avoid the infinite scaling condition. Assuming that the uncertain parameters have the block diagonal structure (15), we consider what we call *structured scalings* satisfying the following conditions: (i) Q_u and S_u commute with $\Theta_u(t)$; (ii) $R_u = -Q_u$ and $R_u > 0$; (iii) $S_u^T = -S_u$. We check that the scheduled part of condition (7) is satisfied. Developing the term gives

$$\Theta_u^2 Q_u + \Theta_u S_u + S_u^T \Theta_u + R_u = (I - \Theta_u^2) R_u \geq 0$$

as required. This choice of the scaling appears rather special and therefore bears the risk of unnecessary conservatism, but its merit is that it greatly reduces the number of decision variables and LMI constraints.

Let us now consider the corresponding questions for the scheduled part of (12). We start with the following technical lemma, which was already used in the proof of Theorem 2; cf. [35].

LEMMA 4. *Suppose the scalings Q_m, S_m, and R_m have been found such that*

$$\begin{pmatrix} \Theta_m \\ I \end{pmatrix}^T \begin{pmatrix} Q_m & S_m \\ S_m^T & R_m \end{pmatrix} \begin{pmatrix} \Theta_m \\ I \end{pmatrix} \geq 0 \quad \text{for every } \Theta \in \boldsymbol{\Theta}. \tag{16}$$

Then there is a choice of the scheduling function ϕ along with appropriate choices of

Q_1, Q_2, R_1, R_2, and S_1, S_2 such that the scheduled part of (7) *is satisfied, i.e.,*

$$\begin{pmatrix} \tilde{\Theta}_m \\ I \end{pmatrix}^T \begin{pmatrix} \mathcal{Q}_m & \mathcal{S}_m \\ \mathcal{S}_m^T & \mathcal{R}_m \end{pmatrix} \begin{pmatrix} \tilde{\Theta}_m \\ I \end{pmatrix} \geq 0 \quad \textit{for every } \Theta \in \mathbf{\Theta}. \tag{17}$$

Proof. As shown in [35], if $\Phi_m := [Q_m S_m; S_m^T R_m]$ satisfies (16), it is always possible to adjust the extended scalings $\mathcal{Q}_m = [Q_1 Q_2; Q_2^T Q_m]$, $\mathcal{R}_m = [R_1 R_2; R_2^T R_m]$, $\mathcal{S}_m = [S_1 S_2; S_2^T S_m]$ in such a way that, with an appropriate choice of the scheduling function ϕ, the scheduled part (17) of condition (12) holds true. An explicit formula for ϕ is given in [35]. ∎

This means that we are left to define a class of scalings Q_m, R_m S_m, which allows the reduction of the infinite set of LMIs (16) to a finite set. If Θ_m has a block diagonal structure

$$\Theta_m = \operatorname{diag}(\theta_{m1} I_{\ell_1}, \ldots, \theta_{ms} I_{\ell_s}) \tag{18}$$

and if prior bounds $|\theta_{mj}(t)| \leq 1$ like for the uncertain parameters are available, this may be done in exactly the same way as for the uncertain part.

Assuming block diagonal structures (15), (18) for both types of parameters, we find it useful in practice to pursue different strategies for the two types of parameters. We use the vertex idea to render the uncertain part of (7) finite, and we use structured scalings for the scheduled parameters. This avoids numerical difficulties which may arise when constructing K if a complicated scheduling function ϕ is required. The use of structured scalings allows the choice $\phi(x) = x$. Notice that for the scheduled parameters, due to conditions (i) above, choosing structured scalings implies that each of the subblocks Q_1, Q_2, Q_m of $\mathcal{Q}_m$ has the block diagonal structure with diagonal blocks of sizes $\ell_1, \ldots, \ell_s$ in (18). This option finally was a good compromise in our numerical tests, and we recommend its use for the type of problem under investigation.

4. Sequential semidefinite programming. In this section, we cast the robust gain-scheduling control design problem as an optimization problem and present an algorithmic approach to its solution.

Recall from Theorem 2 that the complete vector of decision variables for design is $x = (\gamma, Q, R, S, \tilde{Q}, \tilde{R}, \tilde{S}, X, Y)$. We find it notationally useful to point to parts of the vector x by introducing the notation

$$\Phi_u = \begin{pmatrix} Q_u & S_u \\ S_u^T & R_u \end{pmatrix}, \qquad \tilde{\Phi}_u = \begin{pmatrix} \tilde{Q}_u & \tilde{S}_u \\ \tilde{S}_u^T & \tilde{R}_u \end{pmatrix},$$

involving the uncertain blocks of the scaling variables. Similarly, Φ_m, $\tilde{\Phi}_m$ regroup the scheduled parts of Q, R, S, $\tilde{Q}$, $\tilde{R}$, $\tilde{S}$.

Let $\mathcal{A}(x) \leq 0$ represent the LMI constraints (9)–(12), where (12), using one of the techniques from the previous section, has been replaced with a finite set of LMIs, along with $Q \leq 0$ and $\tilde{Q} \leq 0$ required for these procedures. Finally, let $\mathcal{B}(x) = \Phi_u \tilde{\Phi}_u - I = 0$ represent the nonlinear algebraic constraint (13). Then the robust gain-scheduling control problem may be cast in the form (D). More generally, we consider an augmented version (D_c) of (D) for a penalty parameter $c \geq 0$:

$$(D_c) \quad \begin{array}{ll} \text{minimize} & f_c(x) = \gamma + \frac{c}{2} \|\Phi_u \tilde{\Phi}_u - I\|^2 \\ \text{subject to} & \mathcal{A}(x) \leq 0, \\ & \mathcal{B}(x) = \Phi_u \tilde{\Phi}_u - I = 0. \end{array}$$

Remark. Notice that problems (D) and (D_c) are equivalent since the penalty term $\frac{c}{2}\|\Phi_u\tilde{\Phi}_u - I\|^2$ added in (D_c) will vanish at the optimal x. Using (D_c) instead of (D), as we shall see, may add some numerical stability.

Remark. We observe that the variables Q_m, R_m, S_m, $\tilde{Q}_m$, $\tilde{R}_m$, $\tilde{S}_m$, and X, Y occur only in the LMI constraint, which strongly indicates that we expect redundancies in the decision parameters. In fact, our experiments indicate that this is a strong point for using structured scalings in the Θ_m block, as this tends to limit these redundancies. In general, we propose to put bounds $\|\cdot\|_\infty \le M$ on the free variables in order to avoid degeneracy or failure of the successive LMI subproblems. As these additional constraints may be included among the LMIs, $\mathcal{A}(x) \le 0$, we do not change the notation here.

Remark. Notice that the trick used in Examples 1 and 2 of the introductory section does not apply in the robust synthesis case, as the matrices Φ_u, $\tilde{\Phi}_u$ are indefinite. This shows that the problem is as a rule numerically harder than, e.g., static output feedback design or reduced order design.

Let us now extend the idea of SQP to the augmented program (D_c). As we aim at a *primal-dual* method, this requires maintaining estimates for the decision *and* Lagrange multiplier variables. Consider the Lagrangian associated with (D_c):

$$L_c(x;\Lambda,\lambda) = f_c(x) + \operatorname{trace}(\Lambda\cdot\mathcal{A}(x)) + \lambda^T \operatorname{vec}(\Phi_u\tilde{\Phi}_u - I), \tag{19}$$

where $\Lambda \ge 0$ is a positive semidefinite dual matrix variable, λ is a traditional Lagrange multiplier variable whose dimension is m^2, and m is the size of the matrices Φ_u, $\tilde{\Phi}_u$. Given the current iterate x and the current Lagrange multiplier estimates λ, $\Lambda \ge 0$, we define the *tangent problem*

$$(T)\quad \begin{array}{ll} \text{minimize} & \nabla f_c(x)^T\Delta x + \frac{1}{2}\Delta x^T\nabla^2 L_c(x;\Lambda,\lambda)\Delta x \\ \text{subject to} & \mathcal{A}(x+\Delta x) \le 0, \\ & \Phi_u\tilde{\Phi}_u + \Phi_u\Delta\tilde{\Phi}_u + \Delta\Phi_u\tilde{\Phi}_u - I = 0, \end{array}$$

which consists of minimizing the second-order Taylor polynomial of $L_c(x+\Delta x;\Lambda,\lambda)$ about the current x for possible steps Δx, subject to the LMI constraints $\mathcal{A} \le 0$ and the equality constraint $\mathcal{B} = 0$ linearized about the current Φ_u, $\tilde{\Phi}_u$. Notice that the equality constraint above is given in matrix notation. The equivalent expression in long vector notation using the Kronecker product $\otimes$ is

$$(\tilde{\Phi}_u \otimes I)\ \operatorname{vec}(\Delta\Phi) + (I\otimes\Phi_u)\ \operatorname{vec}(\Delta\tilde{\Phi}_u) - \operatorname{vec}(I - \Phi_u\tilde{\Phi}_u) = 0. \tag{20}$$

Here $\tilde{\Phi}_u \otimes I$ is invertible as soon as $\tilde{\Phi}_u$ has maximal rank, while $I\otimes\Phi_u$ is invertible as soon as Φ_u has maximal rank.

Remark. If either Φ_u or $\tilde{\Phi}_u$ is positive definite, we may symmetrize the equality constraint, as considered, e.g., in [15]. As mentioned before, this is typically not possible in the robust synthesis case but may help in different cases.

The choice of (T) is understood by inspecting the necessary optimality conditions, which show that the solution Δx of (T) may be considered as the Newton step from the current point x to the new iterate $x^+ = x+\Delta x$. The Lagrange multipliers $\Lambda^+ \ge 0$ and λ^+ belonging to the linear constraints in (T) are the updates for Λ and λ. Notice that $\Lambda^+ \ge 0$ as a consequence of the Kuhn–Tucker conditions for (T). Further notice that, despite the notation, Λ does not explicitly appear in the Hessian $\nabla^2 L_c(x;\Lambda,\lambda)$ of the Lagrangian, a fact which is due to the linearity of $\mathcal{A} \le 0$. On the other hand,

due to nonlinearity of the equality constraint, λ appears explicitly in the Hessian of the Lagrangian. Updating Λ is then still mandatory to obtain the update λ^+.

Remark. At this stage, we observe that due to the linearity of the LMI constraints, the iterates produced by the SSDP scheme will always satisfy the LMIs, while the nonlinear equality constraint will of course be only approximately satisfied. The fact that we iterate on decision variables satisfying the LMIs is an advantage of our method since it may render even suboptimal solutions of the optimization problem (D_c) useful for the design (cf. the termination phase in the robust control design algorithm presented at the end of this section).

The special structure and the moderate size of the variable $(\Phi_u, \tilde{\Phi}_u)$ occurring in the equality constraint $\mathcal{B} = 0$ suggest using a reduced Hessian technique. For fixed x, respectively, Φ_u, $\tilde{\Phi}_u$, we can eliminate either $\Delta\Phi_u$ or $\Delta\tilde{\Phi}_u$ from the linearized equality constraint in (T) as long as we maintain iterates x with full rank Φ_u, $\tilde{\Phi}_u$. In that event, the matrix $B = [\tilde{\Phi}_u \otimes I \; I \otimes \Phi_u]$ has full row rank m^2, m the size of the matrix Φ_u, and eliminating the equality constraint therefore reduces the problem size by m^2.

Following the standard notation in SQP, let Z be a matrix whose columns form a basis (preferably orthogonal) of the null space of the matrix B belonging to (20), and let the columns of Y form a basis for the range of B^T. Then we may write the displacement Δx as $\Delta x = Z\Delta\tilde{x} + Yw_0 = Z\Delta\tilde{x} + p_0$ for the fixed vector $p_0 = Y(BY)^{-1}\mathrm{vec}(I - \Phi_u\tilde{\Phi}_u)$, where $\Delta\tilde{x}$ is now the reduced decision vector.

With this notation, the reduced tangent problem is

$$(\tilde{T}) \qquad \begin{array}{ll} \text{minimize} & (\nabla f_c(x)^T Z + p_0^T \nabla^2 L_c Z)\Delta\tilde{x} + \frac{1}{2}\Delta\tilde{x} Z^T \nabla^2 L_c Z \Delta\tilde{x} \\ \text{subject to} & \mathcal{A}_* \circ Z(\Delta\tilde{x}) \leq -\mathcal{A}(x + p_0), \end{array}$$

where $\mathcal{A}_*$ is the linear part of $\mathcal{A}$. Notice that in general $(\tilde{T})$ is not yet an SDP since the reduced Hessian $Z^T\nabla^2 L_c Z$ may be indefinite. In order to obtain a convex program, we have to convexify the reduced Hessian, which may be done in several ways. We comment on these at the end of the section.

When the correction is done, the subproblem is convex and may easily be transformed into an SDP problem. Ideally, the solution $\Delta\tilde{x}$ gives rise to a step Δx in the original tangent problem, and the new iterate x^+ is obtained as $x + \Delta x$, but in practice a line search using an appropriate merit function is required. For appropriate choices avoiding the Maratos effect, we refer to the vast literature on the subject (see, e.g., [10], [12]).

In order to obtain the Lagrange multiplier updates, we have to inspect the necessary optimality conditions for $(\tilde{T})$. Let $\tilde{\Lambda}^+ \geq 0$ be the Lagrange multiplier matrix variable in $(\tilde{T})$ associated with the constraint $\mathcal{A} \leq 0$, and let $\Delta\tilde{x}$ be the optimal solution of $(\tilde{T})$. Then the optimal Δx is readily obtained via (20), Λ^+ is chosen as $\tilde{\Lambda}^+$, while λ^+ is found through

$$Y^T\nabla f_c(x) + Y^T\nabla^2 L_c(Z\Delta\tilde{x} + p_0) + Y^T\mathcal{A}_*^T\Lambda^+ + Y^T B^T\lambda^+ = 0, \tag{21}$$

which determines λ^+ uniquely if B has full rank. Conceptually, the SSDP algorithm proposed to solve (D) may be described as follows.

SSDP Algorithm.

1. Find an initial point x^0, such that $\mathcal{A}(x^0) \leq 0$ and such that Φ_u^0, $\tilde{\Phi}_u^0$ are full rank. Select Lagrange multiplier estimates λ^0 and $\Lambda^0 \geq 0$ using formula (21).
2. Given the iterate x^k with Φ_u^k and $\tilde{\Phi}_u^k$ nonsingular and multiplier estimates $\Lambda^k \geq 0$, λ^k, form the reduced tangent problem $(\tilde{T}_k)$ about the current data.

Render the reduced Hessian positive definite if required. Obtain the reduced step $\Delta\tilde{x}^k$ as a solution to the SDP problem, and let $\Delta x^k = Z_k\Delta\tilde{x}^k + p_0^k$. Obtain Lagrange multipliers $\Lambda^\sharp \geq 0$ and $\lambda^\sharp$ from $(\tilde{T}_k)$ using (21).

3. Do a line search in direction Δx^k using an appropriate merit function, and determine the new iterate $x^{k+1} = x^k + \alpha_k\Delta x^k$. Set $\Lambda^{k+1} = \Lambda^k + \alpha_k(\Lambda^\sharp - \Lambda^k)$ and $\lambda^{k+1} = \lambda^k + \alpha_k(\lambda^\sharp - \lambda^k)$. Choose α_k so that Φ_u^{k+1} and $\tilde{\Phi}_u^{k+1}$ are nonsingular.
4. Check the stopping criteria. Either halt or replace k by $k+1$, and go back to step 2.

In order to compute the Hessian $\nabla^2 L(x;\Lambda,\lambda)$ of the Lagrangian in step 2, only second-order derivatives with respect to Φ_u and $\tilde{\Phi}_u$ are required, as $f_c(x)$ is linear in γ and does not depend on the other decision variables. Using the Kronecker product $\otimes$, we have the following formulae (cf. also [16]).

LEMMA 5.

$$\nabla^2_{\tilde{\Phi}_u\Phi_u} L_c = c(\tilde{\Phi}_u \otimes I)^T(\tilde{\Phi}_u \otimes I),\ \nabla^2_{\tilde{\Phi}_u\tilde{\Phi}_u} L_c = c(I \otimes \Phi_u)^T(I \otimes \Phi_u),$$
$$\nabla^2_{\tilde{\Phi}_u\Phi_u} L_c = (I \otimes \mathrm{mat}(\lambda))^T + c((\Phi_u\tilde{\Phi}_u - I)^T \otimes I + (I \otimes \Phi_u)^T(\tilde{\Phi}_u \otimes I)).$$

Remark. Let us comment on the convexification of the reduced tangent problem $(\tilde{T})$ required to obtain an SDP problem. Recent trends in optimization indicate that one should dispense with this procedure. It is considered important to take the directions of negative curvature of the (reduced) Hessian into account, e.g., by using a trust region strategy or by doing sophisticated line searches which combine the Newton direction and the dominant direction of negative curvature. While the second idea could be at least partially realized, a trust region approach is not feasible as yet in the presence of LMI constraints as optimizing a nonconvex quadratic function subject to LMIs is presently too difficult numerically to become a functional scheme. We therefore have to use the well-known convexification methods used in nonlinear optimization over many years, and we refer to [9, 23] for several such strategies.

In our numerical experiments, we tested Powell's idea of doing a Cholesky factorization, and adding correction terms as soon as negative square roots appear, and a direct method which used the QR-factorization to correct negative eigenvalues of the reduced Hessian. A third method adapted to the structure of the problem which we found even more efficient consisted of a Gauss–Newton-type idea. We neglect the term $\Phi_u\tilde{\Phi}_u - I$ in the Hessian matrix (22), performing the modified Cholesky factorization on the remaining term. This is motivated by the fact that dropping this term leaves a positive semidefinite matrix, which is still close to the correct Hessian as long as the neglected term $\Phi_u\tilde{\Phi}_u - I$ is small. This is the case when the nonlinear constraint (13) is approximately satisfied, and the matrix is therefore asymptotically close to the correct (reduced) Hessian. As a consequence, and in contrast with the true Gauss–Newton method, this procedure therefore does not destroy the superlinear quadratic convergence of the scheme.

Observe that in all these procedures, the augmented form (D_c) of the program helps. In fact, the penalty term renders the Hessian more convex than in the original form (D), and so the corrections are often very mild in practice and, according to the theory in polyhedral programming, are not even required asymptotically (cf. [8, 10, 12]. This observation is corroborated in our experiments with LMI constraints.

We summarize the result of this section by presenting the following algorithmic approach to the robust gain-scheduling design problem.

Algorithm for robust gain-scheduling control design.

- *Step* 1. *Initialization.* Locate a strictly feasible decision vector x^0 for the LMI constraints: For fixed large enough $\gamma = \gamma_0$, render the LMIs (9)–(12) maximally negative by solving the SDP problem

$$\min\{t : \text{LMIs (9)–(12)} < tI\}.$$

  Then determine X_0, Y_0, Φ_u^0, and $\tilde{\Phi}_u^0$ so that $\Phi_u^0\tilde{\Phi}_u^0 - I$ is as close as possible to zero. Then initialize the Lagrange multiplier estimates λ_0 and $\Lambda_0 \geq 0$.
- *Step* 2. *Optimization.* Solve the optimization problem (D_c) via SSDP, using $(x^0, \Lambda^0, \lambda^0)$ as a primal-dual starting point. The primal solution is x.
- *Step* 3. *Terminating phase.* Due to nonlinearity, the algebraic constraint (13) is never exactly satisfied at the solution x. It is, however, possible to terminate the program without strict satisfaction of the nonlinear constraints by a simple *perturbation technique* [5], which is applicable as long as the LMIs (9)–(12) are strictly satisfied. One can then replace Φ_u with $\tilde{\Phi}_u^{-1}$ and check whether the LMI constraints (9)–(12) hold, possibly with new X and Y. In this case, a controller is readily obtained. Dually, we can replace $\tilde{\Phi}_u$ with Φ_u^{-1} and check the LMI constraints (9)–(11), with (7), respectively, (17) suitably replaced with its dual form

$$\begin{pmatrix} I \\ -\Theta_i^T \end{pmatrix}^T \tilde{\Phi}_u \begin{pmatrix} I \\ -\Theta_i^T \end{pmatrix} < 0 \quad \forall i = 1, \ldots, N.$$

  If the test fails, the numerical solution to (D_c) is unsatisfactory and has to be improved, e.g., by changing the stopping criteria or by increasing the penalty constant c and rerunning step 2.

Remark. Notice that strict feasibility < 0 is a priori not guaranteed by SSDP but may easily be forced if we replace ≤ 0 in the corresponding LMIs by the stronger $\leq -\varepsilon I$ for a small $\varepsilon > 0$. Moreover, if the SDP subproblem is solved by the notorious interior point techniques, the LMIs are automatically strictly satisfied, and the above perturbation argument is applicable.

5. Fast local convergence of SSDP. In this section, we prove local superlinear and quadratic convergence of the SSDP method under mild regularity hypotheses. It is interesting to recall the history of the SQP method, which was already popular during the late 1970s, even though the first proof of superlinear and quadratic convergence under realistic assumptions was published as late as 1994 by Bonnans [11]. A more compact version of that proof is published in [12]. Both versions are based on techniques introduced by Robinson in the 1980s.

The time interval is the more remarkable, as the equality constrained case was settled much earlier, apparently first by Boggs and Tolle around 1982. See [10] and the references given there. Early proofs of the general case existed but always reduced the situation to the equality constrained case under the (unrealistic) assumption of strict complementarity at the optimal pair.

Inspecting the convergence proofs for Newton's method in [11, 12] shows that they heavily depend on the polyhedrality of the order cone in classical nonlinear programming, so a natural extension to the present case of SSDP does not seem near at hand. Our present approach is nevertheless inspired by Bonnans's paper [11]. It turns out that our method of proof applies even to more general situations, and we present the method in a fairly general context.

We consider the nonlinear programming problem of the form

$$(P)\quad \begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g_E(x) = 0, \\ & g_I(x) \in K^0, \end{array}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $g_E : \mathbb{R}^n \to \mathbb{R}^m$, and $g_I : \mathbb{R}^n \to \mathbb{R}^p$ are $\mathcal{C}^2$-functions, K is a cone in $\mathbb{R}^p$, and K^0 is its polar cone defined as

$$K^0 = \{y \in \mathbb{R}^p : \langle x, y\rangle \le 0 \text{ for each } x \in K\}.$$

In the classical nonlinear programming case, $K = \mathbb{R}^p_+$, $K^0 = \mathbb{R}^p_-$, and the constraint $g_I(x) \in K^0$ becomes $g_i(x) \le 0$ componentwise, while in the semidefinite case, $\mathbb{R}^p \cong \mathbb{S}^r$ (with $p = r(r+1)/2$), the space of symmetric $r \times r$-matrices, $K = \mathbb{S}^r_+$ (with $K^0 = \mathbb{S}^r_-$), the cone of positive semidefinite matrices, and the constraint $g_i(x) \in K^0$ means that the matrix $g_i(x)$ is negative semidefinite. We use the notation $\langle \cdot, \cdot \rangle$ for the scalar product employed, since this may include the classical case $\langle x, y\rangle = \sum_i x_i y_i$ as well as $\langle x, y\rangle = \operatorname{trace}(x \cdot y)$ in the semidefinite case. The adjoint of an operator A with respect to this scalar product is denoted A^*; derivatives with respect to $\langle \cdot, \cdot \rangle$ in the x-variable are indicated by primes. Notice also that $g_I(x)$ was an affine matrix valued function in our applications, but we prefer to include the general nonlinear case as applications of this type are eminent.

We suppose that $\bar{x}$ is a local minimum of (P) and that there exists a Lagrange multiplier $\bar{\lambda} = (\bar{\lambda}_E, \bar{\lambda}_I)$ satisfying the necessary optimality condition

$$(KT)\quad \begin{array}{ll} (1) & f'(\bar{x}) + g'(\bar{x})^* \bar{\lambda} = 0, \\ (2) & g_E(\bar{x}) = 0, \\ (3) & g_I(\bar{x}) \in K^0, \lambda_I \in K, \langle g_I(x), \bar{\lambda}_I\rangle = 0. \end{array}$$

Observe that the existence of $\bar{\lambda}$ is guaranteed under a weak regularity assumption like, for instance, Robinson's constraint qualification hypothesis (cf. [32]). The Lagrangian associated with (P) is

$$L(x; \lambda) = f(x) + \langle g(x), \lambda\rangle = f(x) + \langle g_E(x), \lambda_E\rangle + \langle g_I(x), \lambda_I\rangle. \tag{22}$$

We consider Newton's method for solving the Kuhn–Tucker system (KT), which generates a sequence (x^k, λ^k) approximating the optimal pair $(\bar{x}, \bar{\lambda})$. Given the kth iterate (x^k, λ^k), the $(k+1)$st iterate is obtained by solving the tangent problem

$$(T_k)\quad \begin{array}{ll} \text{minimize} & \langle f'(x^k), \Delta x\rangle + \frac{1}{2}\langle \Delta x, L''(x^k; \lambda^k)\Delta x\rangle \\ \text{subject to} & g_E(x^k) + g'_E(x^k)\Delta x = 0, \\ & g_I(x^k) + g'_I(x^k)\Delta x \in K^0. \end{array}$$

If Δx is the solution to (T_k), then $x^{k+1} = x^k + \Delta x$. The Lagrange multiplier update $\lambda^{k+1} = (\lambda_E^{k+1}, \lambda_I^{k+1})$ is just the Lagrange multiplier belonging to the linearized constraints in (T_k). The Kuhn–Tucker conditions for (T_k) are the following:

$$(KT_k)\quad \begin{array}{l} L'(x^k; \lambda^{k+1}) + L''(x^k; \lambda^k)(x^{k+1} - x^k) = 0, \\ g_E(x^k) + g'_E(x^k)(x^{k+1} - x^k) = 0, \\ g_I(x^k) + g'_I(x^k)(x^{k+1} - x^k) \in K^0, \lambda_I^{k+1} \in K, \\ \qquad \langle g_I(x^k) + g'_I(x^k)(x^{k+1} - x^k), \lambda_I^{k+1}\rangle = 0. \end{array}$$

The aim of the following analysis is to give sufficient conditions for local quadratic or superlinear convergence of the sequence (x^k, λ^k).

Remark. The usual choice of quasi-Newton methods is easily obtained from our scheme by approximating the Hessian $L''(x^k; \lambda^k)$ of the Lagrangian of (P) by a matrix M^k. In order to account for modifications of $L''(x^k; \lambda^k)$ like convexifications as proposed in our experimental section, we include the quasi-Newton approach into our convergence analysis. We shall use the notation $(T_k(M^k))$ for the modified tangent problem with M^k replacing the Hessian of the Lagrangian.

Inspecting classical approaches for the usual polyhedral cone in nonlinear programming shows that local convergence of Newton's method usually requires two types of hypothesis: (a) the second-order sufficient optimality condition and (b) a constraint qualification. As we mentioned before, a third type of condition, strict complementarity, is often used but should be avoided since it is artificial as a rule. At the core is the second-order sufficient optimality condition, saying that the Hessian of the Lagrangian $L''(\bar{x}, \bar{\lambda})$ is positive definite along critical directions. We adopt the definition of critical directions from [32, 11], which in the presence of a multiplier leads to the following.

Definition. The direction $h \neq 0$ is critical at $\bar{x}$ with respect to the Lagrange multiplier $\bar{\lambda}$ if the following hold:

1. $g'_E(\bar{x})h = 0$.
2. There exist $h^k \to h$, $x^k = \bar{x} + t_k h^k$ with $t_k \to 0^+$ in tandem with $\lambda_I^k \to \bar{\lambda}_I$, $\lambda_I^k \in K$ such that for some v^k with $v^k = \mathrm{o}(t_k)$, $g_I(x^k) - v^k \in K^0$ and $\langle g_I(x^k) - v^k, \lambda_I^k \rangle = 0$.

Remark. Recall that in the case of the polyhedral cone $K = \mathbb{R}^p_+$ in nonlinear programming, a critical direction h satisfies conditions (1) for the equality constraints along with the following condition $(2')$ for inequalities: $g'_i(\bar{x})h = 0$ for active constraints $i \in I$ having multiplier $\bar{\lambda}_i > 0$, and $g'_i(\bar{x})h \leq 0$ for active constraints $i \in I$, where $\bar{\lambda}_i = 0$. It is an easy exercise to show that, in this case, (2) is equivalent to this classical definition of criticality $(2')$.

Let us now start analyzing the Newton step for (P) via the following perturbation result.

LEMMA 6. *Suppose there exist sequences $x^k \to \bar{x}$, $\lambda^k \to \bar{\lambda}$, $\delta_k \to 0^+$, and u^k, $v^k = (v_E^k, v_I^k)$ satisfying $u^k = \mathcal{O}(\delta_k)$, $v^k = \mathcal{O}(\delta_k)$ such that*

1. *$L'(x^k; \lambda^k) = u^k$;*
2. *$g_E(x^k) = v_E^k$;*
3. *$g_I(x^k) - v_I^k \in K^0$, $\langle g_I(x^k) - v_I^k, \lambda_I^k \rangle = 0$, $\lambda_I^k \in K$.*

Further suppose that the second-order sufficient optimality condition is satisfied at $(\bar{x}, \bar{\lambda})$, i.e., $\langle h, L''(\bar{x}; \bar{\lambda})h \rangle > 0$ for every critical direction $h \neq 0$, and that $g'(\bar{x})$ has maximal rank. Then $x^k - \bar{x} = \mathcal{O}(\delta_k)$ and $\lambda^k - \bar{\lambda} = \mathcal{O}(\delta_k)$.

Proof. Subtracting equation (1) in the Kuhn–Tucker equations from the perturbed equation (1) above gives

$$L'(x^k; \bar{\lambda}) - L'(\bar{x}; \bar{\lambda}) + g'(x^k)^*(\lambda^k - \bar{\lambda}) = u^k. \tag{23}$$

Now it suffices to show $x^k - \bar{x} = \mathcal{O}(\delta_k)$, for then the first term $L'(x^k; \bar{\lambda}) - L'(\bar{x}; \bar{\lambda})$ on the left-hand side of (23) is $\mathcal{O}(\delta_k)$, and hence so is the second term. Since $g'(x^k) = g'(\bar{x}) + \mathcal{O}(x^k - \bar{x})$, this implies $g'(\bar{x})^*(\lambda^k - \bar{\lambda}) = \mathcal{O}(x^k - \bar{x})$, and since $g'(\bar{x})$ has maximal rank, we conclude $\lambda^k - \bar{\lambda} = \mathcal{O}(x^k - \bar{x}) = \mathcal{O}(\delta_k)$.

Suppose now that the result is incorrect, so $u^k / \|x^k - \bar{x}\| \to 0$, $v^k / \|x^k - \bar{x}\| \to 0$. Picking a subsequence if necessary, we may assume that $(x^k - \bar{x})/\|x^k - \bar{x}\| \to h$ with

$\|h\| = 1$. We show that h is a critical direction.

Notice first that subtracting the perturbed condition (2) from condition (2) in the Kuhn–Tucker ensemble gives

$$\frac{g_E(x^k) - g_E(\bar{x})}{\|x^k - \bar{x}\|} = \frac{v_E^k}{\|x^k - \bar{x}\|} \to 0;$$

hence the equality part (1) of criticality is satisfied. As for the inequality part, observe that the perturbed conditions (1)–(3) just match the second part of the definition of criticality if we use the standing hypothesis that $x^k \to \bar{x}$ slower than $\delta_k \to 0$. Hence h is critical.

To conclude the proof, let us multiply (23) by $x^k - \bar{x}$ and divide by $\|x^k - \bar{x}\|^2$. The right-hand term of the modified equation is then

$$\frac{\langle u^k, x^k - \bar{x}\rangle}{\|x^k - \bar{x}\|^2} \to 0$$

by our standing hypothesis, so the left-hand side of the modified equation also has to converge to 0. The first term on the left-hand side of the modified equation is

$$\frac{\langle L'(x^k; \bar{\lambda}) - L'(\bar{x}; \bar{\lambda}), x^k - \bar{x}\rangle}{\|x^k - \bar{x}\|^2},$$

which converges to $\langle L''(\bar{x}; \bar{\lambda})h, h\rangle$. Since the direction h was seen to be critical, this term is strictly positive. We shall now obtain the sought for contradiction by showing that the remaining term on the left-hand side of the modified equation is asymptotically nonnegative. This is verified by splitting this term into its equality and inequality parts.

The equality part of the term in question is

$$\frac{\langle g_E'(x^k)(x^k - \bar{x}), \lambda_E^k - \bar{\lambda}_E\rangle}{\|x^k - \bar{x}\|^2},$$

which, due to $g_E'(\bar{x})h = 0$, tends to 0. This argument uses the fact that $\lambda_E^k - \bar{\lambda}_E = \mathcal{O}(x^k - \bar{x})$, which itself is a consequence of the standing hypothesis (23) and the constraint qualification.

Inspecting the inequality term remains. Via Taylor expansion, the latter is

$$\frac{\langle g_I(x^k) - g_I(\bar{x}), \lambda_I^k - \bar{\lambda}_I\rangle}{\|x^k - \bar{x}\|^2} + \mathrm{o}(1), \tag{24}$$

again using $\lambda^k - \bar{\lambda} = \mathcal{O}(x^k - \bar{x})$. The left-hand term of (24) is recast as

$$\frac{\langle g_I(x^k) - v_I^k - g_I(\bar{x}), \lambda_I^k - \bar{\lambda}_I\rangle}{\|x^k - \bar{x}\|^2} + \frac{\langle v_I^k, \lambda_I^k - \bar{\lambda}_I\rangle}{\|x^k - \bar{x}\|^2}, \tag{25}$$

and the second term in (25) tends to 0 due to the standing hypothesis. The first term in (25) is nonnegative, for expanding its nominator gives

$$\langle g_I(x^k) - v_I^k, \lambda_I^k\rangle - \langle g_I(x^k) - v_I^k, \bar{\lambda}_I\rangle - \langle g_I(\bar{x}), \lambda_I^k\rangle + \langle g_I(\bar{x}), \bar{\lambda}_I\rangle.$$

Here the first and the last terms vanish as a consequence of the complementarity condition (3) in the Kuhn–Tucker ensemble (KT) and the perturbed condition (3)

above, while the two terms with the negative signs are themselves negative, again due to the conditions (3) above and in (KT). Indeed, λ_I^k, $\bar{\lambda}_I \in K$ and $g_I(\bar{x}) \in K^0$, $g_I(x^k) - v_I^k \in K^0$ imply $\langle g_I(x^k) - v_I^k, \bar{\lambda}\rangle \leq 0$ and $\langle g_i(\bar{x}), \lambda_I^k\rangle \leq 0$. This settles the case by providing the desired contradiction. ∎

With this observation, we are now ready to state our first result.

LEMMA 7. *Suppose Newton's method for solving (P) via a successive solution of $(T_k(M^k))$ with a choice of matrices M^k generates a sequence of iterates (x^k, λ^k) which converges to the Kuhn–Tucker pair $(\bar{x}, \bar{\lambda})$. Further suppose that $g'(\bar{x})$ has maximal rank and that the second-order sufficient optimality condition is satisfied at $(\bar{x}, \bar{\lambda})$.*

1. *If $M^k \to L''(\bar{x}; \bar{\lambda})$, convergence $(x^k, \lambda^k) \to (\bar{x}, \bar{\lambda})$ is superlinear.*
2. *If $M^k - L''(\bar{x}; \bar{\lambda}) = \mathcal{O}(x^k - \bar{x})$, then convergence $(x^k, \lambda^k) \to (\bar{x}, \bar{\lambda})$ is even quadratic.*

Proof. We observe that with $x^{k+1} = x^k + \Delta x$, and λ^{k+1} the Lagrange multiplier in $(T_k(M^k))$, the quasi-Newton step about the current iterate (x^k, λ^k) may be represented as

(i) $L'(x^{k+1}; \lambda^{k+1}) = u^k$,
(ii) $g_E(x^{k+1}) = v_E^k$,
(iii) $g_I(x^{k+1}) - v_I^k \in K^0$, $\lambda_I^{k+1} \in K$, $\langle g_I(x^{k+1}) - v_I^k, \lambda_I^{k+1}\rangle = 0$,

where the perturbation terms u^k and $v^k = (v_E^k, v_I^k)$ are as follows:

$$\begin{aligned} u^k &= L'(x^{k+1}; \bar{\lambda}) - L'(x^k; \bar{\lambda}) - L''(x^k; \bar{\lambda})(x^{k+1} - x^k) + (L''(\bar{x}; \bar{\lambda}) - M^k)(x^{k+1} - x^k) \\ &\quad + (L''(x^k; \bar{\lambda}) - L''(\bar{x}; \bar{\lambda}))(x^{k+1} - x^k) + (g'(x^{k+1}) - g'(x^k))^*(\lambda^{k+1} - \bar{\lambda}), \\ v^k &= -g(x^{k+1}) + g(x^k) + g'(x^k)(x^{k+1} - x^k). \end{aligned}$$

As we wish to bring in the perturbation Lemma 6 above, we let $\delta_k \to 0$ be the speed of convergence of $(u^k, v^k) \to (0, 0)$; then $(x^k, \lambda^k) - (\bar{x}, \bar{\lambda}) = \mathcal{O}(\delta_k)$ as a consequence of that lemma.

Now observe that $v^k = \mathrm{o}(x^{k+1} - x^k)$ and, similarly, $u^k = \mathrm{o}(x^{k+1} - x^k)$ if we use the hypothesis $M^k - L''(\bar{x}; \bar{\lambda}) = \mathrm{o}(1)$. Altogether, $\delta_k = \mathrm{o}(x^{k+1} - x^k)$. The perturbation lemma therefore implies

$$x^{k+1} - \bar{x} = \mathcal{O}(\delta_k) = \mathrm{o}(x^{k+1} - x^k) = \mathrm{o}(\|x^{k+1} - \bar{x}\| + \|x^k - \bar{x}\|).$$

Similarly, as $g'(\bar{x})$ has maximal rank,

$$\lambda^{k+1} - \bar{\lambda} = \mathcal{O}(\delta_k) = \mathrm{o}(\|x^{k+1} - \bar{x}\| + \|x^k - \bar{x}\|).$$

These estimates prove superlinear convergence.

The argument giving quadratic convergence under the stronger hypothesis in (2) is standard and left to the reader (see, for instance, [11]). ∎

As a consequence of Lemma 7, what remains to be checked is mere convergence of Newton's method under the same regularity hypotheses. Here we shall be able to follow a known line of argument already present in Robinson's approach [32]. Let us consider the *limiting tangent problem*

$$(T_\infty) \qquad \begin{aligned} &\text{minimize} && \langle f'(\bar{x}), d\rangle + \tfrac{1}{2}\langle d, L''(\bar{x}; \bar{\lambda})d\rangle \\ &\text{subject to} && g_E(\bar{x}) + \langle g_E'(\bar{x}), d\rangle = 0, \\ & && g_I(\bar{x}) + \langle g_I'(\bar{x}), d\rangle \in K^0, \end{aligned}$$

whose optimal solution is $\bar{d} = 0$, and for which $\bar{\lambda}$ is a Lagrange multiplier. Observe that the second-order optimality conditions for (T_∞) are identical with those of (P),

so if we adopt the constraint qualification from before and the second-order sufficient optimality condition for (P), they also hold for (T_∞). Using a result obtained by Robinson [32, Theorems 2.3, 3.1], we have the following lemma.

LEMMA 8. *Suppose the second-order sufficient optimality condition for (P) is satisfied at the optimal pair $(\bar{x}, \bar{\lambda})$. Further suppose that $g'(\bar{x})$ has maximal rank. Then, given $\varepsilon > 0$, there exists $\delta > 0$ such that if $\|x^k - \bar{x}\| < \delta$, $\|\lambda^k - \bar{\lambda}\| < \delta$, and $\|M^k - L''(\bar{x}; \bar{\lambda})\| < \delta$, then the tangent problem $(T_k(M^k))$ has a local minimum x^{k+1} and an associated Lagrange multiplier λ^{k+1} satisfying $\|x^{k+1} - \bar{x}\| < \varepsilon$ and $\|\lambda^{k+1} - \bar{\lambda}\| < \varepsilon$.*

Proof. Notice that the tangent subproblem $(T_k(M^k))$ may be considered a perturbed version of the ideal tangent problem (T_∞) in the sense of [32, (2.7)]. Now by assumption $g'(\bar{x})$ has maximal rank, and hence (T_∞) is regular in the sense of [32]. Second, since (P) satisfies the second-order sufficient optimality condition at $(\bar{x}, \bar{\lambda})$, so does (T_∞) at the optimal pair $(0, \bar{\lambda})$. Using [32, Theorem 3.1], there exist neighborhoods N_1 of $\bar{x}$, N_2 of $\bar{\lambda}$, and N_3 of $L''(\bar{x}; \bar{\lambda})$ such that for $x^k \in N_1$, $\lambda^k \in N_2$, and $M^k \in N_3$ the tangent problem $(T_k(M^k))$ has a solution x^{k+1}. We may, in addition, choose N_1 small enough to guarantee that $g'(x^k)$ has maximal rank, and therefore $(T_k(M^k))$ also admits Lagrange multipliers λ^{k+1}.

Now using Theorem 2.3 of the same paper, the set valued operator mapping the datum (x^k, λ^k, M^k) of $(T_k(M^k))$ into the set of possible optimal pairs (x^{k+1}, λ^{k+1}) is upper semicontinuous. By second-order sufficient optimality, $(\bar{x}, \bar{\lambda})$ is locally unique. Therefore, upper semicontinuity translates into the following statement: Given $\varepsilon > 0$, there exists $\delta > 0$ such that if (x^k, λ^k, M^k) is in the δ-neighborhood of $(\bar{x}, \bar{\lambda}, L''(\bar{x}; \bar{\lambda}))$, then any (x^{k+1}, λ^{k+1}) lies in the ε-neighborhood of $(\bar{x}, \bar{\lambda})$. This is just what we claimed. ∎

With these auxiliary results, we are now ready to state our local convergence theorem for Newton's method.

THEOREM 9. *Let $(\bar{x}, \bar{\lambda})$ be a Kuhn–Tucker pair for (P) satisfying the second-order sufficient optimality condition, and suppose $g'(\bar{x})$ has maximal rank. Then there exists $\delta > 0$ such that if $\|x^0 - \bar{x}\| < \delta$, $\|\lambda^0 - \bar{\lambda}\| < \delta$, $\|M^k - L''(\bar{x}; \bar{\lambda})\| < \delta$ for every k, and $M^k \to L''(\bar{x}; \bar{\lambda})$, then the sequence (x^k, λ^k) obtained by successive solution of the tangent subproblems $(T_k(M^k))$ is well defined and converges superlinearly to $(\bar{x}, \bar{\lambda})$. Convergence is even quadratic if $M^k - L''(\bar{x}; \bar{\lambda}) = \mathcal{O}(\|x^k - \bar{x}\| + \|\lambda^k - \bar{\lambda}\|)$.*

Proof. (1) Observe that the perturbation Lemma 6 tells that, due to second-order sufficient optimality, the Kuhn–Tucker conditions for (P) follow a Lipschitz-type behavior with respect to specific perturbations u^k, v^k. Let us quantify this: There exist $\delta_1 > 0$ and $\alpha > 0$, $\beta > 0$ such that if u^k, v^k are sufficiently small in the sense that $\|v^k\|$, $\|u^k\| < \delta_1$ and if x^k, λ^k along with u^k, v^k satisfy (1)–(3) in the perturbation Lemma 6, then $\|x^k - \bar{x}\| + \|\lambda^k - \bar{\lambda}\| < \alpha$, and $\|x^k - \bar{x}\| + \|\lambda^k - \bar{\lambda}\| < \beta(\|u^k\| + \|v^k\|)$.

(2) Let $\delta_3 \leq \min(\alpha, \frac{1}{3\beta})$. According to Lemma 8, there exists $\delta_2 > 0$ such that whenever $\|\hat{x} - \bar{x}\| < \delta_2$, $\|\hat{\lambda} - \bar{\lambda}\| < \delta_2$, and $\|\hat{M} - L''(\bar{x}; \bar{\lambda})\| < \delta_2$, the result (x, λ) of the Newton step with datum $(\hat{x}, \hat{\lambda}, \hat{M})$ satisfies $\|(x, \lambda) - (\bar{x}, \bar{\lambda})\| < \delta_3$.

(3) Choose $\delta_4 > 0$ such that the following five conditions are satisfied. First,

$$\|g(x^1) - g(x^2) - g'(x^2)(x^1 - x^2)\| \leq \frac{1}{24\beta}\|x^1 - x^2\|$$

whenever x^1, $x^2 \in B(\bar{x}, \delta_4)$. Second,

$$\|L'(x^1; \bar{\lambda}) - L'(x^2; \bar{\lambda}) - L''(x^2; \bar{\lambda})(x^1 - x^2)\| \leq \frac{1}{24\beta}\|x^1 - x^2\|$$

whenever x^1, $x^2 \in B(\bar{x}, \delta_4)$. Third, $\delta_4 < 1/24\beta$, and fourth

$$\|L''(x^1; \bar{\lambda}) - L''(\bar{x}, \bar{\lambda})\| \leq \frac{1}{24\beta}$$

whenever $x^1 \in B(\bar{x}, \delta_4)$. Finally,

$$\|(g'(x^2) - g'(x^1))(\lambda^1 - \lambda^2)\| \leq \frac{1}{24\beta}\|\lambda^1 - \lambda^2\|$$

whenever λ^1, $\lambda^2 \in B(\bar{\lambda}, \delta_4)$ and x^1, $x^2 \in B(\bar{x}, \delta_4)$.

(4) Now choose $\delta = \min(\delta_1, \delta_2, \delta_3, \delta_4)$; then the conclusion of the theorem holds. In fact, let (x^k, λ^k, M^k) be the datum of the kth Newton step. As $\delta \leq \delta_1$, an optimal pair (x^{k+1}, λ^{k+1}) exists and satisfies $\|(x^{k+1}, \lambda^{k+1}) - (\bar{x}, \bar{\lambda})\| < \delta_3$.

As in the proof of Lemma 7, let us write the Newton step in the form

(i) $L(x^{k+1}; \lambda^{k+1}) = u^k$,

(ii) $g_E(x^{k+1}) = v_E^k$,

(iii) $g_I(x^{k+1}) - v_I^k \in K^0$, $\lambda_I^{k+1} \in K$, $\langle g_I(x^{k+1}) - v_I^k, \lambda_I^{k+1} \rangle = 0$,

where u^k, v^k have the meaning given there. Then $\delta \leq \delta_2$ and $\delta \leq \delta_1$ and step (1) imply $\|u^k\| \leq \frac{1}{6\beta}\|x^{k+1} - x^k\|$ and $\|v^k\| \leq \frac{1}{6\beta}(\|x^{k+1} - x^k\| + \|\lambda^{k+1} - \lambda^k\|)$. Therefore, step (1) implies

$$\|x^{k+1} - \bar{x}\| \leq \beta\frac{1}{6\beta}(\|x^{k+1} - x^k\| + \|\lambda^{k+1} - \lambda^k\|)$$
$$\leq \frac{1}{6}(\|x^k - \bar{x}\| + \|x^{k+1} - \bar{x}\| + \|\lambda^k - \bar{\lambda}\| + \|\lambda^{k+1} - \bar{\lambda}\|),$$

and similarly for $\|\lambda^{k+1} - \bar{\lambda}\|$. Adding both estimates gives

$$\|(x^{k+1}, \lambda^{k+1}) - (\bar{x}, \bar{\lambda})\| \leq \frac{1}{3}(\|(x^k, \lambda^k) - (\bar{x}, \bar{\lambda})\| + \|(x^{k+1}, \lambda^{k+1}) - (\bar{x}, \bar{\lambda})\|).$$

Therefore,

$$\|(x^{k+1}, \lambda^{k+1}) - (\bar{x}, \bar{\lambda})\| \leq \frac{1}{2}\|(x^k, \lambda^k) - (\bar{x}, \bar{\lambda})\|,$$

and this proves linear convergence of the sequence. This settles the case, since it proves, in particular, that the situation needed to start this argument is reproduced at each step. □

Remarks. (1) Notice that Newton's method $M^k = L(x^k; \lambda^k)$ satisfies hypothesis (2) and therefore converges quadratically.

(2) As is well known, superlinear and quadratic convergence of the primal-dual pair (x^k, λ^k) does not imply superlinear or quadratic convergence of the primal sequence x^k. In order to establish primal superlinear convergence, an extra argument is needed, and this leads to a result in the style of the classical Dennis–Moré characterization of superlinear convergence for unconstrained optimization. The result is similar to [12, Theorem 11.5], and we do not present the details here.

In the same vein, a more thorough analysis of the SSDP method will have to address the following elements not considered here due to the lack of space: a global convergence analysis based on an appropriate merit function and an extension of the known result in polyhedral programming saying that the Hessian of the augmented Lagrangian is positive definite at the optimal pair if the penalty parameter c is properly chosen (cf., for instance, [12, Proposition 12.2]). In particular, in our numerical

tests, we observed this effect, and it should be proven rigorously in order to fully justify our approach via SSDP.

(3) Let us point to the main difference of our approach to the setting [11]. Following Robinson's methods, Bonnans embeds the Newton step for (P) into the Newton step of a suitably formulated variational inequality and then establishes a perturbation Lemma in the style of Lemma 6 but with perturbations based on the variational formulation. Consequently, more (and in fact, too many) perturbations are allowed, and the Lipschitz-type behavior is then established only under polyhedrality.

6. Existing techniques and comparison. In principle, the optimization step in our control design algorithm may be replaced with any optimization technique adapted to deal with LMI constraints. Here we shall compare SSDP to two other methods which we have previously used in robust control design: the augmented Lagrangian method and an approach via concave programming. Numerical experiments based on interior-point methods have been reported by Leibfritz and Mostafa [25, 26] for a different but related type of application. Our own experiments with the interior-point approach will be presented in [3].

A thorough investigation of nonlinear optimization techniques in robust control synthesis should include comparison with existing techniques like the D-K iteration scheme. This has already been addressed in [16], where test examples similar to the ones here were used to compare these approaches. As a result, we observed that the D-K iteration scheme was not at ease with these seemingly innocent cases and very often could not even be started due to the lack of a useful initial controller.

A partially augmented Lagrangian scheme for solving (D) was discussed in [16], and we reproduce it here for the convenience of the reader.

Augmented Lagrangian method.

1. Select an initial penalty parameter $c_0 > 0$, a Lagrange multiplier estimate λ^0, and an initial decision vector x^0 satisfying the LMIs, $\mathcal{A}(x^0) \le 0$.
2. For given c_k, λ^k, and x^k, solve

$$\begin{aligned} &\text{minimize} && L_{c_k}(x; 0, \lambda^k) \\ &\text{subject to} && \mathcal{A}(x) \le 0, \end{aligned} \tag{26}$$

   and let x^{k+1} be the solution to (26).
3. Update the Lagrange multiplier using the first-order update formula

$$\lambda^{k+1} = \lambda^k + c_k \mathcal{B}(x^{k+1}). \tag{27}$$

4. Update the penalty parameter such that $c_{k+1} \ge c_k$, increase k, and go back to step 2.

This scheme is often called the *first-order method of multipliers*. It takes the constraint set $\{x : \mathcal{A}(x) \le 0\}$ as an unstructured set and does not attempt to exploit its special LMI structure, which would require attaching a matrix Lagrange multiplier variable $\Lambda \ge 0$ to the LMI. As a consequence, its rate of convergence is only linear if the penalty parameter $c_k = c$ is held fixed, while superlinear convergence is guaranteed if $c_k \to \infty$. The latter is of minor practical importance due to the inevitable ill-conditioning for large c.

Remark. Aiming at good theoretical local convergence properties, we should certainly avoid the augmented Lagrangian method. To ensure superlinear convergence with fixed large enough c, we have to use second-order methods like the proposed SSDP method. Nevertheless, the augmented Lagrangian method has some merits as

it is robust in practice and, similar to the case of SSDP, may be tackled by a series of SDP subproblems if the Newton step called for to solve (26) is suitably convexified. In contrast with the tangent subproblem (T) in SSDP, these SDP subproblems may be solved by primal methods, as Lagrange multipliers Λ are not required. This may be an advantage of the augmented Lagrangian approach since, for instance, in our experiments, a well-implemented primal SDP solver like [20] often outperformed existing primal-dual software, even though the latter is preferred by theory.

Let us finally recall an approach to (D) discussed in [4]. Primarily, this scheme is suited for the feasibility problem (find x such that $\mathcal{B}(x) = 0$, $\mathcal{A}(x) \leq 0$) but may be modified to apply to (D).

Consider (D) with a nonlinear equality constraint of the form $\mathcal{B}(x) = P\tilde{P} - I = 0$ as encountered in our applications. Introducing a slack matrix variable Z, problem (D) may be replaced by the concave program (cf. [4] for a proof)

$$(C)\qquad \begin{aligned} &\text{minimize} && f_c(x) = \gamma + c\ \operatorname{trace}(Z_1 - Z_3^T Z_2^{-1} Z_3) \\ &\text{subject to} && \mathcal{A}(x) \leq 0, \\ & && \mathcal{L}(x) = \begin{pmatrix} Z_1 & Z_3^T & P & I \\ Z_3 & Z_2 & I & \tilde{P} \\ * & * & I & 0 \\ * & * & 0 & I \end{pmatrix} \geq 0. \end{aligned}$$

We may solve (C) by a sequence of subproblems, each of which minimizes the first-order Taylor polynomial of $f_c(x)$ about the inner current iterate x and over the convex set $\{\mathcal{A} \leq 0, \mathcal{L} \geq 0\}$. This procedure is known as the conditional gradient or Frank and Wolfe method. In order to improve its performance, second-order information is at least partially included by approximating the concave second-order term of the objective $f_c(x)$ by a linear underestimate (see [30]). This modification improves convergence but still has the inconvenience of a high CPU cost. Altogether, concave methods cannot compete with the SSDP or augmented Lagrangian techniques, as they are very slow and, due to the slack variable Z, lead to large size problems. We use the concave programming approach in order to check on the quality of our local optimal solutions. In a reasonable number of tests, SSDP did, in fact, terminate with values of γ close to the global optimum. Yet another way to test the quality of the gain γ is to establish a lower bound γ_ℓ for the optimal γ_{opt} by solving (D) without the nonlinear constraint $\mathcal{B} = 0$.

7. Numerical experiments. In this section, two typical test examples are used to compare the SSDP method to the augmented Lagrangian method proposed in [16] for a related situation and a concave programming approach.

7.1. Robust control of a flexible actuator. Consider the unbalanced oscillator described in Figure 3. The plant is built with a cart of weight M, fixed to a vertical plane by a linear spring k and constrained to move only along the z axis. An embedded pendulum with mass m and moment of inertia I is attached to the cart's center of mass and can be rotated in the vertical plane. The cart is submitted to an external disturbance F, and a control torque N is applied to the pendulum to stabilize its movement. The nonlinear equations of motion are

$$(M+m)\ddot{Z} + me\ddot{\vartheta}\cos\vartheta = me\dot{\vartheta}^2\sin\vartheta - kZ + F, \qquad me\ddot{Z}\cos\vartheta + (I+me^2)\ddot{\vartheta} = N,$$

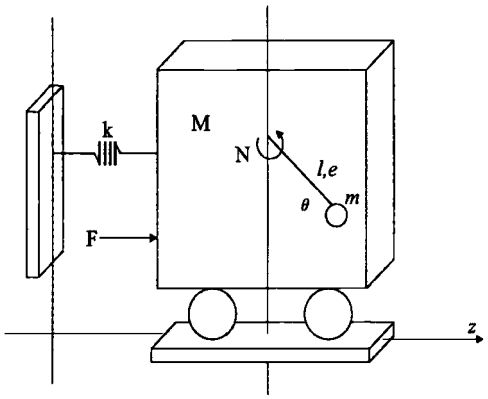


FIG. 3. *Flexible actuator.*

TABLE 1
Behavior of SSDP for the flexible actuator (computations on PC with CPU Pentium II 333 *MHz).*

	SSDP method			Augmented Lagrangian		
Step	γ	$\|P\tilde{P}-I\|_F^2$	c	γ	$\|P\tilde{P}-I\|_F^2$	c
0	7	1.152 e+002	0.5	7	1.152 e+002	0.5
1	4.429	1.935 e−000		3.771	1.085 e+001	
2	2.976	1.529 e+001		2.870	1.156 e+001	
3	1.795	1.717 e−000		2.083	1.297 e+001	
4	1.287	6.214 e−000		1.849	1.415 e+001	
5	1.262	1.762 e−000		1.276	7.169 e−000	
6	1.259	7.276 e−001		1.245	2.615 e−000	
7	1.261	4.679 e−001		1.246	4.716 e−001	
8	1.262	1.526 e−002		1.249	1.274 e−001	2
9	—	2.647 e−004	2	1.251	4.247 e−002	
10		1.796 e−006		1.254	1.676 e−002	
11				—	7.462 e−003	8
12					1.179 e−003	
13					9.584 e−005	32
14					2.145 e−005	
15					1.217 e−006	128

where ϑ and $\dot{\vartheta}$ denote the angular position and velocity of the pendulum and Z, $\dot{Z}$ denote the position and velocity of the cart. We normalize these equations as in [14]:

$$\ddot{\zeta} + \varepsilon\ddot{\vartheta}\cos\vartheta = \varepsilon\dot{\vartheta}^2\sin\vartheta - \zeta + w, \qquad \varepsilon\ddot{\zeta}\cos\vartheta + \ddot{\vartheta} = u,$$

where $[\zeta\dot{\zeta}\vartheta\dot{\vartheta}]^T$ is the new state vector. We assume $\theta_m = \cos\vartheta$ is measured, and we express the remaining nonlinear term in the left-hand equation through the uncertain parameter $\theta_u = \dot{\vartheta}\sin\vartheta$. The parameter block becomes $\Theta = \mathrm{diag}(\theta_m, \theta_u I_3)$. The LFT model of the plant is then derived, and numerical data are given below in order to allow testing of our results with different approaches. Table 1 displays the behavior of the SSDP algorithm. We can see that SSDP achieves good values of γ already after a few iterations. The nonlinear constraints decrease with an approximately linear rate. In practice, one may stop the algorithm whenever γ is no longer reduced over a certain number of iterations and the nonlinear constraint norm is sufficiently small, say, smaller than 10^{-6} or 10^{-7}. The final steps in the table are only for illustration of the asymptotic behavior of the method. Note that the number of decision variables in

TABLE 2
Behavior of modified conditional gradient algorithm for flexible actuator.

	Modified conditional gradient						
Step	γ	$\|P\tilde{P}-I\|_F^2$	c	Step	γ	$\|P\tilde{P}-I\|_F^2$	c
0	7	1.152 e+002	0.5	11		1.725 e−002	512
1	1.295	2.169 e+001		12	1.315	7.642 e−003	
2	1.292	2.576 e+001		13	1.319	2.764 e−003	1024
3	1.302	1.145 e+001	2	14	1.321	9.476 e−004	
4	1.307	5.872 e+000	8	15	—	6.125 e−004	2048
5	1.309	2.057 e+000		16		4.679 e−004	
6	1.311	8.451 e−001		17	1.325	2.762 e−004	
7	—	4.251 e−001	32	18	—	1.927 e−004	
8	1.312	2.745 e−001		19	1.322	2.169 e−004	
9	—	7.567 e−002	128	20	1.324	1.742 e−004	
10		4.571 e−002					

this example was 94. The gain $\gamma_{\text{opt}} = 1.262$ obtained by SSDP was close to the lower estimate $\gamma_\ell = 1.18$ obtained by solving (D) without the constraint $\mathcal{B} = 0$. Note that these results improve on those of the modified conditional gradient which is slower and leads to higher cost values in Table 2.

The numerical data for the flexible actuator LFT plant are

$$P(s) \cong \left(\begin{array}{cccc|cccc|cc} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 5\varepsilon & \varepsilon & -\varepsilon & -\varepsilon & 1 & -0.2 \\ 0 & 0 & 0 & 1.02 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.2 & 0 & 0 & 0 & -\varepsilon & \varepsilon & \varepsilon & 0 & -0.2 & 1 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.84 & 0 & 0 & 0 & -4\varepsilon & -\varepsilon & 0 & \varepsilon & -0.84 & 0.16 \\ 1.23 & 0 & 0 & 0 & -6\varepsilon & 0 & 2\varepsilon & 2\varepsilon & -1.23 & 0.23 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \hline 0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2 \\ \hline 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right),$$

where ε, a coupling parameter, is chosen in this problem to equal 0.01. The vector of regulated variables z consists of three components: z_ζ, z_θ are the damping specifications on ζ, θ; and z_u serves to limit the control activity. The exogenous input w is the external force F.

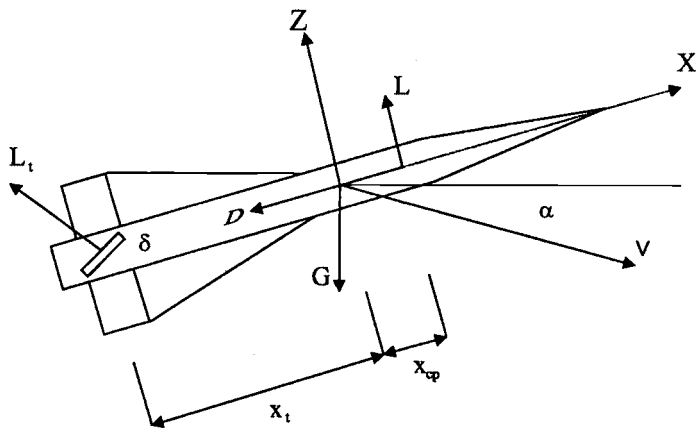


FIG. 4. *Aerodynamic model for air-to-air missile.*

7.2. Robust autopilot of a missile. Consider the missile-airframe control problem illustrated in Figure 4, where the missile is flying with an angle of attack α. The control problem requires that the autopilot generate the elevator deflection δ to maintain the angle of attack α_c called for by the guidance law. The tail-fin actuator is modeled as a first-order system

$$\dot{\delta} = \tau(u - \delta)$$

with time constant $\tau = 1/150$ seconds, so δ itself becomes a state of the system. The nonlinear dynamics of the missile are adopted from [31]:

$$\dot{\alpha} = f\frac{\cos(\alpha/f)}{mV}Z + q, \qquad \dot{q} = fM_\ell/I_y,$$

where m is the mass, $V = M/V_s$ is the speed, I_y is the pitch moment of inertia, $Z = C_Z(\alpha, \delta, M)QS$ is the normal force, $M_\ell = C_m(\alpha, \delta, M)QSd$ is the pitch moment, Q is the dynamic pressure, and S, d reference area and diameter. The normal force and pitch moment aerodynamic coefficients are approximated by third-order polynomials in α and first-order polynomials in δ and M.

Sensor measurements y for feedback include the pitch rate q and α, while the state of the actuator deflection δ is unmeasured. The robust control scheme for the missile autopilot is shown in Figure 5. The time-varying matrix valued parameter is $\Theta = \mathrm{diag}(\theta_m I_4, \theta_u)$, where θ_u is used to translate the nonlinearity in α in the left-hand equation into an LFT with uncertainty. The scheduled parameter θ_m is the variation in Mach number M about nominal flight conditions, $M_0 = 3$. The Mach number is slowly time-varying and is easily measured on-line.

The vector of regulated variables z consists of two components (see Figure 5). The first, z_1, corresponds to a frequency-weighted sensitivity design goal, for tracking error accuracy, while $z_2 = c_\delta\dot{\delta} = c_\delta\tau(u - \delta)$ serves to limit the tail-fin actuator rate $\dot{\delta}$ and indirectly to bound the controller bandwidth in order to avoid trouble with unmodeled flexible modes. The vector of exogenous inputs w includes the command α_c and the pitch rate sensor noise n.

For the problem considered, it is desired to track the step input command $\alpha_c = 20^0$ with a steady state accuracy of 2%, to achieve a rise time of less than 0.3 seconds,

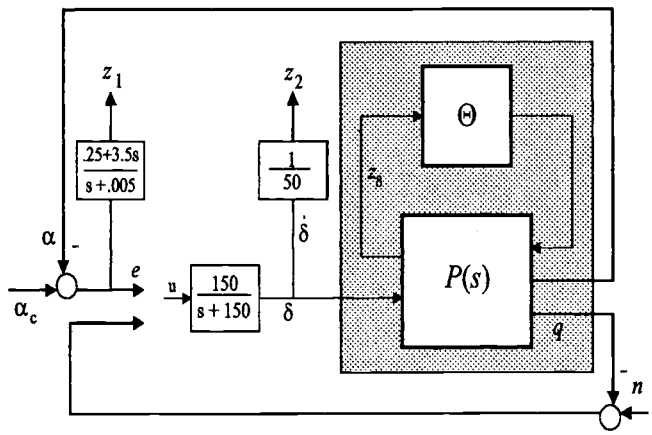


FIG. 5. *Configuration for synthesis.*

TABLE 3
Behavior of SSDP for the missile autopilot (computations on PC with CPU Pentium II 333 *MHz).*

	SSDP method			Augmented Lagrangian		
Step	γ	$\|P\tilde{P}-I)\|_F^2$	c	γ	$\|P\tilde{P}-I\|_F^2$	c
0	7	6.254 e+003	0.25	7	6.254 e+003	0.25
1	1.124	0.476 e−001		1.552	2.147 e+002	
2	0.854	1.876 e−000		1.467	6.345 e−001	
3	0.762	1.287 e−000		0.745	1.827 e−000	
4	0.631	2.655 e−001		0.642	1.845 e−000	
5	0.609	1.425 e−002		0.598	1.475 e−000	
6	0.597	1.721 e−005	2	0.617	7.857 e−001	2
7				0.607	5.749 e−003	8
8				0.596	4.671 e−003	
9				0.597	2.612 e−005	64
10				—	1.682 e−006	

and to limit overshoot to 5% for a wide range of angles of attach ±20 deg and under variations in Mach number ranging from 2 to 4.

For comparison, the numerical data for this LFT model of the missile are reproduced below. Our optimization techniques are then readily applicable, and the results are shown in Table 3. SSDP achieves good values of γ after a few iterations with a similar rate of decrease for the nonlinear constraints. The autopilot example involved 132 decision variables. Again the gain $\gamma_{\text{opt}} = 0.597$ observed at the optimum was close to the lower estimate $\gamma_\ell = 0.57$. The optimal controller obtained by SSDP was then tested in a time domain simulation based on the nonlinear model. The results are shown in Figure 6. The upper curve shows the tracking in α, and the lower curve

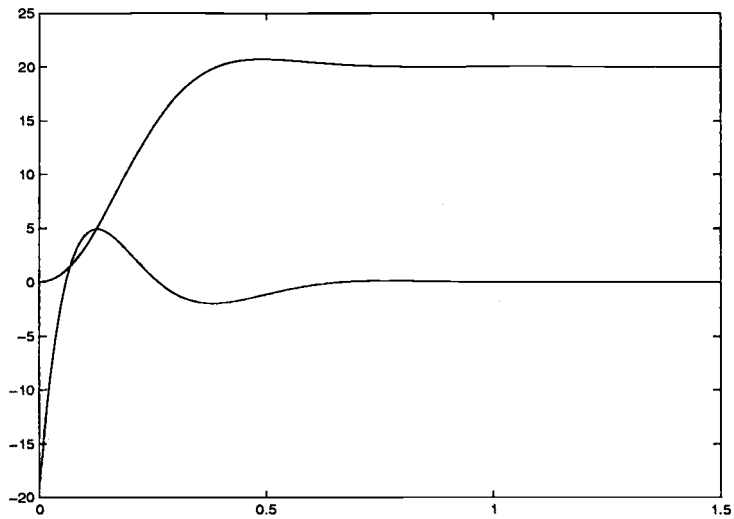


FIG. 6. *Nonlinear simulation for missile autopilot example.*

shows the corresponding elevator deflection δ.

$$P(s) \cong \left(\begin{array}{cccc|cccccc|cc}
-0.876 & 1 & -0.1209 & 0 & 0.201 & 1.185 & 0 & 0 & 0.273 & 0 & 0 & 0 \\
{}^{M}8.9117 & 0 & -130.75 & 0 & 86.32 & 0 & 3 & 1 & 23.46 & 0 & 0 & 0 \\
0 & 0 & -150 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 150 \\
-1 & 0 & 0 & -0.05 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
\hline
0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-0.123 & 0 & -0.017 & 0 & 0.028 & 0 & 0 & 0 & 0.038 & 0 & 0 & 0 \\
0.495 & 0 & -7.264 & 0 & 4.796 & 0 & 0 & 0 & 1.303 & 0 & 0 & 0 \\
1.485 & 0 & -21.79 & 0 & 14.38 & 0 & .5 & 0 & 3.91 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
-0.25 & 0 & 0 & 3.487 & 0 & 0 & 0 & 0 & 0 & 0 & .25 & 0 \\
0 & 0 & -3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 \\
\hline
-1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .01 & 0 & 0
\end{array}\right).$$

Remark. Computational experience with a larger set of typical design examples indicates that the number of iterations (in terms of semidefinite programs) required by SSDP is almost independent of the problem dimension, whereas the CPU of course strongly depends on the efficiency of the SDP solver. As it turns out, in its actual state, the bottleneck of SSDP *is* the SDP solver. The public domain software for SDP we tested could be reliably used to problem sizes of up to 500–1000 decision variables. For larger sizes, the method may fail due to failure of the SDP solver, often already at the stage of finding feasible starting values, or while trying to solve one of the LMI subproblems. Solvers exploiting at best the structure of the problem under consideration may then be required.

Remark. A special type of LMI solver which replaces the SDP by an eigenvalue optimization and uses the bundle method from nonsmooth optimization was presented by Lemaréchal and Oustry [27] and was reported to work well for certain large-size

LMI problems. On the other hand, for large-size problems where most SDP solvers are at ill, the direct approach via interior-point methods may be preferable.

8. Concluding remarks. In this paper, we have developed SSDP, a technique for finding local solutions to robust control design problems. SSDP is an extension of (and is inspired by) SQP, a method in nonlinear optimization known since the late 1970s. Expanding on SQP, SSDP comprises LMI constraints, which are handled explicitly in the course of the algorithm. The method is comfortably implemented with available SDP codes if the Hessian or reduced Hessian is suitably convexified. We found the approach highly reliable (as we demonstrated on a set of test examples), exhibiting local superlinear convergence properties, and applicable to a rich list of problems in robust control theory.

REFERENCES

[1] P. Apkarian and P. Gahinet, *A convex characterization of gain-scheduled H_∞ controllers*, IEEE Trans. Automat. Control, 40 (1995), pp. 853–864. See also p. 1681.

[2] P. Apkarian, P. Gahinet, and G. Becker, *Self-scheduled H_∞ control of linear parameter-varying systems: A design example*, Automatica J. IFAC, 31 (1995), pp. 1251–1261.

[3] P. Apkarian and D. Noll, *A Prototype Primal-Dual LMI-Interior Point Algorithm for Non-convex Robust Control Problems*, working paper.

[4] P. Apkarian and H. D. Tuan, *Robust control via concave minimization—local and global algorithms*, IEEE Trans. Automat. Control, 45 (2000), pp. 299–305.

[5] P. Apkarian and H. D. Tuan, *Concave programming in control theory*, J. Global Optim., 15 (1999), pp. 343–370.

[6] G. J. Balas, J. C. Doyle, K. Glover, A. Packard, and R. Smith, *μ-Analysis and Synthesis Toolbox: User's Guide*, The MathWorks, Natick, MA, 1991.

[7] G. Becker, *Parameter-dependent control of an under-actuated mechanical system*, in Proceedings of the Conference on Decision and Control, New Orleans, LA, 1995, pp. 543–548.

[8] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, London, 1982.

[9] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1995.

[10] P. T. Boggs and J. W. Tolle, *Sequential quadratic programming*, in Acta Numerica, Acta. Numer. 1995, Cambridge University Press, Cambridge, UK, 1995, pp. 1–52.

[11] J. Bonnans, *Local analysis of Newton-type methods for variational inequalities and nonlinear programming*, Appl. Math. Optim., 29 (1994), pp. 161–186.

[12] J. Bonnans, J.-C. Gilbert, C. Lemaréchal, and C. Sagastizábal, *Optimisation Numérique*, Math. Appl. 27, Springer-Verlag, Paris, 1997.

[13] J. David, *Algorithms for Analysis and Design of Robust Controllers*, Ph.D., Thesis, Department of Electrical Engineering, Katholieke Universiteit Leuven, Belgium, 1994.

[14] S. Dussy and L. El Ghaoui, *Measurement-scheduled control for the RTAC problem: An LMI approach. A nonlinear benchmark problem*, Internat. J. Robust Nonlinear Control, 8 (1998), pp. 377–400.

[15] F. Alizadeh, J.-P. A. Haeberly, and M. L. Overton, *Primal-dual interior-point methods for semidefinite programming: Convergence rates, stability and numerical results*, SIAM J. Optim., 8 (1998), pp. 746–768.

[16] B. Fares, P. Apkarian, and D. Noll, *An augmented Lagrangian method for a class of LMI-constrained problems in robust control theory*, Internat. J. Control, 74 (2001), pp. 348–360.

[17] E. Feron, P. Apkarian, and P. Gahinet, *Analysis and synthesis of robust control systems via parameter-dependent Lyapunov functions*, IEEE Trans. Automat. Control, 41 (1996), pp. 1041–1046.

[18] P. Gahinet and P. Apkarian, *A linear matrix inequality approach to H_∞ control*, Internat. J. Robust Nonlinear Control, 4 (1994), pp. 421–448.

[19] P. Gahinet, P. Apkarian, and M. Chilali, *Parameter-dependent Lyapunov functions for real parametric uncertainty*, IEEE Trans. Automat. Control, 41 (1996), pp. 436–442.

[20] P. Gahinet, A. Nemirovski, A. J. Laub, and M. Chilali, *LMI Control Toolbox*, The MathWorks, Natick, MA, 1995.

[21] A. Helmersson, *Methods for Robust Gain-Scheduling*, Ph.D. Thesis, Linköping University, Linköping, Sweden, 1995.

[22] J. W. Helton and O. Merino, *Coordinate optimization for bi-convex matrix inequalities*, in Proceedings of the 36th IEEE Conference on Decision and Control, San Diego, CA, 1997, pp. 3609–3613.

[23] N. J. Higham and S. H. Cheng, *Modifying the inertia of matrices arising in optimization*, Linear Algebra Appl., 275/276 (1998), pp. 261–279.

[24] F. Jarre, *A QQP-Minimization Method for Semidefinite and Smooth Nonconvex Programs*, Technical report, University of Notre Dame, South Bend, IN, 1999.

[25] F. Leibfritz and E. S. Mostafa, *An interior point constrained trust region method for a special class of nonlinear semidefinite programming problems*, SIAM J. Optim., to appear.

[26] F. Leibfritz and E. S. Mostafa, *Trust Region Methods for Solving the Optimal Output Feedback Design Problem*, Forschungsbericht Nr. 00-01, Universität Trier, Trier, Germany, 2000.

[27] C. Lemaréchal and F. Oustry, *Nonsmooth algorithms to solve semidefinite programs*, in Advances in Linear Matrix Inequality Methods in Control, Adv. Des. Control 18, SIAM, Philadelphia, 2000, pp. 57–77.

[28] J. Ly, M. G. Safonov, and R. Y. Chiang, *Real/complex multivariable stability margin computation via generalized Popov multiplier—LMI approach*, in Proceedings of the American Control Conference, Baltimore, MD, 1994, pp. 425–429.

[29] A. Packard, *Gain scheduling via linear fractional transformations*, Systems Control Lett., 22 (1994), pp. 79–92.

[30] P. Pardalos and J. Rosen, *Constrained Global Optimization: Algorithms and Applications*, Lecture Notes in Comput. Sci. 268, Springer-Verlag, Berlin, 1987.

[31] R. T. Reichert, *Robust autopilot design using μ-synthesis*, in Proceedings of the American Control Conference, San Diego, CA, 1990, pp. 2368–2373.

[32] S. Robinson, *Generalized equations and their solutions.* II. *Applications to nonlinear programming*, Math. Programming Stud., 19 (1982), pp. 200–221.

[33] M. G. Safonov, K. C. Goh, and J. H. Ly, *Control system synthesis via bilinear matrix inequalities*, in Proceedings of the American Control Conference, Baltimore, MD, 1994, pp. 45–49.

[34] C. W. Scherer, *A full block $\mathcal{S}$-procedure with applications*, in Proceedings of the IEEE Conference on Decision and Control, San Diego, CA, 1997, pp. 2602–2607.

[35] C. W. Scherer, *Robust mixed control and linear parameter-varying control with full block scaling*, in Advances in Linear Matrix Inequality Methods in Control, Adv. Des. Control 20, SIAM, Philadelphia, 2000, pp. 187–207.

[36] S. Wu and S. Boyd, *Software for Semidefinite Programming and Determinant Maximization Problems with Matrix Structure, User's Guide*, Stanford University, Stanford, CA, 1996.

[37] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control*, Prentice-Hall, Upper Saddle River, NJ, 1996.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 6, pp. 1821–1839

ε-EQUILIBRIA FOR STOCHASTIC GAMES WITH UNCOUNTABLE STATE SPACE AND UNBOUNDED COSTS*

ANDRZEJ S. NOWAK† AND EITAN ALTMAN‡

Abstract. We study a class of noncooperative stochastic games with unbounded cost functions and an uncountable state space. It is assumed that the transition law is absolutely continuous with respect to some probability measure on the state space. Undiscounted stochastic games with expected average costs are considered first. It is shown under a uniform geometric ergodicity assumption that there exists a stationary ε-equilibrium for each $\varepsilon > 0$. The proof is based on recent results on uniform bounds for convergence rates of Markov chains [S. P. Meyn and R. L. Tweedie, *Ann. Appl. Probab.*, 4 (1994), pp. 981–1011] and on an approximation method similar to that used in [A. S. Nowak, *J. Optim. Theory Appl.*, 45 (1985), pp. 591–602], where an ε-equilibrium in stationary policies was shown to exist for the bounded discounted costs. The stochastic game is approximated by one with a countable state space for which a stationary Nash equilibrium exists (see [E. Altman, A. Hordijk, and F. M. Spieksma, *Math. Oper. Res.*, 22 (1997), pp. 588–618]); this equilibrium determines an ϵ-equilibrium for the original game. Finally, new results for the existence of stationary ε-equilibrium for discounted stochastic games are given.

Key words. nonzero-sum stochastic games, approximate equilibria, general state space, long run average payoff criterion

AMS subject classifications. Primary, 90D10, 90D20; Secondary, 90D05, 93E05

PII. S0363012900378371

1. Introduction.

This paper treats nonzero-sum stochastic games with general state space and unbounded cost functions. Our motivation for studying unbounded costs comes from applications of stochastic games to queuing theory and telecommunication networks (see [2, 3, 4, 38]). We assume that the transition law is absolutely continuous with respect to some probability measure on the state space. For the expected average cost case, we impose some stochastic stability conditions, considered often in the theory of Markov chains in general state space [25, 26]. These assumptions imply the so-called ν-geometric ergodicity condition for Markov chains governed by stationary multipolicies of players. Using an approximation technique similar to that in [29], we prove the existence of stationary ε-equilibria in m-person average cost games satisfying the mentioned stability conditions and some standard regularity assumptions. A similar result is stated for discounted stochastic games, but then we do not impose any ergodicity assumptions. To obtain an ε-equilibrium, we apply a recent result by Altman, Hordijk, and Spieksma [4] given for nonzero-sum stochastic games with countably many states. Completely different approximation schemes for stochastic games with a separable metric state space were given by Rieder [39] and Whitt [48]. As in [29], they considered only (bounded) discounted stochastic games.

The passage from finite (or even countably infinite) state space with possibly unbounded cost turns out to be quite a tough problem. In fact, the question of the existence of stationary Nash equilibria in nonzero-sum stochastic games with uncountable state space remains open even in the discounted case. Only some special

*Received by the editors September 18, 2000; accepted for publication (in revised form) August 18, 2001; published electronically March 5, 2002.
http://www.siam.org/journals/sicon/40-6/37837.html

†Institute of Mathematics, Zielona Góra University, Podgórna 50, 65-246 Zielona Góra, Poland (nowak@im.pwr.wroc.pl).

‡INRIA, 2004 Route des Lucioles, B.P.93, 06902 Sophia-Antipolis Cedex, France (altman@sophia.inria.fr).

classes of games are known to possess a stationary Nash equilibrium. For example, Parthasarathy and Sinha [37] proved the existence of stationary Nash equilibria in discounted stochastic games with finitely many actions for the players and state independent nonatomic transition probabilities. Their result was extended by Nowak [30] to a class of uniformly ergodic average cost games. There are papers on certain economic games in which a stationary equilibrium is shown to exist by exploiting a very special transition and payoff structure; see, for example, [5, 7]. Mertens and Parthasarathy [24] reported the existence of nonstationary subgame-perfect Nash equilibria in a class of discounted stochastic games with norm continuous transition probabilities. Some results for nonzero-sum stochastic games with additive reward and transition structure (and, in particular, games with complete information) are given by Küenle [19, 20]. Finally, Harris, Reny, and Robson [13] proved the existence of correlated subgame-perfect equilibria in a class of stochastic games with weakly continuous transition probabilities. We would like to point out that the only papers which deal with nonzero-sum average cost stochastic games with uncountable state space are [20] and [30]. In the zero-sum case, the theory of stochastic games with uncountable state spaces is much more complete. Mertens and Neyman [23] provided some conditions for the existence of value, and Maitra and Sudderth [21, 22] developed a general theory of zero-sum stochastic games with limsup payoffs. Stationary optimal strategies exist in the average cost zero-sum games only if some ergodicity conditions are imposed in the model; see, for example, [31, 34, 15, 17, 18].

In this paper, we make use of an extension of Federgruen's work [11] given by Altman, Hordijk, and Spieksma [4]. Other approaches (based on different assumptions) to nonzero-sum stochastic games with countably many states can be found in [6, 40] and [33]. Some results on sensitive equilibria in a class of ergodic stochastic games are discussed in [33, 16, 35]. To close this brief overview of the existing literature, we note that the theory of stochastic games is much more complete in the case of finite state and action spaces. On one hand, many deep existence theorems are available at the moment; see [23, 22, 44, 45] and some references therein. On the other hand, a theory of algorithms for solving special classes of stochastic games with finitely many states and actions is also well developed [12].

In order to study the uncountable state space, we make use of Lyapunov-type techniques [25] (which also allows us to treat unbounded costs) and of approximations based on discretization. Unfortunately, the discretization to a countable state space does not directly yield a setting for which we can apply the existing theory for stochastic games with a countable state [4]. For example, the Foster (or Lyapunov)-type conditions that have been used for countable Markov chains always involved the requirement of a negative drift outside a finite set, whereas our discretization provides a negative drift outside a countable set. Also, ensuring that the approximating game maintains the same type of ergodic structure as the initial game turned out to be a highly complex problem. The fact that our model allows us to handle unbounded costs is very useful in stochastic games occurring in queueing and in networking applications; see [2, 3, 4, 38], in which bounded costs turn out to be unnatural.

The involved process of discretization given in our paper, which requires assumptions that may be restrictive in some applications, may suggest that, when possible, other equilibrium concepts might be sought instead of the Nash equilibrium. Indeed, some results on the existence of stationary *correlated* equilibria are available at the moment [36, 30, 10]. This type of equilibrium allows for some coordination between players, and the proof of existence is considerably simpler.

This paper is organized as follows. In section 2, we describe our game model. Section 3 is devoted to studying the average cost games. In section 4, we examine discounted stochastic games. An appendix is given in section 5, which contains some auxiliary results on piecewise constant policies in controlled Markov chains.

2. The model and notation. Before presenting the model, we collect some basic definitions and notation. Let $(\Omega, \mathcal{F})$ be a measurable space, where $\mathcal{F}$ is the σ-field of subsets in Ω. By $\mathbb{P}(\Omega)$ we denote the space of all probability measures on $(\Omega, \mathcal{F})$. If Ω is a metric space, then $\mathcal{F}$ is assumed to be the Borel σ-field in Ω. Let $(S, \mathcal{G})$ be another measurable space. We write $P(\cdot|\omega)$ to denote a transition probability from Ω into S. Recall that $P(\cdot|\omega) \in \mathbb{P}(S)$ for each $\omega \in \Omega$, and $P(D|\cdot)$ is a measurable function for each $D \in \mathcal{G}$.

We now describe the game model:

(i) S—the *state space*, endowed with a *countably generated* σ-field $\mathcal{G}$.

(ii) X^i—a *compact metric action space* for player i, $i = 1, 2, \ldots, m$. Let $X = X^1 \times X^2 \times \cdots \times X^m$. We assume that X is given the Borel σ-field.

(iii) $c^i : S \times X \to \mathbb{R}$—a product measurable *cost (payoff) function* for player i.

(iv) $Q(\cdot|s, x)$—a (product measurable) transition probability from $S \times X$ into S, called the *law of motion among states*.

We assume that actions are chosen by the players at discrete times $k = 1, 2, \ldots$. At each time k, the players observe the current state s_k and choose their actions independently of one another. In other words, they select a vector $x_k = (x_k^1, \ldots, x_k^m)$ of actions, which results in a cost $c^i(s_k, x_k)$ at time k incurred by player i, and in a transition to a new state, whose distribution is given by $Q(\cdot|s_k, x_k)$. Let $H_1 = S$ and let $H_n = S \times X \times S \times X \times \cdots \times S$ ($2n - 1$ factors) be the space of all n-stage histories of the game, endowed with the product σ-field. A randomized *policy* γ^i for player i is a sequence $\gamma^i = (\gamma_1^i, \gamma_2^i, \ldots)$, where each γ_n^i is a (product measurable) transition probability $\gamma_n^i(\cdot|h_n)$ from H_n into X^i. The *class of all policies* for player i will be denoted by Γ^i. Let U^i be the set of all transition probabilities u^i from S into X^i. A *Markov policy* for player i is a sequence $\gamma^i = (u_1^i, u_2^i, \ldots)$, where $u_k^i \in U^i$ for every k. A Markov policy γ^i for player i is called *stationary* if it is of the form $\gamma^i = (u^i, u^i, \ldots)$ for some $u^i \in U^i$. Every stationary policy $(u^i, u^i, \ldots)$ for player i can thus be identified with $u^i \in U^i$. Denote by $\Gamma = \prod_{i=1}^m \Gamma^i$ the *set of all multipolicies*, and by U the *subset of stationary multipolicies*. Let $H = S \times X \times S \times X \times \cdots$ be the space of all infinite histories of the game, endowed with the product σ-field. For any $\gamma \in \Gamma$ and every initial state $s_1 = s \in S$, a probability measure P_s^γ and a stochastic process $\{S_k, X_k\}$ are defined on H in a canonical way, where the random variables S_k and X_k describe the state and the action, respectively, chosen by the players on the kth stage of the game (see Proposition V.1.1 in [28]). Thus, for each initial state $s \in S$, any multipolicy $\gamma \in \Gamma$, and any finite horizon n, the *expected n-stage cost* of player i is

$$J_n^i(s, \gamma) = E_s^\gamma \left(\sum_{k=1}^n c^i(S_k, X_k) \right),$$

where E_s^γ means the expectation operator with respect to the probability measure P_s^γ. (Later on we make an assumption on the functions c^i that assures that all the expectations considered in this paper are well defined.)

The *average cost per unit time* to player i is defined as

$$J^i(s, \gamma) = \limsup_{n \to \infty} J_n^i(s, \gamma)/n.$$

If β is a fixed real number in $(0,1)$, called the *discount factor*, then the *expected discounted cost* to player i is

$$D^i(s,\gamma) = E_s^\gamma \left(\sum_{k=1}^{\infty} \beta^{k-1} c^i(S_k, X_k) \right).$$

For any multipolicy $\gamma = (\gamma^1, \ldots, \gamma^m) \in \Gamma$ and a policy σ^i for player i, we define (γ^{-i}, σ^i) to be the multipolicy obtained from γ by replacing γ^i with σ^i.

Let $\varepsilon \geq 0$. A multipolicy γ is called an ε-equilibrium for the average cost stochastic game if for every player i and any policy $\sigma^i \in \Gamma^i$,

$$J^i(s,\gamma) \leq J^i(s,(\gamma^{-i},\sigma^i)) + \varepsilon.$$

We similarly define ε-equilibria for the expected discounted cost games. Of course, a 0-equilibrium will be called a Nash equilibrium.

To ensure the existence of ε-equilibrium strategies for the players in the stochastic game, we will accept some regularity conditions on the primitive data, and in the average expected cost case we will also impose some general Lyapunov stability assumptions on the transition structure.

In both the discounted and average cost cases, we make the following assumptions.

C1: For each player i and $s \in S$, $c^i(s,\cdot)$ is continuous on X. Moreover, there exists a measurable function $\nu : S \to [1,\infty)$ such that

$$L \stackrel{\text{def}}{=} \sup_{s\in S, x\in X, i=1,\ldots,m} \frac{|c^i(s,x)|}{\nu(s)} < \infty. \tag{2.1}$$

C2: There exists a probability measure $\varphi \in \mathbb{P}(S)$ such that

$$Q(B|s,x) = \int_B z(s,t,x)\varphi(dt)$$

for each $B \in \mathcal{G}$ and $(s,x) \in S \times X$. Moreover, we assume that if $x_n \to x_0$ in X, then

$$\lim_{n\to\infty} \int_S |z(s,t,x_n) - z(s,t,x_0)|\nu(t)\varphi(dt) = 0,$$

where ν was defined above (2.1).

Remark 2.1. Let w be a measurable function such that $1 \leq w(s) \leq \nu(s) + \delta$ for all $s \in S$ and for some $\delta > 0$. If $x_n \to x_0$ in X as $n \to \infty$, then

$$\int_S |z(s,t,x_n) - z(s,t,x_0)|w(s)\varphi(dt) \to 0.$$

This follows from **C2**, since $\nu \geq 1$ implies that

$$\int_S |z(s,t,x_n) - z(s,t,x_0)|\varphi(dt) \to 0.$$

3. The undiscounted stochastic game. To formulate our further assumptions, we introduce some helpful notation. Let $s \in S$, $u = (u^1, \dots, u^m) \in U$. We set

$$c^i(s,u) = \int_{X^1} \cdots \int_{X^m} c^i(s, x^1, \dots, x^m) u^1(dx^1|s) \cdots u^m(dx^m|s),$$

and, for any set $D \in \mathcal{G}$, we set

$$Q(D|s,u) = \int_{X^1} \cdots \int_{X^m} Q(D|s, x^1, \dots, x^m) u^1(dx^1|s) \cdots u^m(dx^m|s).$$

By $Q^n(\cdot|s,u)$, we denote the n-step transition probability induced by Q and the multipolicy $u \in U$.

C3 (Drift inequality): Let $\nu : S \to [1, \infty)$ be some given measurable function. There exists a set $C \in \mathcal{G}$ such that ν is bounded on C and for some $\xi \in (0,1)$ and $\eta > 0$ we have

$$\int_S \nu(t) Q(dt|s,x) \le \xi \nu(s) + \eta 1_C(s)$$

for each $(s,x) \in S \times X$. Here 1_C is the characteristic function of the set C.

C4: There exist a $\lambda \in (0,1)$ and a probability measure ζ concentrated on the set C such that

$$Q(D|s,x) \ge \lambda \zeta(D)$$

for any $s \in C$, $x \in X$ and for each measurable set $D \subset C$.

For any measurable function $w : S \to R$, we define the ν-weighted norm as

$$\|w\|_\nu = \sup_{s \in S} \frac{|w(s)|}{\nu(s)}.$$

We write L_ν^∞ to denote the Banach space of all measurable functions w for which $\|w\|_\nu$ is finite.

Condition **C3** implies that, outside a set C, the function ν decreases under any stationary multipolicy u; i.e.,

$$E_s^u \left(\nu(S_{k+1}) - \nu(S_k)|S_k\right) \le -(1-\xi)\nu(S_k) \le -(1-\xi). \tag{3.1}$$

This is known as a drift condition. If (i) the state space is countable, (ii) the set C is finite, and (iii) the state space is communicating under a stationary policy u, then (3.1) implies that the Markov chain (when using u) is ergodic. (This is the well known Foster criterion for ergodicity; see, e.g., [27].)

In the uncountable infinite state space, the same drift condition should be used to obtain the ergodicity condition. However, the finiteness of the set C is replaced by a weaker assumption. Namely, C has to be a *small* set or a *petite set* [25]; condition **C4** is a simple sufficient condition for the set C to be small.

Beyond the ergodicity of the Markov chain $\{S_k\}$ under a stationary multipolicy, Foster-type criteria (i.e., conditions **C3**–**C4**) also ensure the finiteness of the expectation $E_s^u \nu(S_k)$ in steady state, as well the finiteness of the expected cost $E_s^u w(S_k)$

for every potential cost function $w \in L_\nu^\infty$; moreover, they provide a geometric rate of convergence of the expected costs at time k to the steady state cost for $w \in L_\nu^\infty$. These statements will be made precise below.

Note that **C3**–**C4** provide uniform conditions for ergodicity; i.e., ξ, ζ, C, and λ do not depend on the actions (or on the policies). This will be needed in order for approximating games (with countable state space) to have stationary Nash equilibria [4].

LEMMA 3.1. *Assume* **C3**–**C4**. *Then the following properties hold.*

C5: *For every $u \in U$, the corresponding Markov chain is aperiodic and ψ_u-irreducible for some σ-finite measure ψ_u on $\mathcal{G}$. (The latter condition means that if $\psi_u(D) > 0$ for some set $D \in \mathcal{G}$, then the chance that the Markov chain (starting at any $s \in S$ and induced by u) ever enters D is positive.) Thus the state process $\{S_n\}$ is a positive recurrent Markov chain with the unique invariant probability measure denoted by π_u.*

C6: *For every stationary multipolicy u,*

(a)

$$\int_S \nu(s)\pi_u(ds) < \infty.$$

(b) *$\{S_n\}$ is ν-uniformly ergodic; that is, there exist $\theta > 0$ and $\alpha \in (0,1)$ such that*

$$\left| \int_S w(t)Q^n(dt|s,u) - \int_S w(t)\pi_u(dt) \right| \le \nu(s)\|w\|_\nu \theta\alpha^n$$

for every $w \in L_\nu^\infty$ and $s \in S$, $n \ge 1$.

Proof. **C3**–**C4** imply that for any stationary u, the chain is positive Harris recurrent (see Theorem 11.3.4 in [25]). It is thus ψ_u-irreducible (see Chapter 9 of [25]). The aperiodicity (and, in fact, strong aperiodicity) follows from condition **C4** (see [25, p. 116]). This establishes **C5**. **C6** follows from Theorem 2.3 in [26]. □

Remark 3.1. From Lemma 3.1 it follows that for any player i and $u \in U$ we have

$$J^i(u) := \int_S c^i(s,u)\pi_u(ds) = J^i(s,u);$$

that is, the expected average cost of player i is independent of the initial state. Theorem 2.3 in [26] implies that the constants α and θ in Lemma 3.1 depend only on ξ, η, λ, and $\nu_C = \sup_{s\in C} \nu(s)$ (and, in particular, they do not depend on u). **C1**, **C3**, and **C4** imply that the expected costs considered in this section are well defined for any multipolicy $\gamma \in \Gamma$; see [34] or [14].

In what follows, whenever we assume **C1**–**C4**, we shall take the same function ν in **C1** as in **C3**. We are now ready to state our first main result.

THEOREM 3.1. *Consider an undiscounted stochastic game satisfying* **C1**–**C4**. *Then for any $\varepsilon > 0$ there exists a stationary ε-equilibrium.*

The proof of this result is based on an approximation technique and consists of several steps which will be described later on. Before proving the result, we briefly mention the approach and the steps we are using, the difficulties, and the way we overcome these difficulties.

Basic idea behind the proof. Our basic goal is to approximate our game by a sequence of m-person games with countable state spaces and compact action spaces

and which have equilibria in stationary policies; based on such approximating games, we shall construct a stationary policy which is an ϵ-equilibrium for the original game. The basic idea here is similar to the one already used in [29] for the problem with discounted cost. However, the situation here is much more involved; indeed, in the discounted case one does not need to bother about the ergodic structure of the approximating games in order to show that they possess equilibrium in stationary policies. Here, in contrast, we need to carefully construct the approximating games so as to ensure that they not only have the required ergodic property but also are uniform ergodic and have some additional "good" properties for the cost. Our first step in the proof will be to construct such approximating games, which will also satisfy conditions **C1**–**C4**. The function ν, as well as the other objects that appear in these assumptions, will be approximated as well. (We will have to show, for example, that the approximation of ξ is indeed within $(0, 1)$, etc.) The approximation of the game in a way that allows conditions similar to **C1**–**C4** to hold is done in the next two subsections.

Properties similar to **C2**–**C4** were used in [4] to establish the existence of equilibria in stationary policies for games with countable state space; the properties imply, for example, that the costs are continuous in the policies. Unfortunately, the counterpart of property **C4** that is used to establish ergodicity in the literature of countable state Markov chains (or for Markov decision processes, or for Markov games) requires that the set C that appears in conditions **C3**–**C4** be finite. Unfortunately, we *were not able to* come up with a direct approximation scheme for which C is finite. To overcome this problem, we first use some results from [26] to obtain uniform ergodicity results for the approximating chains. Using a key theorem from [41], this will be shown to imply that there exist some function (instead of the original approximation of ν) and constants for which properties **C3**–**C4** hold and for which C is a singleton. This is done in subsection 3.3.

3.1. Transition operators and their ν-weighted norms. If $f \in L_\nu^\infty$ and σ is a finite signed measure on $(S, \mathcal{G})$, then for convenience we set

$$\sigma(f) = \int_S f(s)\sigma(ds),$$

provided that this integral exists. Let P_1 and P_2 be transition subprobabilities from S into S. Define

$$\|P_1 - P_2\|_\nu = \sup_{s\in S} \sup_{|f|\le\nu} \frac{|P_1(f|s) - P_2(f|s)|}{\nu(s)}. \tag{3.2}$$

We will also use the definition (3.2) in the case in which P_1 and P_2 are probability measures on $(S, \mathcal{G})$, or when one of them is zero. Note that if $P_2 = 0$ and P_1 is a transition probability, then it follows from (3.2) that

$$\|P_1\|_\nu = \sup_{s\in S} \frac{P_1(\nu|s)}{\nu(s)}.$$

If P_1 and P_2 are transition probabilities and $\|P_1 - P_2\| < \infty$, then $P_1 - P_2$ induces a bounded linear operator from L_ν^∞ into itself, and $\|P_1 - P_2\|_\nu$ is its operator norm (see Lemma 16.1.1 in [25]).

We now come back to our game model and accept the following notation. For any $u \in U$, we use $Q(u)$ to denote the operator on L_ν^∞ defined by $Q(u)f(s) = Q(f|s, u)$,

$s \in S$, and $f \in L^\infty_\nu$. By **C3**, we have

$$\|Q(u)\|_\nu \le \xi + \eta. \tag{3.3}$$

Clearly, (3.3) implies that $Q(u)$ is (under condition **C3**) a bounded linear operator from L^∞_ν into itself. By $\Pi(u)$ we denote the invariant probability measure operator given by

$$\Pi(u)f = \pi_u(f),$$

where π_u is the invariant probability measure for $Q(\cdot|s,u)$, $u \in U$, and $f \in L^\infty_\nu$.

3.2. Approximating games. We define Γ_A to be the class of stochastic games that "resemble" stochastic games with countably many states and can be used to approximate the original game. The games in Γ_A will depend on some parameter $\delta > 0$. The transition probability in a game belonging to Γ_A is denoted by Q_δ, and the cost function of player i is denoted by c^i_δ.

We introduce some notation:

- $\mathbb{N}$—the set of positive integers,
- $C(X)$—the Banach space of all continuous functions on X, endowed with the supremum norm $\|\cdot\|$,
- $L^1_\nu = L^1_\nu(S, \mathcal{G}, \varphi)$—the Banach space of measurable functions $f : S \to \mathbb{R}$ such that $\int_S |f(s)|\nu(s)\varphi(ds) < \infty$.

We assume that each game $G_\delta \in \Gamma_A$ corresponds to some sequences $\{Y_n\}$, $\{c^i_n\}$, $\{z_n\}$, and $\{\nu_n\}$, where n belongs to some subset $\mathbb{N}_1 \subset \mathbb{N}$ and $\{Y_n\}$ is a measurable partition of the state space such that $Y_n \subset C$ or $Y_n \subset S \setminus C$ for each $n \in \mathbb{N}_1$ (the set C is introduced in assumption **C3**),

$$c^i_\delta(s,x) = c^i_n(x), \quad \text{and} \quad Q_\delta(B|s,x) = \int_B z_n(t,x)\varphi(dt)$$

for all $s \in Y_n$, $x \in X$, and $n \in \mathbb{N}_1$. Moreover, ν_n are rational numbers and $\nu_n \ge 1$ for all $n \in \mathbb{N}_1$. Define $\nu_\delta(s) \stackrel{\text{def}}{=} \nu_n$ if $s \in Y_n$.

We will show that for each $\delta > 0$ it is possible to construct a game G_δ such that $c^i_n \in C(X)$ and $z_n(\cdot,x) \in L^1_\nu$ while $z_n(s,\cdot) \in C(X)$ for all $n \in \mathbb{N}_1$, $x \in X$, and $s \in S$.

Because in our approximation we need to preserve (in some sense) condition **C4**, we consider the following subset $\Delta \subset L^1_\nu$: $\phi \in \Delta$ if and only if ϕ is a density function such that

$$\int_D \phi(s)\varphi(ds) \ge \lambda\zeta(D) \tag{3.4}$$

for each $D \in \mathcal{G}$ such that $D \subset C$. Our assumption **C4** implies that $\Delta \ne \emptyset$. It is obvious that Δ is *convex*. Suppose that $\phi_n \in \Delta$ and $\phi_n \to \phi \in L^1_\nu$. Since $\nu \ge 1$, then $\phi_n \to \phi$ in L^1. By Scheffe's theorem, ϕ is a density function. Moreover, ϕ satisfies (3.4). Thus, we have shown that Δ is a *closed* and *convex* subset of L^1_ν.

Let V be the space of all continuous mappings from X into Δ with the metric ρ defined by

$$\rho(\phi_1,\phi_2) = \max_{x\in X} \int_S |\phi_1(x)(s) - \phi_2(x)(s)|\nu(s)\varphi(ds). \tag{3.5}$$

Since $\mathcal{G}$ is countably generated, L_1 is separable. As in [47, Theorem I.5.1], we can prove the following.

LEMMA 3.2. *V is a complete separable metric space.*

Note that the proof of Theorem I.5.1 in [47] makes use of the convexity of the range space of the continuous mappings involved. In our case, the range space Δ is also convex.

For each $s \in \mathbf{S}$, the transition probability density z of the original game induces elements $\phi(s,\cdot)$ of V by

$$\phi(s,x) = z(s,\cdot,x).$$

From the product measurability of z on $S \times S \times X$, it follows that $s \to \phi(s,\cdot)$ is a measurable mapping from S into V.

We introduce the following notation:

- $\{\phi_k\}$—a countable dense subset of V (see Lemma 3.2),
- $\{c_k\}$—a countable dense set in $C(X)$,
- $\{r_k\}$, $r_k \geq 1$, where $\{r_k\}$ is the set of all rational numbers satisfying $r_k \geq 1$.

Let $0 < \delta < 1$ be fixed. Define for any $k, k_1, \ldots, k_m, l$

$$B(k,k_1,\ldots,k_m,l) = \left\{ s \in S : \rho(\phi(s,\cdot),\phi_k) + \sum_{i=1}^{m} \left\| c^i(s,\cdot) - c_{k_i} \right\| + |\nu(s) - r_l| < \delta \right\}.$$

Let τ be a (fixed) one-to-one correspondence between $\mathbb{N}$ and $\mathbb{N} \times \cdots \times \mathbb{N} = \mathbb{N}^{m+2}$. Define $T_n \overset{\text{def}}{=} B(\tau(n))$, $n \in \mathbb{N}$. Next, set $\bar{Y}_1 \overset{\text{def}}{=} T_1$ and $\bar{Y}_k \overset{\text{def}}{=} T_k - \cup_{j<k} \bar{Y}_j$ for $k \geq 2$.

Let $\{Y_n\}$ be the enumeration of all nonempty sets $\bar{Y}_k$. Clearly, $\{Y_n\}$ is a measurable countable partition of the state space, and n belongs to some $\mathbb{N}_1 \subset \mathbb{N}$.

If necessary, we can modify (trivially) this partition in such a way that $Y_n \subset C$ or $Y_n \subset S \setminus C$ for each n. Note that for each $n \in \mathbb{N}_1$ and each set Y_n there correspond some $z_n \in V$ and $c_n^i \in C(X)$, so that we obtain a game $G_\delta \in \Gamma_A$. Moreover, we have

$$\rho(\phi(s,\cdot), z_n) < \delta$$

($\phi(s,x) = z(s,\cdot,x)$, by definition) for each $n \in \mathbb{N}_1$ and $s \in Y_n$. This implies that

$$\|Q(u) - Q_\delta(u)\|_\nu < \delta \tag{3.6}$$

for every $u \in U$. Next, we have

$$\left\| c^i(s,\cdot) - c_n^i \right\| < \delta$$

for each $n \in \mathbb{N}_1$ and $s \in Y_n$. If we set $c_\delta^i(s,x) = c_n^i(x)$ for $s \in Y_n, x \in X$, we obtain

$$\sup_{s \in S} \sup_{x \in X} |c^i(s,x) - c_\delta^i(s,x)| \leq \delta. \tag{3.7}$$

We also have

$$|\nu(s) - \nu_\delta(s)| < \delta \tag{3.8}$$

for every $s \in S$.

3.3. Equivalence with a game with a countable state space. Next, we shall show that the G_δ game has an equilibrium in the class of stationary multipolicies. This will be done in the proof of the following lemma.

LEMMA 3.3. *Assume that the stochastic game satisfies* **C1**–**C4** *and* $\xi_\delta \stackrel{\text{def}}{=} 3\delta+\xi < 1$. *Then*

(i) *the game* G_δ *satisfies* **C3** *with* ξ *and* ν *replaced by* ξ_δ *and* ν_δ, *respectively, and it satisfies* **C4**; *moreover,*

(ii) *it has a Nash equilibrium in the class of stationary multipolicies.*

Proof. (i) From (3.6), it follows that

$$Q_\delta(\nu|s,x) \le Q(\nu|s,x) + \delta\nu(s)$$

for every $s \in S$ and $x \in X$. Since **C3** holds for the original game, this implies that

$$Q_\delta(\nu|s,x) \le (\delta+\xi)\nu(s) + \eta 1_C(s). \tag{3.9}$$

From (3.8) and (3.9), we conclude that

$$Q_\delta(\nu_\delta|s,x) \le \delta + (\delta+\xi)\delta + (\delta+\xi)\nu_\delta(s) + \eta 1_C(s).$$

Hence

$$Q_\delta(\nu_\delta|s,x) \le \xi_\delta\nu_\delta(s) + \eta 1_C(s) \tag{3.10}$$

for every $s \in S$ and $x \in X$; i.e., a condition of the type **C3** holds. Condition **C4** follows from the construction of Δ (above (3.4)).

(ii) Consider the approximating games under the further assumption that every player i restricts to the class U_0^i of policies that are piecewise constant: u^i belongs to U_0^i if and only if $s \to u^i(\cdot|s)$ is constant on each set Y_n of the partition $\{Y_n\}$ of S. Denote by U_0 the set of all stationary piecewise constant multipolicies. Every game G_δ with the above restriction is equivalent to a stochastic game denoted by $\bar{G}$ with the countable state space $\mathbb{N}_1$ (defined in our approximation procedure). Because every stationary multipolicy in $\bar{G}$ corresponds to a multipolicy in U_0, we will use U_0 also to denote the set of *all stationary multipolicies* in $\bar{G}$. The cost functions in $\bar{G}$ are $c_n^i \in C(X)$, where $n \in \mathbb{N}_1$. The transition probabilities in $\bar{G}$ are given by

$$P_{mn}(u) = Q_\delta(Y_n|s,u)$$

for all $s \in Y_m$, $u \in U_0$, and $m,n \in \mathbb{N}_1$. Let $P(u)$ denote the transition probability matrix corresponding to any $u \in U_0$. Finally, the piecewise constant function ν_δ induces a function $\mu : \mathbb{N}_1 \to [1,\infty)$ by $\mu(n) = \nu_\delta(s), s \in Y_n, n \in \mathbb{N}_1$. (Sometimes we will identify μ with the column vector, and $\mu(n)$ with its nth coordinate.)

Fix δ such that $\xi_\delta < 1$. Applying Lemma 3.1 to the game G_δ, we conclude that it satisfies **C5**. By part (i) and Theorem 2.3 in [26], this game also satisfies **C6** (with possibly different constants θ_1 and α_1). A simple translation of **C5** to the game $\bar{G}$ with countably many states says that for any $u \in U_0$ the Markov chain with the transition probability matrix $P(u)$ has a single ergodic class and is aperiodic. On the other hand, a translation of **C6** and the fact that $\|Q_\delta(u)\|_{\nu_\delta} \le \xi_\delta + \eta$ (which follows from (3.10)) mean that the Markov chain is μ-uniform geometric ergodic; see [9, 41]. By Key Theorem II and Lemma 5.3(ii) in [41], there exist a nonempty finite set $M \subset \mathbb{N}_1$, a function $\tilde{\mu} : \mathbb{N}_1 \to [1,\infty)$, and some $b \in (0,1)$ such that

$$\sum_{n \notin M} P_{kn}(u)\tilde{\mu}(n) \le b\tilde{\mu}(k) \tag{3.11}$$

for every $k \in \mathbb{N}_1$ and $u \in U_0$. This property is called the $\tilde{\mu}$-uniform geometric recurrence (see [8, 9, 41]) and is Assumption A2(1) in [4]. The function $\tilde{\mu}$ is given by

$$\tilde{\mu}(k) = \mu(k) + \sup_{u \in U_0} \left(\sum_{n=1}^{\infty} {}_M P^n(u) \mu \right)(k),$$

where $k \in \mathbb{N}_1$ (${}_M P(u)$ is the matrix $P(u)$ in which we replace the columns corresponding to states $m \in M$ by zeros); see [4, pp. 99–100]. Note that this new function $\tilde{\mu}$ is μ-bounded (i.e., $\sup_{n \in \mathbb{N}_1} \tilde{\mu}(n)/\mu(n) < \infty$) and vice versa. Indeed, $\mu(k) \le \tilde{\mu}(k)$ for each k. On the other hand, by (3.11), we have $({}_M P(u)\tilde{\mu})(k) \le b\tilde{\mu}(k)$ for any $k \in \mathbb{N}_1$ and $u \in U_0$. Hence

$$\tilde{\mu}(k) = \mu(k) + \sup_{u \in U_0} \left(\sum_{n=1}^{\infty} {}_M P^n(u) \mu \right)(k) \le \mu(k) + \sup_{u \in U_0} ({}_M P(u)\tilde{\mu})(k) \le \mu(k) + b\tilde{\mu}(k).$$

Thus, $\tilde{\mu}(k) \le \mu(k)/(1-b)$ for every $k \in \mathbb{N}_1$. Since $\tilde{\mu}$ is μ-bounded and vice versa, this implies the $\tilde{\mu}$-continuity of the immediate costs (recall **C1**) and of the transition probabilities (recall **C2** and Remark 2.1). This is Assumption 1^* in [4]. Since both assumptions 1^* as well as 2(1) in [4] hold, it follows that the game $\bar{G}$ has a stationary Nash equilibrium $u^* \in U_0$, and consequently G_δ has a stationary equilibrium (also denoted by u^*) in the class U_0 of all stationary piecewise constant multipolicies. It now follows from Lemma 5.1 in the appendix that u^* is a Nash equilibrium for the game G_δ in the class U of all stationary multipolicies. □

3.4. Uniform convergence of the steady state probabilities and costs, and proof of the main result. Let $J^i_\delta(u)$ be the expected average cost for player i in the game G_δ when a stationary multipolicy u is used (see Remark 3.1 and Lemma 3.3(i)).

Let $Q_\delta(u)$ and $\Pi_\delta(u)$ denote the transition probability and the invariant probability measure operators under any stationary multipolicy $u \in U$ in the approximating game.

LEMMA 3.4. *Under* **C1**–**C4**,

(i)

$$\lim_{\delta \to 0} \|\Pi_\delta(u) - \Pi(u)\|_\nu = 0$$

uniformly in $u \in U$.

(ii)

$$\lim_{\delta \to 0} \left| J^i_\delta(u) - J^i(u) \right| = 0$$

uniformly in $u \in U$.

Proof. (i) If $\xi_\delta = 3\delta + \xi < 1$, then (by Lemma 3.3) the games G_δ satisfy **C4** and **C3**, with ξ replaced by ξ_δ. By (3.10), we have $\|Q_\delta(u)\|_{\nu_\delta} \le \xi_\delta + \eta$ for all $u \in U$. This and Theorem 2.3 in [26] (applied to the games G_δ) imply that there exists a δ_0 such that

$$\sup_{\delta \le \delta_0} \sup_{u \in U} \|\Pi_\delta(u)\|_{\nu_\delta} < \infty.$$

Hence

$$K_0 \stackrel{\text{def}}{=} \sup_{\delta \le \delta_0} \sup_{u \in U} \|\Pi_\delta(u)\|_\nu < \infty. \tag{3.12}$$

The rest of the proof is an adaptation of the proof of Proposition 1 in [42]. By Lemma 3.1 and Remark 3.1, there exist some $\theta > 0$ and $\alpha \in (0,1)$ such that

$$\sup_{u\in U} \|Q^n(u) - \Pi(u)\|_\nu \le \theta\alpha^n$$

for every n. Hence, there exists an n_0 such that

$$\sup_{n\ge n_0}\sup_{u\in U}\left\|\frac{1}{n}\sum_{i=0}^{n-1} Q^i(u) - \frac{1}{n}(I - Q(u)) - \Pi(u)\right\|_\nu < 1;$$

$Q^0 = I$ is the identity operator. Therefore for each $n \ge n_0$ there exists a ν-bounded transition operator

$$\Phi_n(u) \stackrel{\text{def}}{=} \left(I + \Pi(u) - \frac{1}{n}\sum_{i=0}^{n-1} Q^i(u) + \frac{1}{n}(I - Q(u))\right)^{-1}$$

and

$$K_1 \stackrel{\text{def}}{=} \sup_{n\ge n_0}\sup_{u\in U}\|\Phi_n(u)\|_\nu < \infty. \tag{3.13}$$

Define

$$Z_n(u) \stackrel{\text{def}}{=} I + \frac{1}{n}\sum_{i=1}^{n-1}\sum_{j=1}^{i-1}(Q^j(u) - \Pi(u)).$$

We have

$$K_2 \stackrel{\text{def}}{=} \sup_{n\ge n_0}\sup_{u\in U}\|Z_n(u)\|_\nu < \infty. \tag{3.14}$$

A direct calculation yields

$$(I - Q(u) + \Pi(u))Z_n(u)\Phi_n(u) = I. \tag{3.15}$$

Clearly, (3.15) implies that

$$\Pi_\delta(u)(I - Q(u) + \Pi(u))Z_n(u)\Phi_n(u) = \Pi_\delta(u),$$

so that

$$\Pi_\delta(u)(I - Q(u))Z_n(u)\Phi_n(u) + \Pi(u)Z_n(u)\Phi_n(u) = \Pi_\delta(u). \tag{3.16}$$

From (3.15), we infer that

$$\Pi(u)(I - Q(u) + \Pi(u))Z_n(u)\Phi_n(u) = \Pi(u).$$

Therefore

$$\Pi(u)Z_n(u)\Phi_n(u) = \Pi(u).$$

Substituting into (3.16), we obtain

$$\Pi_\delta(u)(I - Q(u))Z_n(u)\Phi_n(u) = \Pi_\delta(u) - \Pi(u),$$

and consequently

$$\|\Pi_\delta(u) - \Pi(u)\|_\nu = \|\Pi_\delta(u)(Q_\delta(u) - Q(u))Z_n(u)\Phi_n(u)\|_\nu .$$

Combining this with (3.6) and (3.12)–(3.14), we obtain

$$\|\Pi_\delta(u) - \Pi(u)\|_\nu \le \|Q_\delta(u) - Q(u)\|_\nu K_0K_1K_2 < \delta K_0K_1K_2.$$

The proof of statement (i) is finished.

(ii) Using L defined in (2.1) and (3.7), we obtain

$$\begin{aligned}
|J^i(u) - J^i_\delta(u)| &= |\Pi(u)c^i(\cdot,u) - \Pi_\delta(u)c^i_\delta(\cdot,u)| \\
&\le |\Pi(u)c^i(\cdot,u) - \Pi_\delta(u)c^i(\cdot,u)| + |\Pi_\delta(u)(c^i(\cdot,u) - c^i_\delta(\cdot,u))| \\
&\le L\nu(s_0) \sup_{|w|\le\nu} \frac{|\Pi(u)w - \Pi_\delta(u)w|}{\nu(s_0)} + \delta \\
&\le L\nu(s_0) \sup_{s\in S} \sup_{|w|\le\nu} \frac{|\Pi(u)w - \Pi_\delta(u)w|}{\nu(s)} + \delta \\
&= L\nu(s_0)\, \|\Pi(u) - \Pi_\delta(u)\|_\nu + \delta,
\end{aligned}$$

where s_0 is an arbitrary state. Now (ii) follows from (i). □

A version of Lemma 3.4 corresponding with a *bounded* function ν was established by Stettner [42]. When ν is bounded, an elementary proof of Lemma 3.4 (stated as an extension of Ueno's lemma [46]) is possible [32].

Proof of Theorem 3.1. Choose some $\epsilon > 0$. According to Lemma 3.4 there exists some δ such that for all $u \in U$,

$$\left|J^i(u) - J^i_\delta(u)\right| \le \epsilon. \tag{3.17}$$

Let $u^* \in U$ be a Nash equilibrium for the game G_δ in the class U of multipolicies (its existence follows from Lemma 3.3). It then follows from (3.17) that u^* is an ϵ-equilibrium (in the class U) for the original game. The fact that u^* is an ϵ-equilibrium in the class Γ of all multipolicies follows from Theorem 3 and Remark 1 in [34] (or [18, 14] in the Borel state space framework). □

4. The discounted stochastic game. In this section, we drop conditions **C3** and **C4**. However, in the unbounded cost case, we make the following assumption.

C7: There exists $\alpha \in [\beta, 1)$ such that

$$\beta Q(\nu|s,x) \le \alpha\nu(s)$$

for each $s \in S$ and $x \in X$.

Using **C7**, we can easily prove that, for any $s \in S$, any multipolicy $\gamma \in \Gamma$, and any number of stages k, we have

$$|\beta^{k-1}E^\gamma_s(c^i(S_k,X_k))| \le \beta^{k-1}E^\gamma_s(|c^i(S_k,X_k)|) \le L\beta^{k-1}E^\gamma_s(\nu(S_k)) \le L\alpha^{k-1}\nu(s),$$

where L is the constant defined in **C1**. This gives us the following lemma.

LEMMA 4.1. *Assume* **C1** *and* **C7**. *Then for every player i the expected discounted cost $D^i(s,\gamma)$ is well defined (absolutely convergent) for each $s \in S$ and $\gamma \in \Gamma$.*

We are ready to formulate our main result in this section.

THEOREM 4.1. *Any discounted stochastic game satisfying conditions* **C1**, **C2**, *and* **C7** *has a stationary ε-equilibrium for any $\varepsilon > 0$.*

Before we give the proof of this theorem, we state some auxiliary results. Let Δ_1 be the set of *all density functions* in L^1_ν. Clearly, Lemma 3.2 holds true if Δ is replaced by Δ_1. Applying the approximation scheme from section 3 to the present situation, we construct a game G_δ for any $\delta > 0$ such that (3.7) holds and, moreover, we have

$$\sup_{|f|\le\nu} |Q(f|s,u) - Q_\delta(f|s,u)| \le \delta \tag{4.1}$$

for each $s \in S$ and any stationary multipolicy $u \in U$.

Fix player i and set

$$K^n(s,u) = E^u_s(c^i(S_n, X_n)) \quad \text{and} \quad K^n_\delta(s,u) = E^u_s(c^i_\delta(S_n, X_n)),$$

where $s \in S$ and $u \in U$. Clearly, $K^n(s,u)$ is the nth stage cost for player i under stationary multipolicy u when the game starts at an initial state $s \in S$.

LEMMA 4.2. *Assume* **C1** *and* **C7**. *Then, for each* $s \in S$ *and* $u \in U$, *we have*

$$|K^n(s,u) - K^n_\delta(s,u)| \le \delta(1 + (n-1)L)\left(\frac{\alpha}{\beta}\right)^{n-1}.$$

Proof. The proof proceeds by induction. For $n = 1$ the inequality follows immediately from (3.7). We now give the induction step. Note that

$$\begin{aligned} |K^{n+1}(s,u) - K^{n+1}_\delta(s,u)| &= |Q(K^n(\cdot,u)|s,u) - Q_\delta(K^n_\delta(\cdot,u)|s,u)| \\ &\le |Q(K^n(\cdot,u)|s,u) - Q_\delta(K^n(\cdot,u)|s,u)| \\ &\quad + |Q_\delta(K^n(\cdot,u)|s,u) - Q_\delta(K^n_\delta(\cdot,u)|s,u)|. \end{aligned}$$

Using (4.1), our induction hypothesis, and the obvious inequality

$$K^n(s,u) \le L\left(\frac{\alpha}{\beta}\right)^{n-1}\nu(s),$$

which holds for every $s \in S$ and $u \in U$, we obtain

$$\begin{aligned} |K^{n+1}(s,u) - K^{n+1}_\delta(s,u)| &\le \delta L\left(\frac{\alpha}{\beta}\right)^{n-1} + \delta(1 + (n-1)L)\left(\frac{\alpha}{\beta}\right)^{n-1} \\ &= \delta(1+nL)\left(\frac{\alpha}{\beta}\right)^{n-1} \le \delta(1+nL)\left(\frac{\alpha}{\beta}\right)^{n}, \end{aligned}$$

which ends the proof. □

From Lemmas 4.1 and 4.2, we infer the following result.

LEMMA 4.3. *Assume* **C1** *and* **C7**. *If* $D^i_\delta(s,u)$ *is the expected* β*-discounted cost for player* i *in the game* G_δ, *then*

$$|D^i(s,u) - D^i_\delta(s,u)| \le \delta(1 + \alpha(L-1))(1-\alpha)^{-2}$$

for each $s \in S$ *and* $u \in U$.

The game G_δ is characterized by the cost functions c^i_δ, transition probability Q_δ, and the function ν_δ. Note that if δ is sufficiently small, then the game G_δ satisfies condition **C1** with L replaced by $2L$. From our approximation scheme (the new

definition of the space V) and Remark 2.1, it follows also that **C2** is satisfied in our game G_δ. Since $|\nu(s) - \nu_\delta(s)| < \delta$ for all $s \in S$, we have (by (4.1))

$$\beta \int_S \nu_\delta(t) Q_\delta(dt|s,x) \le \alpha \nu_\delta(s) + \alpha\delta + 2\beta\delta < \alpha_0 \nu_\delta(s),$$

where $\alpha_0 = \alpha + \alpha\delta + 2\delta$ and $s \in S$, $x \in X$. Note that $\beta < \alpha_0$, and if δ is sufficiently small, then $\alpha_0 < 1$, and thus G_δ satisfies condition **C7** with α replaced by α_0. Let δ_0 be a positive number such that for every $\delta < \delta_0$ the game G_δ satisfies conditions of type **C1**, **C2**, and **C7**. In particular, we have $\beta < \alpha_0 < 1$.

LEMMA 4.4. *If $\delta < \delta_0$, then the game G_δ has a Nash equilibrium in the class U of all stationary multipolicies.*

Proof. We use a transformation to bounded cost games similar to that of [43, p. 101]. One may define the new discount factor $\tilde\beta \stackrel{\text{def}}{=} \alpha_0$ and the functions

$$\tilde c^i(s,x) = \frac{c^i_n(x)}{\nu_\delta(s)}, \qquad \tilde z(s,t,x) = \frac{\beta z_n(t,x)\nu_\delta(t)}{\alpha_0 \nu_\delta(s)},$$

where $s \in Y_n$, $t \in S$, and $x \in X$. This transformation ensures that the new costs $\tilde c^i$ are bounded and that

$$q(\cdot|s,x) \stackrel{\text{def}}{=} \int_S \tilde z(s,t,x)\varphi(dt)$$

is a transition subprobability such that $q(Y_n|s,x)$ is continuous in x for each n and $s \in S$. Moreover, it implies that

$$\tilde D^i(s,u) = \frac{D^i_\delta(s,u)}{\nu_\delta(s)}, \tag{4.2}$$

where $\tilde D^i(s,u)$ is the expected discounted cost for player i under any $u \in U$ in the transformed (bounded) game. Similarly, as in section 3 we can recognize the game G_δ as a game with countably many states. By [11], such a game has a stationary Nash equilibrium. In other words, our bounded game has an equilibrium $u^\star$ in the class U_0 of all piecewise constant multipolicies. It now follows from Lemma 5.2 in the appendix that $u^\star$ is an equilibrium for the bounded game in the class U. By (4.2), we infer that $u^\star$ is also an equilibrium (in the class U of all stationary multipolicies) for the game G_δ. □

Proof of Theorem 4.1. Fix $\varepsilon > 0$. By Lemma 4.3, there exists $\delta < \delta_0$ such that

$$|D^i(s,u) - D^i_\delta(s,u)| \le \varepsilon/2$$

for each $s \in S$ and $u \in U$. It follows from Lemma 4.4 that the game G_δ has an equilibrium $u^\star$ in the class U. Clearly, $u^\star$ is an ε-equilibrium in the class U for the original game. The fact that $u^\star$ is also an ε-equilibrium in Γ follows from Theorem 2 and Remark 1 in [34] (or [14] in the case of Borel state space games). □

5. Appendix. In this section we restrict our attention to the approximating games and state some auxiliary results on sufficiency of piecewise constant policies in the sense that they can be used to dominate any other policy. Related statements are proven in [1] for countable state space models. Their extension to the present situation would require new notation and some additional measure theoretic work.

Therefore, in this section we restrict ourselves to stationary policies, and in such a case we can use different methods which are based on some standard arguments from the dynamic programming literature [14].

Let G_δ be an approximating game corresponding to a partition of the state space. Fix player i and a stationary *piecewise constant* multipolicy u^{-i} for the other players. For any $s \in S$ and $f \in U^i$ set

$$c(s,f) = c^i_\delta(s,(u^{-i},f)) \quad \text{and} \quad q(\cdot|s,f) = Q_\delta(\cdot|s,(u^{-i},f)).$$

Recall that U^i_0 denotes the set of all piecewise constant stationary policies for player i.

Consider the Markov decision process (MDP) with the state space S, the action space X^i, the cost function c, and the transition probability q.

The average cost case. We assume that δ is sufficiently small so that the MDP satisfies conditions **C1**–**C4** (restricted to the one-player case). Let $J_n(s,f)$ ($J(f)$) denote the expected n-stage (expected average) cost (in the MDP) under stationary policy f.

LEMMA 5.1. *Assume* **C1**–**C4**, *and consider the average cost MDP described above. Then for each $f \in U^i$ there exists some $f_0 \in U^i_0$ such that*

$$J(f_0) \le J(f).$$

Proof. Let $f \in U^i$ and $g = J(f)$. By Lemma 3.1, our MDP satisfies condition **C6** with ν replaced by ν_δ and possibly different constants. It is well known that in such a case the function

$$h(s) \stackrel{\text{def}}{=} E^f_s \left[\sum_{n=1}^{\infty} (c(S_n, X^i_n) - g) \right]$$

is well defined and $h \in L^\infty_{\nu_\delta}$. Moreover, we have

$$g + h(s) = c(s,f) + q(h|s,f) \quad \text{for each} \quad s \in S.$$

For the details, see [14] and [25]. Our approximating game (and thus the MDP) satisfies continuity conditions **C1**–**C2**. Because the cost function c and the transition probability correspond to a partition of the state space (and, in addition, the other players use stationary piecewise constant multipolicy u^{-i}), this implies that one can find some $f_0 \in U^i_0$ such that

$$c(s,f_0) + q(h|s,f_0) \le c(s,f) + q(h|s,f) = g + h(s)$$

for all $s \in S$. Iterating this inequality, we obtain

$$J_n(s,f_0) + q^n(h|s,f_0) \le ng + h(s)$$

for all $s \in S$. Hence

$$\frac{J_n(s,f_0)}{n} + \frac{q^n(h|s,f_0)}{n} \le g + \frac{h(s)}{n}$$

for each n, and consequently

$$J(f_0) \le g = J(f).$$

For a detailed discussion of the fact that **C3** implies that $q^n(h|s, f_0)/n \to 0$ as $n \to \infty$, consult [14] or [34]. □

The discounted cost case. We now assume that the stochastic game satisfies **C1**, **C2**, and **C7**. If δ is sufficiently small, then both G_δ and the aforementioned MDP satisfy **C1**, **C2**, and **C7**, but with different constants (see section 4). Let $f \in U^i$. By $D_n(s, f)$ ($D(s, f)$) we denote the expected n-stage discounted (total discounted) cost in the MDP under policy f.

LEMMA 5.2. *Assume* **C1**, **C2**, *and* **C7**, *and consider the discounted MDP described above. Then for each $f \in U^i$ there exists some $f_0 \in U_0^i$ such that*

$$D(s, f_0) \le D(s, f) \quad \textit{for every} \quad s \in S.$$

Proof. Set $d(s) = D(s, f)$, $s \in S$. Under our assumptions, we have

$$d(s) = c(s, f) + \beta q(d|s, f)$$

for all $s \in S$ (see Lemma 4.1). From our compactness and continuity conditions, **C7**, and the construction of the approximating game, it follows that there exists some $f_0 \in U_0^i$ such that

$$c(s, f_0) + \beta q(d|s, f_0) \le c(s, f) + \beta q(d|s, f) = d(s)$$

for each $s \in S$. Hence

$$D_n(s, f_0) + \beta^n q^n(d|s, f_0) \le d(s)$$

for each n and $s \in S$, and consequently

$$D(s, f_0) \le D(s, f) \quad \text{for each} \quad s \in S.$$

The fact that $\beta^n q^n(d|s, f_0) \to 0$ as $n \to \infty$ follows easily from **C7**. □

REFERENCES

[1] E. ALTMAN, *Constrained Markov Decision Processes*, Chapman and Hall, London, 1998.

[2] E. ALTMAN, *Non zero-sum stochastic games in admission, service and routing control in queueing systems*, QUESTA, 23 (1996), pp. 259–279.

[3] E. ALTMAN AND A. HORDIJK, *Zero-sum Markov games and worst-case optimal control of queueing systems*, QUESTA, 21, (1995), pp. 415–447.

[4] E. ALTMAN, A. HORDIJK, AND F. M. SPIEKSMA, *Contraction conditions for average and α-discount optimality in countable state Markov games with unbounded rewards*, Math. Oper. Res., 22 (1997), pp. 588–618.

[5] R. AMIR, *Continuous stochastic games of capital accumulation with convex transitions*, Games Econom. Behav., 15 (1996), pp. 111–131.

[6] V. S. BORKAR AND M. K. GHOSH, *Denumerable state stochastic games with limiting average payoff*, J. Optim. Theory Appl., 76 (1993), pp. 539–560.

[7] L. O. CURTAT, *Markov equilibria of stochastic games with complementarities*, Games Econom. Behav., 17 (1996), pp. 177–199.

[8] R. DEKKER AND A. HORDIJK, *Recurrence conditions for average and Blackwell optimality in denumerable state Markov decision chains*, Math. Oper. Res., 17 (1992), pp. 271–290.

[9] R. DEKKER, A. HORDIJK, AND F. M. SPIEKSMA, *On the relation between recurrence and ergodicity properties in denumerable Markov decision chains*, Math. Oper. Res., 19 (1994), pp. 539–559.

[10] D. DUFFIE, J. GEANAKOPLOS, A. MAS-COLELL, AND A. MCLENNAN, *Stationary Markov equilibria*, Econometrica, 62 (1994), pp. 745–782.

[11] A. FEDERGRUEN, *On N-person stochastic games with denumerable state space*, Adv. in Appl. Probab., 10 (1978), pp. 452–471.

[12] J. A. Filar and K. Vrieze, *Competitive Markov Decision Processes*, Springer-Verlag, New York, 1997.
[13] C. Harris, P. Reny, and A. Robson, *The existence of subgame-perfect equilibrium in continuous games with almost perfect information: A case for public randomization*, Econometrica, 63 (1995), pp. 507–544.
[14] O. Hernández-Lerma and J. B. Lasserre, *Further Topics in Discrete-Time Markov Control Processes*, Springer-Verlag, New York, 1999.
[15] O. Hernández-Lerma and J. B. Lasserre, *Zero-sum stochastic games in Borel spaces: Average payoff criteria*, SIAM J. Control Optim., 39 (2001), pp. 1520–1539.
[16] A. Jaśkiewicz, *On strong* 1*-optimal policies in Markov control processes with Borel state spaces*, Bull. Polish Acad. Sci., 48 (2000), pp. 439–450.
[17] A. Jaśkiewicz, *Zero-sum semi-Markov games*, SIAM J. Control Optim., to appear.
[18] A. Jaśkiewicz and A. S. Nowak, *On the optimality equation for zero-sum ergodic stochastic games*, Math. Methods Oper. Res., 54 (2001), pp. 291–301.
[19] H.-U. Küenle, *Equilibrium strategies in stochastic games with additive cost and transition structure*, Int. Game Theory Rev., 1 (1999), pp. 131–147.
[20] H.-U. Küenle, *Stochastic games with complete information and average cost criterion*, Ann. Internat. Soc. Dynam. Games, 5 (2000) pp. 325–338.
[21] A. Maitra and W. D. Sudderth, *Borel stochastic games with limsup payoff*, Ann. Probab., 21 (1993), pp. 861–885.
[22] A. Maitra and W. D. Sudderth, *Discrete Gambling and Stochastic Games*, Springer-Verlag, New York, 1996.
[23] J.-F. Mertens and A. Neyman, *Stochastic games*, Internat. J. Game Theory, 10 (1981), pp. 53–66.
[24] J.-F. Mertens and T. Parthasarathy, *Equilibria for discounted stochastic games*, Research paper 8750, CORE, Université Catholique de Louvain, Louvain-la-Neuver, Belgium, 1987.
[25] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, New York, 1993.
[26] S. P. Meyn and R. L. Tweedie, *Computable bounds for geometric convergence rates of Markov chains*, Ann. Appl. Probab., 4 (1994), pp. 981–1011.
[27] S. P. Meyn and R. L. Tweedie, *State-dependent criteria for convergence of Markov chains*, Ann. Appl. Probab., 4 (1994), pp. 149–168.
[28] J. Neveu, *Mathematical Foundations of the Calculus of Probability*, Holden–Day, San Francisco, 1965.
[29] A. S. Nowak, *Existence of equilibrium stationary strategies in discounted noncooperative stochastic games with uncountable state space*, J. Optim. Theory Appl., 45 (1985), pp. 591–602.
[30] A. S. Nowak, *Stationary equilibria for nonzero-sum average payoff ergodic stochastic games with general state space*, Ann. Internat. Soc. Dynam. Games, 1 (1994), pp. 231–246.
[31] A. S. Nowak, *Zero-sum average payoff stochastic games with general state space*, Games Econom. Behav., 7 (1994), pp. 221–232.
[32] A. S. Nowak, *A generalization of Ueno's inequality for n-step transition probabilities*, Applicationes Mathematicae, 25 (1998), pp. 295–299.
[33] A. S. Nowak, *Sensitive equilibria for ergodic stochastic games with countable state spaces*, Math. Methods Oper. Res., 50 (1999), pp. 65–76.
[34] A. S. Nowak, *Optimal strategies in a class of zero-sum ergodic stochastic games*, Math. Methods Oper. Res., 50 (1999), pp. 399–420.
[35] A. S. Nowak and A. Jaśkiewicz, *Remarks on Sensitive Equilibria in Stochastic Games with Additive Reward and Transition Structure*, Technical report, Institute of Mathematics, University of Zielona Góra, Zielona Góra, Poland, 2002.
[36] A. S. Nowak and T. E. S. Raghavan, *Existence of stationary correlated equilibria with symmetric information for discounted stochastic games*, Math. Oper. Res., 17 (1992), pp. 519–526.
[37] T. Parthasarathy and S. Sinha, *Existence of stationary equilibrium strategies in non-zero-sum discounted stochastic games with uncountable state space and state independent transitions*, Internat. J. Game Theory, 18 (1989), pp. 189–194.
[38] O. Passchier, *The Theory of Markov Games and Queueing Control*, Ph.D. thesis, Department of Mathematics and Computer Science, Leiden University, Leiden, The Netherlands, 1998.
[39] U. Rieder, *Equilibrium plans for nonzero-sum Markov games*, in Game Theory and Related Topics, O. Moeschlin and D. Pallaschke, eds., North–Holland, Amsterdam, 1979, pp. 91–102.
[40] L. I. Sennott, *Nonzero-sum stochastic games with unbounded costs: discounted and average*

cost cases, Z. Oper. Res., 40 (1994), pp. 145–162.

[41] F. M. Spieksma, *Geometrically Ergodic Markov Chains and the Optimal Control of Queues*, Ph.D. thesis, Department of Mathematics and Computer Science, Leiden University, Leiden, The Netherlands, 1990.

[42] L. Stettner, *On nearly self-optimizing strategies for a discrete-time uniformly ergodic adaptive model*, Appl. Math. Optim., 27 (1993), pp. 161–177.

[43] J. Van der Wal, *Stochastic Dynamic Programming*, Mathematical Center Tracts 139, Mathematisch Centrum, Amsterdam, 1981.

[44] N. Vieille, *Equilibrium in* 2*-player stochastic games* 1: *A reduction*, Israel J. Math., 119 (2000), pp. 55–91.

[45] N. Vieille, *Equilibrium in* 2*-player stochastic games* 2: *The case of recursive games*, Israel J. Math., 119 (2000), pp. 93–126.

[46] T. Ueno, *Some limit theorems for temporally discrete Markov processes*, J. Fac. Sci. Univ. Tokyo, 7 (1957), pp. 449–462.

[47] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1977.

[48] W. Whitt, *Representation and approximation of noncooperative sequential games*, SIAM J. Control Optim., 18 (1980), pp. 33–48.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 6, pp. 1840–1857

EXISTENCE, MULTIPLICITY, AND REGULARITY FOR SUB-RIEMANNIAN GEODESICS BY VARIATIONAL METHODS*

ROBERTO GIAMBÒ†, FABIO GIANNONI†, AND PAOLO PICCIONE‡

Abstract. We develop a variational theory for geodesics joining a point and a one dimensional submanifold of a sub-Riemannian manifold. Given a Riemannian manifold (M, g), a smooth distribution $\Delta \subset TM$ of codimension one in M, a point $p \in M$, and a smooth immersion $\gamma : \mathbb{R} \to M$ with closed image in M and which is everywhere transversal to Δ, we look for curves in M that are stationary with respect to the Riemannian energy functional among all of the absolutely continuous curves *horizontal* with respect to Δ and that join p and γ. If (M, g) is complete, such extremizers exist, and they are curves of class C^2 characterized as the solutions of an integro-differential equation or by a system of ordinary differential equations. We present some results concerning a sort of *exponential map* relative to the integro-differential equation and some applications. In particular, we obtain that if p and γ are sufficiently close in M, then there exists a unique length minimizer. We obtain existence and multiplicity results by means of the Ljusternik–Schnirelman theory.

Key words. sub-Riemannian geodesics, Ljusternik–Schnirelman theory

AMS subject classifications. 53C17, 53C22, 58E05, 58E10, 58E25

PII. S0363012900367242

1. Introduction.

The goal of this paper is to develop a variational theory for sub-Riemannian geodesics in a general context and to obtain existence and multiplicity results, in analogy with the corresponding theory for Riemannian geodesics.

The interest in the existence of (local) length minimizers in a sub-Riemannian manifold comes essentially from control theory, where such minimizers represent optimal solutions of a system with linear constraints on the first derivatives of the admissible paths.

A sub-Riemannian manifold consists of a triple (M, Δ, g), where M is a smooth manifold, $\Delta \subset TM$ is a smooth distribution in M, and g is a positive definite metric tensor on Δ. The kind of sub-Riemannian geodesics that we are interested in are the so-called *normal* geodesics, which are curves obtained as solutions of the sub-Riemannian Hamiltonian $H = \frac{1}{2}g\big(p|_\Delta, p|_\Delta\big)$ on TM^*. It is well known that normal sub-Riemannian geodesics are horizontal with respect to Δ, i.e., $\dot{x} \in \Delta$, and that they locally minimize their length (see [7]); such curves will be studied in this paper as solutions of a Lagrangian variational problem. There is a class of geodesic curves, called *abnormal*, which satisfies Hamiltonian equations with abnormal Hamiltonians. These equations are determined by the distribution Δ and not by the metric g. The first example of an abnormal sub-Riemannian length minimizer has been constructed by Montgomery in [9]. Wide classes of abnormal minimizers for 2-distributions have been described in [7]. Good references for the basics of sub-Riemannian geodesics are [8, 10]. A wide study of abnormal geodesics can be found in [1].

*Received by the editors February 2, 2000; accepted for publication (in revised form) September 25, 2001; published electronically March 5, 2002.
http://www.siam.org/journals/sicon/40-6/36724.html

†Dipartimento di Matematica e Informatica, Università di Camerino, Italy (roberto.giambo@unicam.it, fabio.giannoni@unicam.it).

‡Departamento de Matemática, Universidade de São Paulo, Brazil (piccione@ime.usp.br). This author is partially sponsored by CNPq (Processo n. 301410/95) and by FAPESP (Processo n. 00/09277-5).

An approach to multiplicity results in optimal control theory by means of global analysis techniques (Morse theory and Ljusternik–Schnirelman theory) has been developed in a different context by the Russian school (see Vakhrameev [13] and its extensive bibliography). The main assumption of [13] is that the control system is of *constant rank*, i.e., that the *endpoint mapping* defined on the space of horizontal paths is of constant rank. This assumption is satisfied, for instance, when Δ is strongly bracket-generating (or when Δ is integrable in the sense of Frobenius).

The main obstruction for developing a variational theory for sub-Riemannian geodesics between two points is that, in general, the set of horizontal curves joining two fixed points does not have a differentiable structure unless one poses strong nonintegrability conditions on the distribution Δ, or, for instance, when the system is of constant rank in the sense of Vakhrameev [13]. In this paper, we do not make such assumptions on Δ, but rather we allow that the final endpoint of a trial path for our variational problem is free to move on a submanifold of M which is *transversal* to Δ. With such a choice we overcome the *rigidity* problem of the fixed endpoint case, and we obtain a smooth manifold structure for the set of horizontal curves satisfying suitable regularity conditions (see Proposition 2.1 and Remark 2.2).

As a first approach to this technique, we will initially consider the case of a distribution Δ of codimension one in TM, and we will assume that Δ is *transversally oriented* in TM, i.e., that the quotient bundle TM/Δ is orientable. One can extend the sub-Riemannian metric defined on Δ to a Riemannian metric in M; such an extension is of course noncanonical. By the transversal orientation, it is not restrictive to assume that Δ is the orthogonal distribution to a unit vector field on M. Moreover, if the original sub-Riemannian structure is complete, one can assume that the Riemannian extension is also complete. We will therefore consider the following geometric setup.

Let (M, g) be a complete Riemannian manifold, let Y be a never vanishing smooth vector field on M, and let $\Delta = Y^{\perp}$ denote the orthogonal distribution to Y. For all $q \in M$, we set $\Delta_q = \Delta \cap T_q M$; moreover, we will denote by $\langle \cdot, \cdot \rangle$ the positive definite inner product on each tangent space $T_q M$ given by $g(q)$ and by $|\cdot|$ the corresponding length. We will assume without loss of generality that Y is normalized on M:

$$\langle Y, Y \rangle = 1. \tag{1}$$

Let $\gamma : \mathbb{R} \to M$ be a smooth curve in M which is a closed immersion of $\mathbb{R}$ in M (i.e., $\gamma'(t) \neq 0$ for all t and γ has closed image $\mathrm{Im}(\gamma)$ in M) and which is everywhere *transversal* to Δ, i.e., $\dot\gamma(t) \notin \Delta_{\gamma(t)}$ for all t.

Let ∇ denote the covariant derivative relative to the Levi–Civita connection of g; given a smooth function α on M, we denote by $\nabla\alpha$ the gradient of α with respect to the metric g.

Let p be a fixed point in M, set $\mathrm{Im}(\gamma) = \gamma(\mathbb{R})$, and let $\mathcal{C}^1_{p,\gamma}$ denote the set of all curves of class C^1 in M parameterized on $[0, 1]$ joining p and γ:

$$\mathcal{C}^1_{p,\gamma} = \{z \in C^1([0,1], M) : z(0) = p,\ z(1) \in \mathrm{Im}(\gamma)\};$$

by $\mathcal{C}^1_{p,\gamma}(\Delta)$ we will denote the subset of $\mathcal{C}^1_{p,\gamma}$ consisting of *horizontal* curves:

$$\mathcal{C}^1_{p,\gamma}(\Delta) = \{z \in \mathcal{C}^1_{p,\gamma} : \dot z(t) \in \Delta_{z(t)} \text{ for all } t\}. \tag{2}$$

The set $\mathcal{C}^1_{p,\gamma}$ has a natural structure of an infinite dimensional Banach differentiable manifold; the differentiable structure of (a suitable completion of) $\mathcal{C}^1_{p,\gamma}(\Delta)$ will be discussed in section 2. We denote by L and E, respectively, the Riemannian *length* and *energy* functionals on $\mathcal{C}^1_{p,\gamma}$ defined by

$$L(z) = \int_0^1 \sqrt{\langle \dot z, \dot z\rangle}\, \mathrm{d}t, \quad E(z) = \frac{1}{2}\int_0^1 \langle \dot z, \dot z\rangle\, \mathrm{d}t. \tag{3}$$

In this paper, we will be interested in studying the curves in $\mathcal{C}^1_{p,\gamma}(\Delta)$ that are local length minimizers, i.e., in those horizontal curves z between p and γ such that, for $a, b \in [0,1]$ sufficiently close, the restriction $z|_{[a,b]}$ is a horizontal curve of minimal length between $z(a)$ and $z(b)$. Such curves will be called local *sub-Riemannian length minimizers* between p and γ. More in particular, we will consider the stationary points of the functional E in $\mathcal{C}^1_{p,\gamma}(\Delta)$: they are geodesics in the sub-Riemannian manifold $(\mathcal{M}, \Delta, g)$. Namely, the critical points of E in $\mathcal{C}^1_{p,\gamma}(\Delta)$ are normal sub-Riemannian geodesics, and therefore they locally minimize their length (see, for instance, [7, Appendix C]). Moreover, the minima in $\mathcal{C}^1_{p,\gamma}(\Delta)$ of the functionals E and of L coincide up to parameterization.

PROPOSITION 1.1. *A curve $x \in \mathcal{C}^1_{p,\gamma}(\Delta)$ is a minimal point for E on $\mathcal{C}^1_{p,\gamma}(\Delta)$ if and only if it is a sub-Riemannian length minimizer between p and γ satisfying $|\dot x(t)| = \text{const.}$ on $[0,1]$.*

We have a first existence result concerning the minima of the length functional.

THEOREM 1.2. *Let (M, g) be a complete Riemannian manifold, let Y be a never vanishing smooth vector field on M, let $\Delta = Y^\perp$ be its orthogonal distribution, and let $\gamma : \mathbb{R} \to M$ be a closed immersion which is transversal to Δ. Then there exists at least one minimizer x for L in $\mathcal{C}^1_{p,\gamma}(\Delta)$, with $|\dot x(t)|$ constant on $[0,1]$.*

Some results concerning the characterization of the normal geodesics in a sub-Riemannian manifold, connecting submanifolds $\mathcal{P}$ and $\mathcal{Q}$ of any codimension, as critical points of the action functional can be found, for instance, in [12]. Existence and multiplicity results can be obtained if $\mathcal{P}$ and $\mathcal{Q}$ are closed submanifolds of $\mathcal{M}$ and one of them is compact. The proof is essentially the same as that of Theorem 1.8 (thanks to the results proved in [12]).

As will be observed in section 2, for the proof of Theorem 1.2 it is not restrictive to assume that γ is an integral line of the vector field Y. For the other results of the paper, we will explicitly make such assumption.

Given a smooth vector field W on M, we denote by $(\nabla W)^*$ the transpose of the covariant derivative of W, which is the $(1,1)$ tensor field on M whose value at a point $q \in M$ is the linear map on T_qM defined by

$$\big\langle (\nabla W)^*[v_1], v_2\big\rangle = \big\langle \nabla_{v_2} W, v_1\big\rangle \quad \text{for all } v_1, v_2 \in T_qM. \tag{4}$$

For all x in $\mathcal{C}^1_{p,\gamma}(\Delta)$, let $\lambda_x : [0,1] \to \mathbb{R}$ be the map of class C^2 given by

$$\lambda_x(t) = e^{\int_0^t \langle \dot x, \nabla_Y Y\rangle\, \mathrm{d}s} \cdot \left[\int_t^1 \big\langle \dot x, \nabla_{\dot x} Y\big\rangle e^{-\int_0^s \langle \dot x, \nabla_Y Y\rangle\, \mathrm{d}r}\, \mathrm{d}s\right]. \tag{5}$$

THEOREM 1.3. *Suppose that γ is an integral line of Y. If x is a critical point of E in $\mathcal{C}^1_{p,\gamma}(\Delta)$, then x is a curve of class C^2, and it satisfies the equation*

$$\nabla_{\dot x}\dot x - \nabla_{\dot x}\left(\lambda_x \cdot Y\right) + \lambda_x \cdot (\nabla Y)^*[\dot x] = 0. \tag{6}$$

Remark 1.4. Observe that the integro-differential equation (5)–(6) is not *local*, in the following sense. Given any subinterval $[a,b] \subseteq [0,1]$, one can consider an alternative integro-differential problem given by (6) and λ_x given by

$$\lambda_x = e^{\int_a^t \langle \dot{x}, \nabla_Y Y\rangle \, \mathrm{d}s} \cdot \left[\int_t^b \left\langle \dot{x}, \nabla_{\dot{x}} Y \right\rangle e^{-\int_0^s \langle \dot{x}, \nabla_Y Y\rangle \, \mathrm{d}r} \, \mathrm{d}s \right]. \tag{7}$$

Given a solution x of (5)–(6), the restriction of x to the interval $[a,b]$ is not, in general, a solution of (6)–(7). The interpretation of this fact is that, even though the critical points of E in $\mathcal{C}^1_{p,\gamma}(\Delta)$ locally minimize their length, in general they do *not* minimize locally the distance between a point and an integral line of Y. The fact that sufficiently small portions of solutions of (6)–(7) minimize their length can be deduced from the fact that (6)–(7) are equivalent to the Hamilton equations satisfied by the normal sub-Riemannian geodesics (see Appendix B); for normal sub-Riemannian geodesics the local minimality is proven, for instance, in [7, Appendix C]. However, a direct proof of the local minimality for solutions of (6)–(7) can also be given without the use of Hamiltonian formalism.

Note that if the pair (x, λ_x) satisfies (6)–(7), clearly it satisfies the system of ordinary differential equations

$$\begin{cases} \nabla_{\dot{x}} \dot{x} - \nabla_{\dot{x}} (\lambda Y) + \lambda \, (\nabla Y)^* [\dot{x}] = 0, \\ \lambda' - \lambda \left\langle \nabla_Y Y, \dot{x} \right\rangle + \left\langle \dot{x}, \nabla_{\dot{x}} Y \right\rangle = 0. \end{cases} \tag{8}$$

At the end of section 2, we point out that, if $(x(t), \lambda_x(t))$ is a solution of the above system, then $\langle \dot{x}(t), \dot{x}(t) \rangle$ and $\langle \dot{x}(t), Y(x(t)) \rangle$ are constant. From this point of view (in analogy with Riemannian geodesics), we could say that the sub-Riemannian geodesics are the solution of (8) with the initial condition $\langle \dot{x}(0), Y(x(0)) \rangle = 0$. Observe also that the pair (x, λ) plays the role of the *Hamiltonian lift* of the sub-Riemannian geodesic x.

Remark 1.5. Suppose that Y is a conformal Killing vector field. Then, for all $v \in Y^\perp$, it is $\langle \nabla_v Y, v \rangle = 0$. Hence, for all $x \in \mathcal{C}^1_{p,\gamma}(\Delta)$, it is $\langle \nabla_{\dot{x}} Y, \dot{x} \rangle \equiv 0$, and so $\lambda_x \equiv 0$. From (6), this implies that if x is a critical point of E in $\mathcal{C}^1_{p,\gamma}(\Delta)$, then it is a Riemannian geodesic.

Remark 1.6. Changing the point of view, the stationary paths x for the functional E in $\mathcal{C}^1_{p,\gamma}(\Delta)$ can be thought of as *constrained* critical points of E in $\mathcal{C}^1_{p,\gamma}$ subject to the linear constraint on the first derivative $\dot{x} \in \Delta$. From this viewpoint, given such a constrained critical point x, the map λ_x of formula (5) can be interpreted as the corresponding *Lagrange multiplier* (see Appendix A).

Remark 1.7. It is well known (see for instance, [7]) that, as a consequence of the Pontryagin maximum principle, the sub-Riemannian extremals are either abnormal or they satisfy the Hamilton equations corresponding to the Hamiltonian function $H : TM^* \to \mathbb{R}$ given by

$$H(q,p) = \frac{1}{2} g^{-1}\big(p|_\Delta, p|_\Delta\big),$$

where $g^{-1} : \Delta^* \times \Delta^* \to \mathbb{R}$ is the inverse of the metric $g|_\Delta$. We will prove in Appendix B that (5) and (6) (or, equivalently, system (8)) are equivalent to the Hamilton equations of H.

Let us recall the following definition: let $\mathcal{X}$ be a topological space, and let $\mathcal{Y} \subset \mathcal{X}$ be a subspace. The *Ljusternik–Schnirelman category* $\mathrm{cat}_{\mathcal{X}}(\mathcal{Y})$ of $\mathcal{Y}$ in $\mathcal{X}$ is the possibly

infinite minimal number of closed contractible subsets of $\mathcal{X}$ that form a covering of $\mathcal{Y}$. We set $\operatorname{cat}(\mathcal{X}) = \operatorname{cat}_{\mathcal{X}}(\mathcal{X})$.

We have a multiplicity result for sub-Riemannian geodesics between p and γ given in terms of the Ljusternik–Schnirelman category of the Hilbert manifold $\Omega_{p,\gamma}(\Delta)$ defined in (10) as follows.

THEOREM 1.8. *Under the assumptions of Theorem* 1.2, *if* γ *is an integral line of* Y, *there are at least* $\operatorname{cat}(\Omega_{p,\gamma}(\Delta))$ *normal geodesics between* p *and* γ. *Moreover, if* $\operatorname{cat}(\Omega_{p,\gamma}(\Delta))$ *is infinite, then there exists a sequence* $\{x_n\}_{n\in\mathbb{N}}$ *of normal geodesics between* p *and* γ *such that*

$$\lim_{n\to\infty} E(x_n) = +\infty.$$

Under suitable assumptions on the flow of Y, one proves that the inclusion map $\Omega_{p,\gamma}(\Delta)$ into $\Omega_{p,\gamma}$ (defined in (9)) is a homotopy equivalence; it follows, in particular, that $\operatorname{cat}(\Omega_{p,\gamma}(\Delta))$ is equal to $\operatorname{cat}(\Omega_{p,\gamma})$ (see Proposition 3.3). If M is not contractible, by a well known result of Fadell and Husseini (see [3]), it is $\operatorname{cat}(\Omega_{p,\gamma}) = +\infty$, and we have a class of examples where Theorem 1.8 gives the existence of infinite normal geodesics between p and γ.

Let us now look abstractly at the integro-differential equation (6). Observe that it makes perfect sense to consider solutions of (6) that are not horizontal curves. We will prove in section 2 that a solution x of (6) is a horizontal curve if and only if $\dot{x}(0) \in \Delta$ (Theorem 2.3). Moreover, we will see that a solution of (6) satisfying $x(0) = v \in T_pM$ exists and is unique, provided that v is small enough (Proposition 4.1). This fact allows us to introduce a *sub-Riemannian* point-to-line exponential map $\mathfrak{exp}_p$, defined in a neighborhood of $0 \in T_pM$ by $\mathfrak{exp}_p(v) = x_v(1)$, where x_v is the unique solution of (6) satisfying the initial condition $\dot{x}_v(0) = v$.

In perfect analogy with the well-known properties of the Riemannian exponential map, the map $\mathfrak{exp}_p$ is a diffeomorphism between a neighborhood of $0 \in T_pM$ and a neighborhood of $p \in M$. As a first important (and trivial) consequence of this fact, we obtain the following local uniqueness result for sub-Riemannian length minimizers between a point and an integral line of Y.

THEOREM 1.9. *Under the hypotheses of Theorem* 1.2, *if* γ *is sufficiently close to* p, *then there is a unique minimizer* x *for* L *in* $\mathcal{C}^1_{p,\gamma}(\Delta)$ *with* $|\dot{x}(t)|$ *constant on* $[0,1]$.

We conclude with a remark that sub-Riemannian geodesics of the kind considered in this paper have appeared recently in a general relativistic context (see [4, 5]). Given a *stationary Lorentzian manifold*, which represents the model for a relativistic spacetime with gravitational field stationary with respect to a distinguished observer field Y, one has a natural sub-Riemannian metric defined on the orthogonal distribution to Y by taking the restriction of the spacetime metric tensor. In this situation, the sub-Riemannian geodesics (in a suitable conformal perturbation of the metric) joining an event p of M and an integral line of Y represent the *travel time* brachistochrones between a source and an observer in the spacetime.

We briefly outline the structure of the paper. In order to be able to apply techniques from critical point theory and global analysis on manifolds, our variational problem has to be cast in a Hilbert manifold setting. In section 2, we define the variational framework by proving the existence of a Hilbert manifold structure on the set $\Omega_{p,\gamma}(\Delta)$ of horizontal curves from p to γ of Sobolev class H^1; then we study the first variation of the sub-Riemannian energy functional E defined on $\Omega_{p,\gamma}(\Delta)$.

In section 3 we prove that $E : \Omega_{p,\gamma}(\Delta) \to \mathbb{R}$ satisfies a certain compactness property, called the Palais–Smale condition, which is the key property for all of the

results of existence and multiplicity of critical points for functions on noncompact manifolds. By standard arguments of critical point theory we will then obtain the proof of Proposition 1.1, Theorem 1.2, Theorem 1.3, and Theorem 1.8.

In section 4 we look at the flow on M defined by the integro-differential equation (6), and we study the sub-Riemannian exponential map.

Finally, the paper has two short appendices. In Appendix A, we show that the map λ_x of formula (5) and the integro-differential equation (6) appear naturally also when one uses the method of Lagrange multipliers. In Appendix B we prove that the integro-differential equation (5)–(6) is equivalent to the normal extremal equations coming from the maximum principle of Pontryagin.

2. The variational framework. First variation and the critical points of E. We assume hereafter that (M, g) is a complete Riemannian manifold and that Y is a normalized smooth vector field on M; we set $\Delta = Y^{\perp}$. For each $q \in M$, we set $\Delta_q = \Delta \cap T_q M$.

Let $I \subset \mathbb{R}$ be any interval, and suppose that $\gamma : I \to M$ is a smooth curve having values in a compact subset of M. Suppose that γ is everywhere transversal to Δ. By the transversality of γ and a partition of unity argument, it is easy to prove that we can find a complete Riemannian metric $\tilde{g}$ and a smooth vector field $\tilde{Y}$ on M such that

- the orthogonal distribution $\tilde{Y}^{\perp}$ with respect to $\tilde{g}$ coincides with Δ;
- g and $\tilde{g}$ coincide on Δ;
- $\dot{\gamma}(t) = \tilde{Y}(\gamma(t))$ for all $t \in I$.

Now, since (M, g) is complete, all curves starting at the fixed point p whose length is bounded above by a given constant remain inside a compact subset of M. Then, to prove the results announced in the introduction, it suffices to treat the case that γ is a maximal integral line of the vector field Y.

We will assume hereafter that $\gamma : \mathbb{R} \to M$ is an integral curve of Y.

As is customary, if $I \subseteq \mathbb{R}$ is any interval, we will denote by $H^1(I, \mathbb{R}^n)$ the Sobolev space of absolutely continuous curves $z : I \longmapsto \mathbb{R}^n$ such that the integral $\int_I |\dot{z}|^2 \, \mathrm{d}t$ is finite, where $|\cdot|$ denotes the Euclidean norm in $\mathbb{R}^n$.

Given *any* differentiable manifold N, the set $H^1([0,1], N)$ is defined as the set of all absolutely continuous curves $z : [0,1] \longmapsto N$ such that, for every local chart (V, φ) on N, with $\varphi : U \longmapsto \mathbb{R}^n$ a diffeomorphism, and for every closed subinterval $I \subseteq [0,1]$ such that $z(I) \subset V$, it is $\varphi \circ z \in H^1(I, \mathbb{R}^n)$. For all differentiable manifold N, with $\dim(N) = n$, the set $H^1([0,1], N)$ has the structure of an infinite dimensional manifold, modeled on the Hilbert space $H^1([0,1], \mathbb{R}^n)$. We will denote by TN the tangent bundle of N and by $\pi : TN \to N$ the canonical projection; for $p \in N$, $T_pN = \pi^{-1}(p)$ denotes the tangent space of N at p. A vector field along a curve $z : [0,1] \to N$ is a map $\zeta : [0,1] \to TN$ with $\pi(\zeta(t)) = z(t)$ for all t. Given any $z \in H^1([0,1], N)$, the tangent space $T_z H^1([0,1], N)$ is identified with the set

$$T_z H^1([0,1], N) = \{\zeta \in H^1([0,1], TN) : \zeta \text{ vector field along } z\},$$

which is an infinite dimensional vector space with a topology that makes it into a *Hilbertable* space.

We introduce the sets

$$\Omega_{p,\gamma} = \{z \in H^1([0,1], M) : z(0) = p, \ z(1) \in \mathrm{Im}(\gamma)\}, \tag{9}$$

and

$$\Omega_{p,\gamma}(\Delta) = \{z \in \Omega_{p,\gamma} : \langle \dot{z}, Y \rangle = 0 \text{ a.e. in } [0,1]\}. \tag{10}$$

It is well known that $\Omega_{p,\gamma}$ is a smooth Hilbert submanifold of $H^1([0,1],M)$; for all $z \in \Omega_{p,\gamma}$ the tangent space $T_z\Omega_{p,\gamma}$ is given by

$$T_z\Omega_{p,\gamma} = \{V \in T_zH^1([0,1],M) : V(0) = 0,\ V(1) \parallel Y(z(1))\}.$$

We endow $T_z\Omega_{p,\gamma}$ with the Hilbert space structure induced by the inner product:

$$\langle\!\langle V, V \rangle\!\rangle = \int_0^1 \langle \nabla_{\dot z} V, \nabla_{\dot z} V \rangle \, \mathrm{d}t; \tag{11}$$

then $\Omega_{p,\gamma}$ becomes an infinite dimensional Riemannian manifold with the metric defined by (11).

The functionals E and L defined in formula (3) have a continuous extension to the space $H^1([0,1],M)$. The energy functional E is smooth, and so is its restriction to $\Omega_{p,\gamma}$; the length functional L is only Lipschitz continuous. For $z \in H^1([0,1],M)$, the Gateaux derivative $\mathrm{d}E(z)$ is given by the bounded linear map on $T_zH^1([0,1],M)$:

$$\mathrm{d}E(z)[V] = \int_0^1 \langle \nabla_{\dot z} V, \dot z \rangle \, \mathrm{d}t. \tag{12}$$

PROPOSITION 2.1. *$\Omega_{p,\gamma}(\Delta)$ is a smooth Hilbert submanifold of $\Omega_{p,\gamma}$. For all $z \in \Omega_{p,\gamma}(\Delta)$, the tangent space $T_z\Omega_{p,\gamma}(\Delta)$ is given by the Hilbert subspace of $T_z\Omega_{p,\gamma}$:*

$$T_z\Omega_{p,\gamma}(\Delta) = \{V \in T_z\Omega_{p,\gamma} : \langle \nabla_{\dot z} V, Y \rangle + \langle \dot z, \nabla_V Y \rangle = 0\}. \tag{13}$$

The restriction of E to $\Omega_{p,\gamma}(\Delta)$ is smooth.

Proof. We consider the map $F : \Omega_{p,\gamma} \to L^2([0,1],\mathbb{R})$ defined by

$$F(z) = \langle \dot z, Y \rangle. \tag{14}$$

It is easy to see that F is smooth, that $\Omega_{p,\gamma}(\Delta) = F^{-1}(0)$, and that the Gateaux derivative $\mathrm{d}F(z)$ of F at z is given by

$$\mathrm{d}F(z)[V] = \langle \nabla_{\dot z} V, Y \rangle + \langle \dot z, \nabla_V Y \rangle. \tag{15}$$

By the implicit function theorem (see [6]), to prove the proposition we need to show that, for all $z \in \Omega_{p,\gamma}(\Delta)$, the differential $\mathrm{d}F(z) : T_z\Omega_{p,\gamma} \to L^2([0,1],\mathbb{R})$ is surjective. To this aim, let $h \in L^2([0,1],M)$ be fixed; consider the vector field $V_h = \phi_h \cdot Y$ along z, where ϕ_h is the function

$$\phi_h(t) = e^{-\int_0^t \langle \nabla_Y Y, \dot z \rangle \, \mathrm{d}s} \cdot \left[\int_0^t h(s) \cdot e^{\int_0^s \langle \nabla_Y Y, \dot z \rangle \, \mathrm{d}r} \, \mathrm{d}s \right]. \tag{16}$$

It is easily checked that $\phi_h \in H^1([0,1],\mathbb{R})$ and $\phi_h(0) = 0$, which implies that $V_h \in T_z\Omega_{p,\gamma}$. Moreover, $\mathrm{d}F(z)[V_h] = h$, which proves that $\mathrm{d}F(z)$ is surjective. Finally, for $z \in \Omega_{p,\gamma}(\Delta)$, the tangent space $T_z\Omega_{p,\gamma}(\Delta)$ is given by the kernel of $\mathrm{d}F(z)$, and (13) is proven. □

Remark 2.2. The regularity of the set of horizontal curves joining a fixed point p and a curve γ can be studied alternatively considering the *endpoint mapping* $\Omega_p(\Delta) \ni z \mapsto z(1) \in M$ defined on the set of horizontal curves starting at p. It is well known that $\Omega_p(\Delta)$ has the structure of a smooth Hilbert manifold and that the image of the differential of the endpoint map at the point z contains the distribution $\Delta_{z(1)}$.

By the inverse mapping theorem $\Omega_{p,\gamma}(\Delta)$ is a smooth submanifold of $\Omega_p(\Delta)$ provided that γ is transversal to Δ. Using the same argument, one proves (see [12]) that for distributions Δ of arbitrary rank, the set $\Omega_{P_1,P_2}(\Delta)$ of horizontal paths joining two given submanifolds $P_1, P_2 \subset M$ has the structure of an infinite dimensional Hilbert manifold provided that either P_1 or P_2 is everywhere transversal to Δ.

THEOREM 2.3. *The critical points of E in $\Omega_{p,\gamma}(\Delta)$ are curves of class C^2. They are characterized as the solutions on the interval $[0,1]$ of the following integro-differential equation:*

$$\nabla_{\dot{x}}\dot{x} - \nabla_{\dot{x}}\left(\lambda_x \cdot Y\right) + \lambda_x \cdot (\nabla Y)^*[\dot{x}] = 0, \tag{17}$$

where

$$\lambda_x(t) = e^{\int_0^t \langle \nabla_Y Y, \dot{x}\rangle \,\mathrm{d}s} \cdot \left[\int_t^1 \left\langle \dot{x}, \nabla_{\dot{x}} Y\right\rangle e^{-\int_0^s \langle \nabla_Y Y, \dot{x}\rangle \,\mathrm{d}r}\,\mathrm{d}s\right], \tag{18}$$

and $\dot{x}(0) \in \Delta_p$.

Proof. To determine the integro-differential equation (17)–(18), we argue as follows. Let x be any point in $\Omega_{p,\gamma}(\Delta)$; for all $W \in T_x\Omega_{p,\gamma}$, we define a *projection* V_W of W onto $T_x\Omega_{p,\gamma}(\Delta)$ by setting

$$V_W = W - \psi_W \cdot Y, \tag{19}$$

where

$$\psi_W(t) = e^{-\int_0^t \langle \nabla_Y Y, \dot{x}\rangle \,\mathrm{d}s} \cdot \left[\int_0^t C_W(s) \cdot e^{\int_0^s \langle \nabla_Y Y, \dot{x}\rangle \,\mathrm{d}r}\,\mathrm{d}s\right], \tag{20}$$

and

$$C_W = \left\langle \dot{x}, \nabla_W Y\right\rangle + \left\langle \nabla_{\dot{x}} W, Y\right\rangle = \left\langle W, (\nabla Y)^*[\dot{x}]\right\rangle + \left\langle \nabla_{\dot{x}} W, Y\right\rangle. \tag{21}$$

Observe that, since $\left\langle Y, Y\right\rangle = 1$, then $\left\langle \nabla_{\dot{x}} Y, Y\right\rangle = 0$, and

$$C_Y = \left\langle \dot{x}, \nabla_Y Y\right\rangle.$$

Checking, with the above definitions, that V_W is in $T_x\Omega_{p,\gamma}(\Delta)$ is straightforward, and the details are omitted.

Now, if x is a critical point of E in $\Omega_{p,\gamma}(\Delta)$, it is $\mathrm{d}E(x)[V_W] = 0$ for all $W \in T_x\Omega_{p,\gamma}$, and, since $\left\langle \dot{x}, Y\right\rangle = 0$, (12) gives

$$\begin{aligned} 0 = \mathrm{d}E(x)[W - \psi_W \cdot Y] &= \int_0^1 \left[\left\langle \nabla_{\dot{x}}\left(W - \psi_W \cdot Y\right), \dot{x}\right\rangle\right]\,\mathrm{d}t \\ &= \int_0^1 \left[\langle \nabla_{\dot{x}} W, \dot{x}\rangle - \dot{\psi}_W \cdot \langle Y, \dot{x}\rangle - \psi_W \cdot \langle \nabla_{\dot{x}} Y, \dot{x}\rangle\right]\mathrm{d}t \\ &= \int_0^1 \left[\left\langle \nabla_{\dot{x}} W, \dot{x}\right\rangle - \psi_W \cdot \left\langle \nabla_{\dot{x}} Y, \dot{x}\right\rangle\right]\,\mathrm{d}t. \end{aligned} \tag{22}$$

Now let λ_x be given by (18). Reversing the order of integration with Fubini's theorem and recalling (20) and (21), we get

$$\begin{aligned}
&\int_0^1 \psi_W \cdot \langle \nabla_{\dot x} Y, \dot x\rangle \,\mathrm{d}t \\
&\quad= \int_0^1 \langle \nabla_{\dot x} Y, \dot x\rangle e^{-\int_0^t \langle \nabla_Y Y, \dot x\rangle \,\mathrm{d}r} \cdot \left[\int_0^t C_W(s) e^{\int_0^s \langle \nabla_Y Y, \dot x\rangle \,\mathrm{d}r}\,\mathrm{d}s\right] \mathrm{d}t \\
&\quad= \int_0^1 C_W(s) e^{\int_0^s \langle \nabla_Y Y, \dot x\rangle \,\mathrm{d}r} \cdot \left[\int_s^1 \langle \nabla_{\dot x} Y, \dot x\rangle e^{-\int_0^t \langle \nabla_Y Y, \dot x\rangle \,\mathrm{d}r}\,\mathrm{d}t\right] \mathrm{d}s \\
&\quad= \int_0^1 C_W(s)\cdot \lambda_x(s)\,\mathrm{d}s = \int_0^1 \left[\langle W, (\nabla Y)^*[\dot x]\rangle + \langle \nabla_{\dot x} W, Y\rangle\right]\cdot \lambda_x(s)\,\mathrm{d}s.
\end{aligned} \tag{23}$$

Then (22) becomes

$$\int_0^1 \left[\langle \nabla_{\dot x} W, \dot x\rangle - \lambda_x(s)\cdot \langle \nabla_{\dot x} W, Y\rangle - \lambda_x(s)\cdot \langle W, (\nabla Y)^*[\dot x]\rangle\right]\,\mathrm{d}s = 0. \tag{24}$$

Suppose that x is of class C^2; integrating by parts the terms in (24) containing the covariant derivative $\nabla_{\dot x} W$, we obtain

$$\int_0^1 \langle W, \nabla_{\dot x}(\dot x - \lambda_x \cdot Y) + \lambda_x \cdot (\nabla Y)^*[\dot x]\rangle \,\mathrm{d}t = 0 \tag{25}$$

for all $W \in T_x\Omega_{p,\gamma}$. Observe that there is no boundary term arising from the integration by parts because $W(0) = 0$, $\lambda_x(1) = 0$, and $\langle W(1), \dot x(1)\rangle = 0$. This last equality follows from the fact that $W(1)$ is parallel to $\dot\gamma$, hence to Y, and $\langle \dot x, Y\rangle = 0$.

Since W is arbitrary in (25), the fundamental lemma of calculus of variations tells us that x satisfies (17), and we have proven that the critical points of E in $\Omega_{p,\gamma}(\Delta)$ satisfy the integro-differential problem (17)–(18).

The C^2-regularity of the critical points of E is obtained by a bootstrap argument. If $x \in \Omega_{p,\gamma}(\Delta)$ is a critical point of E, then (24) holds for any C^∞ vector field W along x such that $W(0) = W(1) = 0$. If Z is a vector field of class L^2, let us denote by $\int_0^s Z\,\mathrm{d}s$ the vector field U solving

$$\begin{cases} \nabla_{\dot x} U = Z, \\ U(0) = 0. \end{cases} \tag{26}$$

From now on, all the integrals of vector fields along curves will be understood in this sense. Using a local coordinate system and Gronwall's lemma, it is easily seen that $U \in H^1$ and, therefore, $U \in C^0$.

Note that $\dot x \in L^2$ and $\lambda_x \in C^0$. Integrating by parts the last term inside the integral in (24) (and recalling that W vanishes at both endpoints of x), we get

$$\begin{aligned}
0 &= \int_0^1 \langle \nabla_{\dot x} W, \dot x - \lambda_x(s) Y\rangle - \langle W, \lambda_x(s)(\nabla Y)^*[\dot x]\rangle \mathrm{d}s \\
&= \int_0^1 \left\langle \nabla_{\dot x} W, \dot x - \lambda_x(s) Y + \left[\int_0^s \lambda_x(t)(\nabla Y)^*[\dot x]\,\mathrm{d}t\right]\right\rangle \,\mathrm{d}s.
\end{aligned} \tag{27}$$

Set

$$\chi(s) = \dot x(s) - \lambda_x(s) Y(x(s)) + \int_0^s \lambda_x(t)(\nabla Y)^*[\dot x(t)]\,\mathrm{d}t.$$

Using local coordinates and Christoffel symbols, we obtain by (27) $\chi \in H^1$ and, in particular, $\chi \in C^0$. Since λ_x and Y are C^0, we deduce $\dot{x} \in C^0$. Then λ_x, Y, $\int_0^s \lambda_x(t)(\nabla Y)^*[\dot{x}]\ \mathrm{d}t$ are C^1, while, by (27), $\chi \in C^1$. Hence $\dot{x}$ is C^1, obtaining the desired regularity.

Conversely, if x is a solution of (17)–(18) such that $\dot{x}(0) \in \Delta$, then x is a horizontal curve. Namely, using (17), we compute

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\langle \dot{x}, Y\rangle &= \langle \nabla_{\dot{x}}\dot{x}, Y\rangle + \langle \dot{x}, \nabla_{\dot{x}} Y\rangle \\ &= \lambda_x' - \lambda_x\langle \nabla_Y Y, \dot{x}\rangle + \langle \dot{x}, \nabla_{\dot{x}} Y\rangle = 0, \end{aligned} \tag{28}$$

where the last equality is due to (18). Hence, if $\langle \dot{x}(0), Y\rangle = 0$, then $\langle \dot{x}, Y\rangle \equiv 0$. Moreover, it is easy to see that all elements $V \in T_x\Omega_{p,\gamma}(\Delta)$ are of the form V_W given in formula (19). Indeed, since $C_V = 0$, we can choose $W = V$ and $\psi_W = 0$. So every solution x of (17)–(18) such that $x(0) = p$ and $\dot{x}(0) \in \Delta_p$ is a critical point of E in $\Omega_{p,\gamma}(\Delta)$. This concludes the proof. ∎

Remark 2.4. Observe that, by (28), a solution $(x(t), \lambda_x(t))$ of the system (8) satisfies the conservation law $\langle \dot{x}(t), Y(x(t))\rangle = 0$.

We remark that the map λ_x in Theorem 2.3 can be interpreted as the *Lagrangian multiplier* of the constrained critical point x in $\Omega_{p,\gamma}(\Delta)$ (see Lemma A.1).

COROLLARY 2.5. *If x is a critical point of E in $\Omega_{p,\gamma}(\Delta)$, then it satisfies the conservation law:*

$$|\dot{x}| \equiv \text{const}. \tag{29}$$

Proof. Taking the product of (17) by $\dot{x}$, we obtain

$$\begin{aligned} 0 &= \langle \nabla_{\dot{x}}\dot{x}, \dot{x}\rangle - \lambda_x' \cdot \langle Y, \dot{x}\rangle - \lambda_x \cdot \langle \nabla_{\dot{x}} Y, \dot{x}\rangle + \lambda_x \cdot \langle \nabla_{\dot{x}} Y, \dot{x}\rangle \\ &= \langle \nabla_{\dot{x}}\dot{x}, \dot{x}\rangle = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\langle \dot{x}, \dot{x}\rangle, \end{aligned} \tag{30}$$

which proves the claim. ∎

3. Minimal curves for L. In this section, we show the existence of a minimum for the energy functional E in $\Omega_{p,\gamma}(\Delta)$, whose regularity has already been proven in Theorem 2.3. To this aim, we show that E satisfies a good enough compactness property, namely the Palais–Smale condition. Proposition 1.1, which establishes the relation between minimal points for E and sub-Riemannian length minimizers can be obtained easily by standard arguments.

We recall, given a C^1 functional $\mathfrak{f} : \mathfrak{X} \to \mathbb{R}$ on a Hilbert manifold $\mathfrak{X}$, that $\mathfrak{f}$ is said to satisfy the *Palais–Smale condition* at level $c \in \mathbb{R}$ if every sequence $\{x_n\}_{n\in\mathbb{N}} \subset \mathfrak{X}$ such that

$$\begin{aligned} &\lim_{n\to\infty} \mathfrak{f}(x_n) = c, \\ &\lim_{n\to\infty} \|\mathrm{d}\mathfrak{f}(x_n)\| = 0 \end{aligned} \tag{31}$$

(where $\|\cdot\|$ denotes the norm of bounded linear functionals on the Hilbert space $T_{x_n}\mathfrak{X}$) has a subsequence converging in $\mathfrak{X}$. We will use throughout the paper some well-known results concerning functionals satisfying the Palais–Smale condition (see, for instance, the proofs of Corollary 3.2 and Theorem 1.8); for the reader's convenience, we will briefly give a formal statement of these properties when they are used.

PROPOSITION 3.1. *The functional E satisfies the Palais–Smale condition at every level $c \in \mathbb{R}$.*

Proof. Let $\{x_n\}_{n\in\mathbb{N}}$ be a sequence in $\Omega_{p,\gamma}(\Delta)$ satisfying (31). Since M is complete and $\int_0^1 \langle \dot{x}_n, \dot{x}_n\rangle \mathrm{d}t \leq \text{const.}$, up to subsequences we can assume that x_n is convergent to some curve x uniformly and $\dot{x}_n$ is weakly convergent to $\dot{x}$ in L^2. Observe that $x \in \Omega_{p,\gamma}(\Delta)$; the fact that $x(1) \in \mathrm{I}m(\gamma)$ follows from our assumption that $\mathrm{I}m(\gamma)$ is closed in M.

Then (31) yields

$$\int_0^1 \left\langle \dot{x}_n, \nabla_{\dot{x}_n}\zeta \right\rangle \mathrm{d}t = \int_0^1 \left\langle a_n, \nabla_{\dot{x}_n}\zeta \right\rangle \mathrm{d}t \tag{32}$$

for every admissible variation ζ and for some sequence a_n converging to 0 in L^2. Then, with a similar argument as in Theorem 2.3, (32) gives the existence of a sequence b_n converging to 0 in L^2 such that

$$\dot{x}_n - b_n - \varphi_{x_n} Y(x_n) + \int_0^s \varphi_{x_n} \cdot (\nabla Y)^*[\dot{x}_n]\,\mathrm{d}t = z_n, \tag{33}$$

where

$$\varphi_{x_n}(\tau) = \int_\tau^1 \left\langle \dot{x}_n, \nabla_{\dot{x}_n} Y(x_n)\right\rangle e^{\int_\sigma^\tau \langle \dot{x}_n, \nabla_{Y(x_n)}Y(x_n)\rangle\,\mathrm{d}\rho}\,\mathrm{d}\sigma, \tag{34}$$

and z_n is a sequence in L^2 such that $\nabla_{\dot{x}_n} z_n = 0$. From (34), φ_{x_n} is uniformly bounded in L^∞, and it has uniformly bounded derivative in L^1. Moreover, the covariant integral V_n,

$$V_n = \int_0^s A_n\,\mathrm{d}t, \quad A_n = \varphi_{x_n} \cdot (\nabla Y)^*[\dot{x}_n], \tag{35}$$

solves the equation

$$\begin{cases} \nabla_{\dot{x}_n} V_n = A_n, \\ V_n(0) = 0; \end{cases} \tag{36}$$

since $\dot{x}_n$ is uniformly bounded in L^2, using the coordinate expression of (36), we have that V_n is uniformly bounded in L^∞, and $\dot{V}_n$ is uniformly bounded in L^2; then V_n is uniformly bounded in H^1.

The uniform boundedness of z_n in L^2 implies the existence of a sequence $s_n \in [0,1]$ such that $z_n(s_n)$ is bounded. Using again the coordinate expression for the equation

$$\begin{cases} \nabla_{\dot{x}_n} z_n = 0, \\ z_n(s_n) = B_n, \end{cases} \tag{37}$$

with B_n bounded, we obtain that z_n is uniformly bounded in H^1.

In conclusion, (33) yields the existence of a sequence c_n which is uniformly bounded in L^∞ and has uniformly bounded derivative in L^1 such that

$$\dot{x}_n - b_n = c_n. \tag{38}$$

But c_n has a converging subsequence in L^2 (see [2]), and b_n converges to 0 in L^2; these two facts imply that $\dot{x}_n$ has a converging subsequence in L^2, and then E satisfies the Palais–Smale condition at every level $c \in \mathbb{R}$. ∎

We are now ready to prove the existence of a minimizer for E:

COROLLARY 3.2. *The functional E attains its minimum in $\Omega_{p,\gamma}(\Delta)$.*

Proof. This is a classical argument of calculus of variations, repeated here for the reader's convenience. Suppose that we are given a (possibly infinite dimensional) complete Hilbert manifold $\mathfrak{X}$ and a smooth functional $\mathfrak{f} : \mathfrak{X} \to \mathbb{R}$ satisfying the Palais–Smale condition at every level $c \in \mathbb{R}$. For $c \in \mathbb{R}$, let $\mathfrak{f}^c = \{x \in \mathfrak{X} : \mathfrak{f}(x) \leq c\}$ denote the closed c-sublevel of $\mathfrak{f}$. If $c_0 \in \mathbb{R}$ is *not* a critical value of $\mathfrak{f}$, then there exists $\eta > 0$ such that $\mathfrak{f}^{c_0-\eta}$ is homeomorphic to $\mathfrak{f}^{c_0+\eta}$ (see, e.g., [11]); such a homeomorphism can be given explicitly using the flow of the gradient of $\mathfrak{f}$. It follows that $\mathfrak{f}$ attains its minimum if $\mathfrak{f}$ is bounded from below; namely, if $c_0 = \inf \mathfrak{f}$ were not a critical value, then $\mathfrak{f}^{c_0-\eta} = \emptyset$ would be homeomorphic to $\mathfrak{f}^{c_0+\eta} \neq \emptyset$.

In our situation, we have that the Hilbert manifold $\mathfrak{X} = \Omega_{p,\gamma}(\Delta)$ is complete because (M, g) is complete and because $\mathrm{Im}(\gamma)$ is closed in M; the functional $\mathfrak{f} = E$ satisfies the Palais–Smale condition by Proposition 3.1, and it is clearly bounded from below. The conclusion follows. □

Proof of Proposition 1.1. It follows easily from the existence of a minimizer of E and a reparameterization argument. □

Also, the proof of Theorems 1.2 and 1.3 is a straightforward consequence of Proposition 1.1, Theorem 2.3 and Corollary 2.5.

Proof of Theorem 1.8. It is a classical argument of the theory of Ljusternik and Schnirelman (see [11] for details). We recall briefly the main ideas of the theory; assume that $\mathfrak{X}$ is a complete Hilbert manifold and that $\mathfrak{f} : \mathfrak{X} \to \mathbb{R}$ is a smooth functional satisfying the Palais–Smale condition at every level $c \in \mathbb{R}$ which is bounded from below. Then one proves the following facts:

1. for every $c \in \mathbb{R}$, the Ljusternik–Schnirelman category $\mathrm{cat}_{\mathfrak{X}}(\mathfrak{f}^c)$ of the sublevel $\mathfrak{f}^c$ is finite;
2. if there exists $c \in \mathbb{R}$ such that $\sup\{\mathfrak{f}(z) : \mathfrak{f}'(z) = 0\} < c$, then $\mathrm{cat}_{\mathfrak{X}}(\mathfrak{f}^c) = \mathrm{cat}(\mathfrak{X})$;
3. if $c \in \mathbb{R}$ as in (2) does not exist, then there exists a sequence $\{x_n\}$ of critical points such that $\mathfrak{f}(x_n) \to +\infty = \sup \mathfrak{f}$;
4. if, on the contrary, such a c exists, let Γ_k denote the collection of closed subsets A of $\mathfrak{X}$ whose Ljusternik–Schnirelman category $\mathrm{cat}_{\mathfrak{X}}(A)$ is greater than or equal to k. Then there exists $c \in \mathbb{R}$ such that $\mathfrak{f}^c \in \Gamma_k$ for all $k = 1, \ldots, \mathrm{cat}(\mathfrak{X})$. Since $\sup \mathfrak{f} > -\infty$, one can define the following (possibly finite) sequence of real numbers:
$$c_k = \inf_{A \in \Gamma_k} \left[\sup_{x \in A} \mathfrak{f}(x) \right], \qquad k = 1, 2, \ldots, \mathrm{cat}(\mathfrak{X}); \tag{39}$$
5. each c_k is a critical value of $\mathfrak{f}$; moreover, if for some k and $r > 0$ one has $c_k = c_{k+1} = \cdots = c_r$, then there are at least $r + 1$ critical points in $\mathfrak{f}^{-1}(c_k)$.

By the facts above, one concludes easily that $\mathfrak{f}$ has at least $\mathrm{cat}(\mathfrak{X})$ critical points.

Now, the proof of Theorem 1.8 is an application of the above theory in the case that $\mathfrak{X} = \Omega_{p,\gamma}(\Delta)$ and $\mathfrak{f} = E$, keeping in mind the results of Theorem 2.3 and of Proposition 3.1. □

Finally, we conclude the section with a result that relates the topology of the spaces $\Omega_{p,\gamma}(\Delta)$ and $\Omega_{p,\gamma}$; the latter space is in general an easier object to deal with.

PROPOSITION 3.3. *Assume that the vector field Y is complete on M, and denote by $\psi : M \times \mathbb{R} \to \mathcal{M}$ its flow, i.e., for all $q \in M$, $t \mapsto \psi(q, t)$ is the maximal integral line of Y passing through q at the instant $t = 0$; let $\mathrm{d}_x\psi(q, t)$ denote the differential of the map $m \mapsto \psi(m, t)$ at the point q.*

Assume that there exist continuous maps $A, B : M \to \mathbb{R}^+$ such that the following estimate holds for the norm of $\mathrm{d}_x\psi$:

$$\|\mathrm{d}_x\psi(q,r)\| \le A(q)\cdot|r| + B(q) \quad \textit{for all } (q,r)\in M\times\mathbb{R}. \tag{40}$$

Then there exists a strong deformation retract[1] *from $\Omega_{p,\gamma}$ to $\Omega_{p,\gamma}(\Delta)$; in particular, $\Omega_{p,\gamma}$ and $\Omega_{p,\gamma}(\Delta)$ have the same homotopy type and the same Ljusternik–Schnirelman category.*

Proof. Given an H^1 map $\mu : [0,1] \to \mathbb{R}$ with $\mu(0) = 0$ and a curve $z \in \Omega_{p,\gamma}$, denote by $\mathfrak{c}(z,\mu) \in \Omega_{p,\gamma}$ the curve $t \mapsto \psi\big(z(t),\mu(t)\big)$; clearly, if $\mu \equiv 0$ on $[0,1]$, then $\mathfrak{c}(z,\mu) = z$.

For $z \in \Omega_{p,\gamma}$, denote by $\mu_z : [0,1] \to \mathbb{R}$ the solution of the initial value problem:

$$\mu_z'(t) = -\langle \mathrm{d}_x\psi\big(z(t),\mu_z(t)\big)[\dot z(t)], Y\big(\psi\big(z(t),\mu_z(t)\big)\big)\rangle, \quad \mu_z(0) = 0. \tag{41}$$

The existence of a global solution on the interval $[0,1]$ of (41) follows easily using the estimate (40), observing that z and $\langle Y, Y\rangle$ are bounded and $\dot z \in L^2$.

By standard continuous dependence arguments for ordinary differential equations, the map $\Omega_{p,\gamma} \ni z \mapsto \mu_z \in \mathcal{H}^1([0,1],\mathbb{R})$ is continuous; therefore, we get a continuous map $\Omega_{p,\gamma} \ni z \mapsto \mathfrak{c}(z,\mu_z) \in \Omega_{p,\gamma}$.

Now, using (41), it is easily seen that $\mathfrak{c}(z,\mu_z)$ is almost everywhere orthogonal to Y, and hence $\mathfrak{c}(z,\mu_z) \in \Omega_{p,\gamma}(\Delta)$ for all $z \in \Omega_{p,\gamma}$. A strong deformation retract $h : \Omega_{p,\gamma}\times[0,1] \to \Omega_{p,\gamma}$ from $\Omega_{p,\gamma}$ to $\Omega_{p,\gamma}(\Delta)$ is then obtained by setting

$$h(z,s) = \mathfrak{c}(z, s\cdot\mu_z). \qquad \square$$

For instance, the assumptions of Proposition 3.3 hold if Y is a complete Killing vector field in (M,g), in which case $\mathrm{d}_x\psi$ is an isometry. More generally, if Y is of the form $Y = \langle W,W\rangle^{-\frac{1}{2}}W$ for some *conformal* vector field, then the inequality (40) is satisfied when $\langle W,W\rangle$ is bounded on each integral line of W. Recall that a smooth vector field W on M is conformal if its flow consists of conformal (i.e., angle preserving) maps; equivalently, W is conformal if the Lie derivative $\mathcal{L}_W(g)$ of the metric g is conformally equivalent to g, i.e., of the form $\phi\cdot g$ for some smooth map $\phi : M \to \mathbb{R}^+$.

4. Local theory: The exponential map. In this section, we study the flow on M defined by the integro-differential equation (6), with the aim of proving the local uniqueness of sub-Riemannian length minimizers between a point and an integral line of y.

We start by proving an existence and uniqueness result for local solutions of (6).

PROPOSITION 4.1. *Let $p \in M$ and $v_0 \in T_pM$; suppose that $|v_0|$ is sufficiently small. Then there exists a unique solution of the integro-differential equation* (17)–(18) *satisfying the initial conditions $x(0) = p$ and $\dot x(0) = v_0$.*

Proof. As we have observed in Corollary 2.5, the solutions of the integro-differential problem satisfy $|\dot x| = \text{const.}$, and hence all the solutions of our initial value problem remain inside a ball of radius b around the point p. Using local coordinates, we may therefore assume that $M = \mathbb{R}^n$, $p = 0$.

[1]Recall that, given a topological space $\mathcal{X}$ and a subspace $\mathcal{Y} \subset \mathcal{X}$, a strong deformation retract from $\mathcal{X}$ to $\mathcal{Y}$ is a continuous map $h : \mathcal{X}\times[0,1] \to \mathcal{X}$ such that $h(x,0) = x$ for all $x \in \mathcal{X}$, $h(y,s) = y$ for all $(y,s) \in \mathcal{Y}\times[0,1]$, and $h(x,1) \in \mathcal{Y}$ for all $x \in \mathcal{X}$.

For all $v_1, v_2, v_3 \in \mathbb{R}^n$, we set

$$A(v_1)[v_2] = (\nabla_{v_2} Y)(v_1) - (\nabla Y)^*[v_2](v_1) + \big\langle v_2, \nabla_{Y(v_1)} Y \big\rangle\, Y \tag{42}$$

and

$$B(v_1)[v_2, v_3] = -\Gamma(v_1)[v_2, v_3] - \big\langle v_2, \nabla_{v_3} Y \big\rangle_{v_1}, \tag{43}$$

where $\Gamma(q)[\cdot,\cdot] : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ is the bilinear map given by the Christoffel symbols at the point q of the metric g in local coordinates. Observe that $B(x)[z,z]$ is continuous in x and bilinear in z, while $A(x)[z]$ is continuous in x and linear in z. We now consider the map

$$G : C^0([0,1],\mathbb{R}^n) \times C^0([0,1],\mathbb{R}^n) \times \mathbb{R}^n \to C^0([0,1],\mathbb{R}^n) \times C^0([0,1],\mathbb{R}^n) \tag{44}$$

given by $G = (G_1, G_2)$, where

$$G_1(z,x,v)(t) = z(t) - v - \int_0^t [B(x)[z,z] + \Lambda(z,x)\, A(x)[z]]\, \mathrm{d}t, \tag{45}$$

$$G_2(z,x,v)(t) = x(t) - \int_0^t z\, \mathrm{d}s, \tag{46}$$

the maps A and B are defined in (42) and (43), and, finally,

$$\Lambda(z,x)(t) = e^{\int_0^t \langle \nabla_Y Y, z\rangle\, \mathrm{d}s} \cdot \left[\int_t^1 \big\langle z, \nabla_z Y \big\rangle e^{-\int_0^s \langle \nabla_Y Y, z\rangle\, \mathrm{d}r}\, \mathrm{d}s \right]. \tag{47}$$

Clearly, G is of class C^1, and $G(0,0,0) = 0$.

An elementary calculation shows that $(z, x, w_0) \in G^{-1}(0)$ if and only if x is of class C^2, $z = \dot{x}$, and x is a solution for the integro-differential problem (17)–(18) satisfying $x(0) = p$ and $\dot{x}(0) = z(0) = w_0$. Once we prove that the Jacobian

$$\frac{\partial G}{\partial (z,x)}(0,0,0) : C^0([0,1],\mathbb{R}^{2n}) \to C^0([0,1],\mathbb{R}^{2n}) \tag{48}$$

is invertible, there exists, for v sufficiently small, a C^1 map

$$v \mapsto (z_v, x_v) \tag{49}$$

such that

$$G(z_v, x_v, v) = 0. \tag{50}$$

To prove this, we differentiate formally the maps (45) and (46) at a generic point (z,x,v) in the direction (ω,ξ), obtaining

$$\begin{aligned} &\frac{\partial G_1}{\partial z}(z,x,v)[\omega] + \frac{\partial G_1}{\partial x}(z,x,v)[\xi] \\ &= \omega(t) - \int_0^t \left[\frac{\mathrm{d}B}{\mathrm{d}x}[\xi][z,z] + B(x)[\omega,z] + B(x)[z,\omega] + \frac{\partial \Lambda}{\partial z}[\omega]\, A(x)[z] \right] \mathrm{d}s \\ &\qquad - \int_0^t \left[\frac{\partial \Lambda}{\partial x}[\xi]\, A(x)[z] + \Lambda(z,x)\, \frac{\mathrm{d}A}{\mathrm{d}x}[\xi][z] + \Lambda(z,x)\, A(x)[\omega] \right] \mathrm{d}s \end{aligned} \tag{51}$$

and

$$\frac{\partial G_2}{\partial z}(z,x,v)[\omega] + \frac{\partial G_2}{\partial x}(z,x,v)[\xi] = \xi(t) - \int_0^t \omega \, \mathrm{d}s. \tag{52}$$

When we evaluate (51) and (52) at $(z,x,v) = (0,0,0)$, considering the linearity of the above functions in the variables between brackets and the fact that $\Lambda(0,0) = 0$, we obtain the following simple expression for the Jacobian (48):

$$\frac{\partial G}{\partial(z,x)}(0,0,0)[\omega,\xi] = \left(\omega, \xi - \int_0^t \omega \, \mathrm{d}s\right). \tag{53}$$

Clearly, this is an invertible map, whose inverse is easily computed as

$$\left(\frac{\partial G}{\partial(z,x)}(0,0,0)\right)^{-1}[\omega,\xi] = \left(\omega, \xi + \int_0^t \omega \, \mathrm{d}s\right). \tag{54}$$

This proves that the pair (z_v, x_v) and, in particular, the curve x_v depend regularly on v for v small enough such that (49) is well defined. □

By Proposition 4.1, for all $p \in M$ there exists a neighborhood $\mathcal{U}_p$ of 0 in T_pM such that for all $v \in \mathcal{U}_p$ the integro-differential equation (17)–(18) admits a unique solution x_v satisfying $x_v(0) = p$ and $\dot{x}_v(0) = v$. We can therefore define the following *exponential map* $\mathfrak{exp}_p : \mathcal{U}_p \to M$ by

$$\mathfrak{exp}_p(v) = x_v(1). \tag{55}$$

PROPOSITION 4.2. *For all $p \in M$, $\mathfrak{exp}_p$ is a local diffeomorphism between an open neighborhood of 0 in T_pM and an open neighborhood of p in M. In particular, $\mathfrak{exp}_p$ gives a local diffeomorphism between a neighborhood of 0 in Δ_p and a hypersurface Σ_p of M through p, with $T_p\Sigma_p = \Delta_p$, which is transversal to Y.*

Proof. From Proposition 4.1 we know that $\mathfrak{exp}_p$ is a map of class C^1 around $v = 0$; we now show that its differential $\mathrm{d}\mathfrak{exp}_p(0)$ is the *identity map* on T_pM.

From the implicit function theorem, the differential at $v = 0$ of the map $v \mapsto (z_v, x_v)$ is given by

$$-\left(\frac{\partial G}{\partial(z,x)}(0,0,0)\right)^{-1} \circ \frac{\partial G}{\partial v}(0,0,0),$$

which is easily computed from (45), (46), and (54) as

$$-\left(\frac{\partial G}{\partial(z,x)}(0,0,0)\right)^{-1}\left[\frac{\partial G}{\partial v}(0,0,0)[w]\right] = (w, w\cdot t) \quad \text{for all } w \in \mathbb{R}^n. \tag{56}$$

The derivative of $\mathfrak{exp}_p$ at $v = 0$ in the direction w is given by the evaluation at $t = 1$ of the second component of (56). Hence $\mathrm{d}\mathfrak{exp}_p(0)[w] = w$ and $\mathfrak{exp}_p$ is a local diffeomorphism in a neighborhood $\mathcal{U}_p$ of $0 \in T_pM$, which proves the first part of the statement.

Since Δ_p is a vector subspace of codimension one in T_pM, it follows that the image Σ_p of $\Delta_p \cap \mathcal{U}_p$ through $\mathfrak{exp}_p$ is a codimension one submanifold of M. Since $\mathrm{d}\mathfrak{exp}_p(0)$ is the identity map, it follows that $T_p\Sigma_p = \Delta_p$; moreover, since γ is transversal to Δ_p, by continuity every flow line of Y will be a transversal to Σ_p around the point p, and this concludes the proof. □

Proof of Theorem 1.9. If γ is sufficiently close to p, then, by the transversality, γ intercepts exactly once the hypersurface Σ_p of Proposition 4.2. The conclusion follows immediately from Propositions 1.1 and 4.1. ∎

Appendix A. The method of Lagrange multipliers. In this short appendix, we use the method of Lagrange multipliers to study the solutions of our variational problem, and we show that the map λ_x of formula equation (5) and the integro-differential equation (6) appear naturally also in this context.

We recall that $x \in \Omega_{p,\gamma}(\Delta)$ is a *constrained* critical point of E if and only if there exists a function $\lambda_x \in L^2([0,1],\mathbb{R})$ such that x is a *free* critical point in $\Omega_{p,\gamma}$ for the functional E_{λ_x} defined by

$$E_{\lambda_x}(z) = E(z) - \int_0^1 \lambda_x \cdot \langle \dot{z}, Y \rangle \, \mathrm{d}t. \tag{57}$$

In this case, the function λ_x is necessarily unique, and it is called the *Lagrange multiplier* associated to x.

We have the following.

LEMMA A.1. *Let $x \in \Omega_{p,\gamma}(\Delta)$ be a critical point for E. Then the corresponding Lagrange multiplier λ_x is a C^2-function given by*

$$\lambda_x(t) = e^{\int_0^t \langle \dot{x}, \nabla_Y Y \rangle \, \mathrm{d}s} \cdot \left[\int_t^1 \langle \dot{x}, \nabla_{\dot{x}} Y \rangle e^{-\int_0^s \langle \dot{x}, \nabla_Y Y \rangle \, \mathrm{d}r} \, \mathrm{d}s \right]. \tag{58}$$

Proof. The condition that x is a constrained critical point for E is that the following equation is satisfied for all $V \in T_x\Omega_{p,\gamma}$:

$$\int_0^1 \left[\langle \dot{x}, \nabla_{\dot{x}} V \rangle - \lambda_x \left(\langle \nabla_{\dot{x}} V, Y \rangle + \langle \dot{x}, \nabla_V Y \rangle \right) \right] \, \mathrm{d}t = 0. \tag{59}$$

Taking the covariant integral of $\lambda_x (\nabla Y)^*[\dot{x}]$ vanishing at $s = 1$, we see that $\dot{x} - \lambda_x Y$ is of class C^0. Taking its scalar product with Y, we obtain that λ_x is of class C^0, and therefore $\dot{x}$ is C^0. Repeating the above argument, we have also that $\dot{x}$ is C^1.

At this point, integration by parts in (59) of the terms containing the covariant derivative $\nabla_{\dot{x}} V$ gives

$$\int_0^1 \langle -\nabla_{\dot{x}} \dot{x} + \nabla_{\dot{x}} (\lambda_x \cdot Y) - \lambda_x (\nabla Y)^*[\dot{x}], V \rangle \, \mathrm{d}t - \lambda_x(1) \langle V(1), Y(x(1)) \rangle = 0. \tag{60}$$

Since $V(1)$ is arbitrary, it is easy to see that (60) is satisfied for all $V \in T_x\Omega_{p,\gamma}$ if and only if the following two equations are satisfied:

$$-\nabla_{\dot{x}} \dot{x} + \nabla_{\dot{x}} (\lambda_x \cdot Y) - \lambda_x \cdot (\nabla Y)^*[\dot{x}] = 0, \quad \lambda_x(1) = 0. \tag{61}$$

Taking the product of the differential equation in (61) by Y and considering that, since $\langle \dot{x}, Y \rangle \equiv 0$, it is

$$-\langle \nabla_{\dot{x}} \dot{x}, Y \rangle = \langle \dot{x}, \nabla_{\dot{x}} Y \rangle,$$

we obtain the following Cauchy problem for λ_x (recall that $\langle Y, Y \rangle = 1$):

$$\begin{cases} \lambda_x' - \lambda_x \cdot \langle \dot{x}, \nabla_Y Y \rangle + \langle \dot{x}, \nabla_{\dot{x}} Y \rangle = 0, \\ \lambda_x(1) = 0. \end{cases} \tag{62}$$

The unique solution of (62) is (58), and this concludes the proof. ∎

Appendix B. The system (8) is equivalent to Hamilton's equations of normal geodesics. Using the Pontryagin maximum principle, one proves (see, for instance, [7, Appendix B]) that a sub-Riemannian extremizer either is abnormal or satisfies the Hamilton equations of the Hamiltonian $H : TM^* \to \mathbb{R}$ given by

$$H(q,p) = \frac{1}{2} g^{-1}\big(p|_\Delta, p|_\Delta\big), \tag{63}$$

where $g^{-1} : \Delta^* \times \Delta^* \to \mathbb{R}$ is the metric induced by $g|_\Delta$ on Δ^*. In this appendix, we will show that (5) and (6) (or, equivalently, system (8)) are, in fact, equivalent to the Hamilton equations of H.

A convenient setup for this calculation is obtained by identifying the tangent bundle TM with the cotangent bundle TM^* using the metric g and considering the pull-back of the canonical symplectic form ω of TM^*. If $\pi_\Delta : TM \to \Delta$ denotes the orthogonal projection, the sub-Riemannian Hamiltonian (63) is given as a map $H : TM \to \mathbb{R}$ by

$$H = \frac{1}{2} \left\langle \pi_\Delta(u), \pi_\Delta(u) \right\rangle, \quad u \in TM;$$

since $\Delta = Y^\perp$ and $g(Y,Y) = 1$,

$$\pi_\Delta(u) = u - \langle u, Y \rangle\, Y$$

for all $u \in TM$.

Let us consider the horizontal subspace $T_{\mathrm{hor}}\big(TM\big)$ of $T(TM)$ determined by the Levi–Civita connection of g, and let us write

$$T(TM) = T_{\mathrm{hor}}\big(TM\big) \oplus T_{\mathrm{ver}}\big(TM\big),$$

where $T_{\mathrm{ver}}\big(TM\big)$ is the vertical subspace of $T(TM)$, i.e., the subspace tangent to the fibers of TM. For $v_1, v_2 \in T_{\mathrm{hor}}\big(TM\big)$ and $w_1, w_2 \in T_{\mathrm{ver}}\big(TM\big)$, the canonical symplectic form is given by

$$\omega\big((v_1,w_1),(v_2,w_2)\big) = \langle v_1, w_2 \rangle - \langle w_1, v_2 \rangle.$$

Given a smooth curve $(x,u) : [0,1] \to TM$, the Hamilton equations are then written as

$$\begin{cases} \dot{x} = \vec{H}_{\mathrm{hor}}, \\ \\ \nabla_{\dot{x}} u = \vec{H}_{\mathrm{ver}}, \end{cases} \tag{64}$$

where $\vec{H}$ is the Hamiltonian vector field defined by $\mathrm{d}H = \omega\big(\vec{H}, \cdot\big)$ and $\vec{H}_{\mathrm{hor}}$ and $\vec{H}_{\mathrm{ver}}$ are, respectively, its horizontal and vertical components in $T(TM)$. Moreover, a sub-Riemannian minimizer between a point $p \in M$ and an integral curve of γ must satisfy the boundary conditions

$$x(0) = p, \qquad u(1) \perp Y(x(1)). \tag{65}$$

In order to obtain an explicit expression for the components of $\vec{H}$, we compute as follows. For $u \in TM$, let $t \mapsto z(t)$ be a parallel vector field along a curve in M with velocity v at $t = 0$ and such that $z(0) = u$; then

$$\mathrm{d}H_u(v,0) = \frac{\mathrm{d}}{\mathrm{d}t} \left[\frac{1}{2} \big\langle z - \langle z, Y \rangle Y, z - \langle z, Y \rangle Y \big\rangle \right] = -\langle u, Y \rangle \, \langle u, \nabla_v Y \rangle;$$

moreover,

$$\mathrm{d}H(0,w) = \mathrm{d}\big(H|_{T_{\mathrm{ver}}(TM)}\big)(w) = \big\langle u - g(u,Y)\,Y, w\big\rangle.$$

Hence we get

$$\vec{H}_{\mathrm{ver}} = \langle u, Y\rangle\,(\nabla Y)^*[u], \quad \vec{H}_{\mathrm{hor}} = u - \langle u, Y\rangle\,Y,$$

and (64) becomes

$$\begin{cases} \dot{x} = u - \langle u, Y\rangle\,Y, \\ \\ \nabla_{\dot{x}} u = \langle u, Y\rangle\,(\nabla Y)^*[u]. \end{cases} \tag{66}$$

Setting $\lambda = -\langle u, Y\rangle$, we get from (66)

$$\nabla_{\dot{x}}\dot{x} - \lambda' Y - \lambda\nabla_{\dot{x}} Y = -\lambda\big((\nabla Y)^*[\dot{x}] + \lambda(\nabla Y)^*[Y]\big). \tag{67}$$

Now, since $g(Y,Y) \equiv 1$, we have $(\nabla Y)^*[Y] = 0$, and hence we get

$$\nabla_{\dot{x}}\dot{x} - \lambda' Y - \lambda\nabla_{\dot{x}} Y = -\lambda\,(\nabla Y)^*[\dot{x}]. \tag{68}$$

Multiplying (68) by Y and using the relation $\big\langle \nabla_{\dot{x}}\dot{x}, Y\big\rangle = -\big\langle \dot{x}, \nabla_{\dot{x}} Y\big\rangle$, we get

$$-\big\langle \dot{x}, \nabla_{\dot{x}} Y\big\rangle - \lambda' = -\lambda\big\langle \nabla_Y Y, \dot{x}\big\rangle. \tag{69}$$

Moreover, the boundary condition $\langle u(1), Y(x(1))\rangle = 0$ in (65) gives

$$\lambda(1) = 0; \tag{70}$$

now (68), (69), and (70) are clearly the same as (5) and (6).

Acknowledgments. We wish to thank the referees for their very useful comments and suggestions.

REFERENCES

[1] A. A. Agrachev and A. V. Sarychev, *Abnormal sub–Riemannian geodesics: Morse index and rigidity*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 6 (1996), pp. 635–690.
[2] H. Brezis, *Analyse Fonctionelle*, Masson, Paris, 1983.
[3] E. Fadell and S. Husseini, *Category of loop spaces of open subsets in Euclidean spaces*, Nonlinear Anal., 17 (1991), pp. 1153–1161.
[4] F. Giannoni and P. Piccione, *An existence theory for relativistic brachistochrones in stationary spacetimes*, J. Math. Phys., 39 (1998), pp. 6137–6152.
[5] F. Giannoni, P. Piccione, and J. A. Verderesi, *An approach to the relativistic brachistochrone problem by sub–Riemannian geometry*, J. Math. Phys., 38 (1997), pp. 6367–6381.
[6] S. Lang, *Differential Manifolds*, Springer-Verlag, Berlin, 1985.
[7] W. Liu and H. J. Sussmann, *Shortest Paths for Sub–Riemannian Metrics on Rank–2 Distribution*, Mem. Amer. Math. Soc., 118, (1995).
[8] R. Montgomery, *Singular extremals on Lie groups*, Math. Control Signals Systems, 7 (1994), pp. 217–234.
[9] R. Montgomery, *Abnormal minimizers*, SIAM J. Control Optim., 32 (1994), pp. 1605–1620.
[10] R. Montgomery, *A survey of singular curves in sub–Riemannian geometry*, J. Dynam. Control Systems, 1 (1995), pp. 49–90.
[11] J. Mawhin and M. Willem, *Critical Point Theory and Hamiltonian Systems*, Springer-Verlag, Berlin, 1989.
[12] P. Piccione and D. Tausk, *Variational aspects of the geodesic problem in sub–Riemannian geometry*, J. Geom. Phys., 39 (2001), pp. 183–206.
[13] S. A. Vakhrameev, *Morse theory and Lyusternik–Shnirelman theory in geometric control theory*, J. Math. Sci., 71 (1994), pp. 2434–2485.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 6, pp. 1858–1866

A CONSTRAINT ON THE MAXIMUM REFLECTANCE OF RAPIDLY OSCILLATING DIELECTRIC GRATINGS*

GANG BAO†, DAVID C. DOBSON‡, AND KARIM RAMDANI†

Abstract. When considering optimal design problems involving diffraction gratings, it is useful to have some a priori characterization of the range of possible reflectances one can achieve for given material parameters. Here we consider the limiting case of a rapidly oscillating dielectric grating and show that such gratings can have reflectance no greater than that of a flat interface, regardless of the shape of the grating interface.

Key words. diffraction grating, optimal design, maximum reflectance

AMS subject classifications. 78A45, 93B03

PII. S036301290037435X

1. Introduction.

A diffraction grating is formed by a periodic interface separating two homogeneous materials. In practical applications, one wishes to design the shape of the interface so that time-harmonic waves incident on the interface have a desired reflection and transmission pattern. Such design problems can be solved, for example, by optimization techniques [7] and homogenization [2]. An important question arising in this context is as follows: Given a particular class of admissible designs (interface shapes), which reflection and transmission patterns are attainable? In this paper we provide an answer in the case of *blazed* gratings (i.e., interfaces which can be represented by the graph of a function), which are rapidly oscillating with respect to the incident wavelength. The gratings are required to be dielectric. The basic result is a constraint on the reflectance, which says that in the limit as the grating period goes to zero the grating reflectance can be no greater than the reflectance obtained for a flat interface. This constraint holds regardless of the depth of the grating and the shape of the interface.

While rapidly oscillating gratings may seem to be of limited practical interest, they are, in fact, widely used. Optical engineers have been aware of homogenization effects in gratings for many years and often use high spatial frequency gratings to approximate corresponding multilayered structures (and vice-versa) [12]. The primary practical advantage of this approach is that material "layers" with intermediate refractive indices can be approximated by a grating composed of only two materials. In this way, the use of expensive, unstable, or nonexistent materials can often

*Received by the editors June 22, 2000; accepted for publication (in revised form) September 6, 2001; published electronically March 5, 2002. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Office of Scientific Research or the U.S. Government.

http://www.siam.org/journals/sicon/40-6/37435.html

†Department of Mathematics, Michigan State University, East-Lansing, MI 48824 (bao@math.msu.edu, ramdani@math.msu.edu). The research of the first author was partially supported by the NSF Applied Mathematics Programs grant DMS 98-03604 (99-96416), the NSF University-Industry Cooperative Research Programs grants DMS 98-03809 and DMS 99-72292, the NSF Western Europe Programs grant INT 98-15798, and the Office of Naval Research (ONR) grant N000140010299.

‡Department of Mathematics, Texas A&M University, College Station, TX 77843-3368 (dobson@math.tamu.edu). The research of this author was supported by the Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant F49620-98-1-0005. The research of this author was also supported by NSF grant DMS 0072439 and an Alfred P. Sloan Research Fellowship.

be avoided. One of the primary uses for rapidly oscillating gratings is in so-called *moth-eye* antireflective structures (see, e.g., [1, 8] and references therein), which are widely used to reduce glare on display devices and are commercially available.

Any optical engineer engaged in designing or optimizing a rapidly oscillating grating is faced with the question of whether or not a desired reflectance profile is attainable with given materials. This paper is aimed exactly at that question, showing that high reflectivity designs are generally *not* attainable with simple blazed high spatial frequency gratings. We consider the approach taken here as a first step toward solving the more difficult problem of characterizing attainable reflection/transmission patterns in more general diffraction gratings.

The plan of the paper is as follows. In the next section, we begin by analyzing the case of reflection from a "layered medium," i.e., a medium which has spatial dependence in only one direction. Under the condition that the refractive index of the medium is monotone in that direction, we establish the desired reflectance constraint. We conclude in section 3 by using homogenization theory to reduce the limiting case of a rapidly oscillating grating to the monotone layered medium. We prove that in the limit as the grating period goes to zero, the reflectance constraint is satisfied.

2. Layered medium case. We first consider a layered medium in $\mathbb{R}^2$, characterized by the real dielectric coefficient $k(x_2)$, where $x = (x_1, x_2) \in \mathbb{R}^2$. It is assumed that $k(x_2) \equiv k_a$ for $x_2 \geq 0$ (i.e., in the "air"), and $k(x_2) \equiv k_s$ for $x_2 \leq -b$, (i.e., in the "substrate"), where $0 < b < \infty$ is an arbitrary depth. Consider an incoming plane wave $u_i = e^{i\alpha x_1 + i\beta_a x_2}$, where

$$\alpha = k_a \sin\theta, \quad \beta_a = k_a \cos\theta, \tag{1}$$

and $|\theta| < \pi/2$ is the angle of incidence with respect to the x_2-axis. We wish to find solutions w satisfying the Helmholtz equation $\triangle w + k^2 w = 0$ in $\mathbb{R}^2$, plus appropriate outgoing wave conditions.

To make the problem independent of x_1, one can consider the functions $u = we^{-i\alpha x_1}$. Defining $\beta(x_2) = \sqrt{k(x_2)^2 - \alpha^2}$ and setting $\beta_s = \beta(-b)$, we specify the reflection and transmission conditions

$$\begin{aligned} u(x_2) &= e^{i\beta_a x_2} + re^{-i\beta_a x_2} \quad \text{for } x_2 \geq 0, \\ u(x_2) &= te^{i\beta_s x_2} \qquad\qquad\quad \text{for } x_2 \leq -b, \end{aligned} \tag{2}$$

where the coefficients r and t are to be determined. This leads to the following boundary value problem in x_2:

$$u'' + \beta^2 u = 0 \quad \text{in } (0, -b), \tag{3}$$

$$u'(0) = -i\beta_a u(0) + 2i\beta_a, \tag{4}$$

$$u'(-b) = i\beta_s u(-b). \tag{5}$$

In weak form, we have

$$\int_{-b}^{0} u'\overline{v}' - \int_{-b}^{0} \beta^2 u\overline{v} + i\beta_a u(0)\overline{v}(0) + i\beta_s u(-b)\overline{v}(-b) = 2i\beta_a \overline{v}(0). \tag{6}$$

LEMMA 2.1. *Let $\beta \in L^\infty(-b, 0)$ be real-valued. Then problem* (3)–(5) *admits a unique weak solution $u \in H^1(-b, 0)$.*

Proof. We seek $u \in H^1(-b, 0)$ such that (6) is satisfied for all $v \in H^1(-b, 0)$. It is easy to rewrite this problem as a linear operator equation $u - Au = f$, where A

is compact (see, e.g., [7]). Applying the Fredholm alternative, existence then follows from uniqueness for the homogeneous problem $w - Aw = 0$.

Thus it suffices to prove uniqueness for the homogeneous problem

$$w'' + \beta^2 w = 0 \quad \text{in } (0, -b), \tag{7}$$

$$w'(0) = -i\beta_a w(0), \tag{8}$$

$$w'(-b) = i\beta_s w(-b), \tag{9}$$

with associated weak form

$$\int_{-b}^{0} w'\overline{v}' - \int_{-b}^{0} \beta^2 w\overline{v} + i\beta_a w(0)\overline{v}(0) + i\beta_s w(-b)\overline{v}(-b) = 0. \tag{10}$$

Note that any solution $w \in H^1(-b, 0)$ of (10) is also in $H^2(-b, 0)$ since $w'' = -\beta^2 w$ a.e., and the right-hand side is in L^2. By Sobolev imbedding, $w \in C^1$. Setting $v = w$ in (10) and taking the imaginary part, we find that $w(-b) = w(0) = 0$. From (8), (9) we also have $w'(-b) = w'(0) = 0$. Uniqueness now follows by classical results for the Cauchy problem (see Hörmander [9, section 8.9] or Nirenberg [10]). □

We can now investigate the properties of the reflectance of a given structure defined by $\beta(x_2)$. First, given the solution u to (3)–(5), we define the reflection coefficient $r = u(0) - 1$, and the *reflectance* $\mathcal{R} = |r|^2$. The reflectance represents the proportion of incident energy reflected from the structure. Similarly, we define the transmission coefficient $t = u(-b)$, and the *transmittance* $\mathcal{T} = (\beta_s/\beta_a)|t|^2$. Setting $v = u$ and taking the imaginary part of the resulting equality in (6) yield *conservation of energy*:

$$\mathcal{R} + \mathcal{T} = 1. \tag{11}$$

Now taking $v = u'$ and applying the identities (3)–(5), one finds from (6) that

$$\beta_s^2|u(-b)|^2 - \beta_a^2\{|u(0)|^2 - 4\text{Re } u(0) - 4\} = \int_{-b}^{0} \beta^2(u'\overline{u} + u\overline{u}'). \tag{12}$$

Integrating the last term in (12) by parts, we have

$$\int_{-b}^{0} \beta^2(u'\overline{u} + u\overline{u}') = -\int_{-b}^{0} (\beta^2)'|u|^2 + \beta_a^2|u(0)|^2 - \beta_s^2|u(-b)|^2.$$

Then (12) becomes

$$2\beta_s^2|u(-b)|^2 - 2\beta_a^2|u(0) - 1|^2 - 2\beta_a^2 = -\int (\beta^2)'|u|^2. \tag{13}$$

Applying conservation of energy (11), $|t|^2 = (\beta_a/\beta_s)(1 - |r|^2)$ so that (13) yields

$$\mathcal{R} = \frac{\beta_s - \beta_a}{\beta_s + \beta_a} + \frac{1}{2\beta_a(\beta_a + \beta_s)} \int_{-b}^{0} (\beta^2)'|u|^2. \tag{14}$$

Since β^2 is nonincreasing, we immediately obtain that

$$\mathcal{R} \leq \frac{\beta_s - \beta_a}{\beta_s + \beta_a}.$$

The term on the right is the *square root* of the reflectance in the case of a flat profile (see, e.g., Born and Wolf [4] for a complete discussion of reflectance from flat interfaces). To improve this estimate, we need a lower bound on $|u|^2$.

LEMMA 2.2. *Suppose $\beta(x_2)$ is nonincreasing. Then the solution u of* (3)–(5) *satisfies*

$$|u|^2 \geq |t|^2,$$

where $t = u(-b)$ is the transmission coefficient.

Proof. First suppose that $k(x_2)$ is composed of a finite number of homogeneous layers, with refractive indices $k_a \leq k_1 \leq k_2 \leq \cdots \leq k_n \leq k_s$, with depths $h_1, \ldots, h_n$; i.e., setting $b_j = \sum_{k=1}^{j} h_k$, we have $k(x_2) = k_j$ for $-b_j \leq x_2 \leq -b_{j-1}$. Set $b = b_n$.

Letting $u(-b) = t$, the boundary condition (5) is $u'(-b) = i\beta_s t$. Solving for u in the nth layer, $-b_n \leq x \leq -b_{n-1}$, one obtains

$$u(x) = t(\cos\beta_n(x + b_n) + i(\beta_s/\beta_n)\sin\beta_n(x + b_n))e^{-i\beta_s b_n}.$$

Note that since $\beta_s/\beta_n \geq 1$, we have $|u(x)|^2 \geq |t|^2$. Having obtained u in terms of t the nth layer, one can now continue propagating the solution upward layer by layer, each time obtaining a solution in the form

$$u(x) = \tilde{t}_j(q_1 \cos\theta + iq_2(\beta_j/\beta_{j-1})\sin\theta),$$

where $\tilde{t}_j$ is a complex constant with $|\tilde{t}_j|^2 \geq |t|^2$, and q_1 and q_2 are complex constants in the form

$$\begin{aligned} q_1 &= \cos\phi + i(\beta_{j+1}/\beta_j)\sin\phi, \\ q_2 &= i\sin\phi + (\beta_{j+1}/\beta_j)\cos\phi. \end{aligned}$$

The result follows from the fact that the complex number $Z = q_1\cos\theta + iq_2(\beta_j/\beta_{j-1})\sin\theta$ satisfies $|Z| \geq 1$. Indeed, setting $\gamma_j = \beta_j/\beta_{j-1}$, we have

$$\begin{aligned} |Z|^2 &= (\cos\theta\cos\phi - \gamma_j\sin\theta\sin\phi)^2 + \gamma_{j+1}^2(\cos\theta\sin\phi + \gamma_j\sin\theta\cos\phi)^2 \\ &= \cos^2\theta(\cos^2\phi + \gamma_{j+1}^2\sin^2\phi) + \gamma_j^2\sin^2\theta(\sin^2\phi + \gamma_{j+1}^2\cos^2\phi) \\ &\quad + 2\gamma_j(\gamma_{j+1}^2 - 1)\cos\theta\cos\phi\sin\theta\sin\phi \\ &= \cos^2\theta(1 + (\gamma_{j+1}^2 - 1)\sin^2\phi) + \gamma_j^2\sin^2\theta(1 + (\gamma_{j+1}^2 - 1)\cos^2\phi) \\ &\quad + 2\gamma_j(\gamma_{j+1}^2 - 1)\cos\theta\cos\phi\sin\theta\sin\phi. \end{aligned}$$

Thus $|Z|^2$ can be written as

$$|Z|^2 = \cos^2\theta + \gamma_j^2\sin^2\theta + (\gamma_{j+1}^2 - 1)(\cos\theta\sin\phi + \gamma_j\sin\theta\cos\phi)^2.$$

Since β is a nonincreasing function, we have $\gamma_j = \beta_j/\beta_{j-1} \geq 1$, and thus $|Z| \geq 1$. Consequently, $|u(x)| \geq |t|^2$ in each layer.

In a manner exactly analogous to the procedure above, one can also obtain the estimate

$$|u(x)|^2 \leq |u(0)|^2 + |2 - u(0)|^2 \quad \text{for } x \leq 0. \tag{15}$$

Since $|u(0)-1|^2 = \mathcal{R} \le 1$, it follows that $\int_{-b}^{0} |u|^2 \le C$, where C is independent of the piecewise constant function β, provided only that β is nonincreasing. Taking the real part of the bilinear form (6) with $v = u$, we then find immediately that

$$\|u\|_{H^1} \le C, \tag{16}$$

where C now depends only on b, β_a, β_s.

The general case of nonincreasing $\beta \in L^\infty$ is now handled easily by approximation. Specifically, let $\{\beta^k\}$ be a sequence of nondecreasing, piecewise constant functions converging to a given β in the weak $*$ L^∞ sense. Let u_k be the corresponding sequence of solutions to (3)–(5). By (16), $\|u_k\|_{H^1} \le C$; hence there exists a subsequence (still denoted u_k) converging weakly in H^1 and strongly in L^2 to some $\tilde{u}$. It follows that, for every fixed $v \in H^1$,

$$\begin{aligned} 2i\beta_a \overline{v}(0) &= \int_{-b}^{0} u_k' \overline{v}' - \int_{-b}^{0} (\beta^k)^2 u_k \overline{v} + i\beta_a u_k(0)\overline{v}(0) + i\beta_s u_k(-b)\overline{v}(-b) \\ &\to \int_{-b}^{0} \tilde{u}' \overline{v}' - \int_{-b}^{0} \beta^2 \tilde{u}\overline{v} + i\beta_a \tilde{u}(0)\overline{v}(0) + i\beta_s \tilde{u}(-b)\overline{v}(-b) \end{aligned}$$

so that by Lemma 2.1, $\tilde{u} = u$, the unique solution to (3)–(5). Finally, since u_k, $k = 1, 2, \ldots$, along with u are uniformly bounded in H^2 and hence in C^1, we see that the convergence $u_k \to u$ is actually pointwise. Since the estimate $|u_k(x)|^2 \ge |t_k|^2$ holds for each k, it must also hold for u, t. ∎

LEMMA 2.3. *Suppose $\beta(x_2)$ is nonincreasing. Then*

$$\mathcal{R} \le \left(\frac{\beta_s - \beta_a}{\beta_s + \beta_a} \right)^2. \tag{17}$$

Proof. Using the previous lemma, we find that

$$\int_{-b}^{0} (\beta^2)' |u|^2 \le |t|^2 (\beta_a^2 - \beta_s^2).$$

Noting that $|t|^2 = (\beta_a/\beta_s)(1 - \mathcal{R})$, the identity (14) then yields the desired estimate with a simple manipulation. ∎

The estimate in Lemma 2.3 is sharp. Equality is attained for a sharp interface between two media with refractive indices k_a and k_s. Thus *the reflectance produced by any nonincreasing refractive index $k(x_2)$ with $k(-b) = k_s$ and $k(0) = k_a$ can be no more than the reflectance produced by the piecewise constant $k_c(x_2)$ with $k_c(x_2) = k_s$ for $x_2 < a$ and $k_c(x_2) = k_a$ for $x_2 > a$, $a \in (-b, 0)$.* Incidentally, it is well known in engineering that for a fixed incidence angle one can create a layered structure with $\mathcal{R}$ lying anywhere in the interval $[0, R_{max}]$, with $R_{max} = ((\beta_s - \beta_a)/(\beta_s + \beta_a))^2$. The key point here is that $\mathcal{R}$ cannot exceed R_{max} with nonincreasing β.

3. Rapidly oscillating case. We now consider the case of a rapidly oscillating dielectric grating. Specifically, suppose that we are given a grating structure with period L. By rescaling the problem, it suffices to consider the case $L = 2\pi$. Let $f \in L^\infty(\mathbb{R})$ be 2π-periodic; i.e., let

$$f(x_1) = f(x_1 + 2\pi n) \quad \text{a.e. in } x_1, \text{ for all integers } n$$

and satisfy

$$-b < \inf f \le \sup f < 0. \tag{18}$$

The function f represents an interface between two homogeneous materials with refractive indices k_a and k_s. Define a corresponding refractive index function on $\mathbb{R}^2$:

$$\rho_f(x) = \begin{cases} k_a^2 & \text{if } x_2 > f(x_1), \\ k_s^2 & \text{otherwise.} \end{cases} \tag{19}$$

As in the previous section, given an incoming plane wave from above $u_i = e^{i\alpha x_1 + i\beta_a(x_2)}$ (where α and β_a are as defined in (1)), we seek solutions of the Helmholtz equation $\triangle w + \rho_f w = 0$, where w is a sum of the incoming and scattered fields and satisfies appropriate outgoing wave conditions. The standard approach to solving this problem is to search for "quasi-periodic" solutions, that is, solutions w such that $u = we^{-i\alpha x_1}$ is 2π-periodic in x_1. A well-known procedure exists for formulating the problem variationally. This is outlined, for example, in [2, 7]. The basic idea is to expand the periodic functions u in a Fourier series in x_1 and match the solutions with the fundamental solution in the homogeneous regions $x_2 > 0$ and $x_2 < -b$. This leads naturally to a Fourier series expansion for the Dirichlet-to-Neumann operators on the boundaries $\{x_2 = 0\}$ and $\{x_2 = -b\}$. Defining the cylindrical domain $\Omega = (\mathbb{R} \times (-b, 0))/(2\pi Z \times \{0\})$ and the periodic boundaries Γ_a corresponding to $\{x_2 = 0\}$ and Γ_s corresponding to $\{x_2 = -b\}$, the problem can then be formulated as

$$\begin{aligned} \triangle_\alpha u + \rho_f u &= 0 \qquad \text{in } \Omega, \\ T_a u - \frac{\partial u}{\partial x_2} &= 2i\beta_a \quad \text{on } \Gamma_a, \\ T_s u + \frac{\partial u}{\partial x_2} &= 0 \qquad \text{on } \Gamma_s, \end{aligned}$$

where $\triangle_\alpha = \triangle + 2i\alpha\partial_1 - \alpha^2$. The Dirichlet-to-Neumann operators T_j are defined by

$$(T_j\phi)(x_1) = \sum_{n\in Z} i\beta_j^n \phi_n e^{inx_1}, \quad j = a, s,$$

where

$$\beta_j^n = \begin{cases} \sqrt{k_j^2 - (n+\alpha)^2} & \text{if } k_j^2 \ge (n+\alpha)^2, \\ i\sqrt{k_j^2 - (n+\alpha)^2} & \text{if } k_j^2 < (n+\alpha)^2, \end{cases}$$

and ϕ_n denote the Fourier coefficients of ϕ. To obtain the weak form, we define for $u, v \in H^1(\Omega)$

$$B_{\rho_f}(u, v) \equiv \int_\Omega (\nabla + i\underline{\alpha})u \cdot \overline{(\nabla + i\underline{\alpha})v} - \int_\Omega \rho_f\, u\overline{v} - \int_{\Gamma_a} (T_a u)\overline{v} + \int_{\Gamma_s} (T_s u)\overline{v},$$

where $\underline{\alpha} = (\alpha, 0)$ and

$$g(v) = -2i\beta_a \int_{\Gamma_a} \overline{v}.$$

We then wish to find $u \in H^1(\Omega)$ such that

$$B_{\rho_f}(u, v) = g(v) \quad \text{for all } v \in H^1(\Omega). \tag{20}$$

It is well known that a unique solution $u \in H^1(\Omega)$ of problem (20) exists for all but possibly a discrete set of parameters k_a, k_s (see [3]). In addition, using a perturbation argument, the following lemma is proved in [6].

LEMMA 3.1. *Provided that the incoming wave is sufficiently low-frequency (k_a, k_s are sufficiently small), problem* (20) *admits a unique weak solution for all f satisfying* (18). *Furthermore, the solutions u are bounded in $H^1(\Omega)$ independently of f.*

Remark. Under the conditions of Lemma 3.1, solutions are actually uniformly bounded in H^2, independent of f. This follows immediately from the equation

$$\triangle u = -2i\alpha\partial_1 u + (\alpha^2 - \rho_f)u.$$

The H^1 bound on u and the L^∞ bound on ρ_f guarantee that $\|\triangle u\|_{L^2} \leq C$, giving the H^2 estimate.

Once the solution to problem (20) has been determined, one can easily find the scattered far-field. The *Rayleigh expansion* [11] dictates that the field above $\{x_2 = 0\}$ must be in the form

$$u(x_1, x_2) = \sum_{n=-\infty}^{\infty} r_n e^{i(nx_1 - \beta_a^n x_2)},$$

where the r_n are unknown scalars. Matching this expansion with the boundary conditions for the variational solution, one finds that r_0, which corresponds to the "zero order" reflected mode, must be given by

$$r_0 = \frac{1}{2\pi}\int_0^{2\pi} u(x_1, 0)\, dx_1 - 1. \tag{21}$$

By rescaling the problem, one can see easily that for a sufficiently small grating period L the coefficients β_a^n are real only for $n = 0$. This means that only the zero order mode propagates. Similarly, using the Rayleigh expansion in the region $x_2 < -b$ and the fact that the grating period L is small, one finds the lone transmitted mode $t_0 = \frac{1}{2\pi}\int_0^{2\pi} u(x_1, -b)\, dx_1$.

As in the layered medium case, setting the reflectance $\mathcal{R}_0 = |r_0|^2$ and the transmittance $\mathcal{T}_0 = |t_0|^2$, one can easily verify conservation of energy $\mathcal{R}_0 + \mathcal{T}_0 = 1$ [7]. We would like to show that a reflectance bound similar to (17) holds in the grating case.

For $n = 1, 2, \ldots$, define $\rho_n(x_1, x_2) = \rho_f(nx_1, x_2)$. Thus ρ_n represents a 2π-periodic grating oscillating more and more rapidly as n increases. It is easily verified that $\rho_n \rightharpoonup \tilde{\rho}$ in the weak $*$ $L^\infty(\Omega)$ sense, where

$$\tilde{\rho}(x_1, x_2) = \frac{1}{2\pi}\int_0^{2\pi} \rho_f(x_1, x_2)\, dx_1.$$

Note that $\tilde{\rho}$ is independent of x_1. Furthermore, due to the form of ρ_f (19) and the fact that $k_a^2 \leq k_s^2$, it is easy to see that $\tilde{\rho}$ is nonincreasing in x_2.

We can now state the main result of this paper.

THEOREM 3.2. *Assume the conditions of Lemma* 3.1. *Given an arbitrary grating profile f and any $\epsilon > 0$, there exists a grating period L such that when the profile f is produced with period L or less, the reflectance $\mathcal{R}_0$ resulting from f satisfies*

$$\mathcal{R}_0 \leq \left(\frac{\beta_s^0 - \beta_a^0}{\beta_s^0 + \beta_a^0}\right)^2 + \epsilon. \tag{22}$$

Thus, analogous to the layered medium case, for rapidly oscillating gratings the reflectance can be no more than the reflectance of a sharp interface between materials k_a and k_b plus a small error, regardless of the grating shape.

Proof. Since the bound (17) holds for $\tilde{\rho}$, inequality (22) is simply a statement of the continuity of $\mathcal{R}_0(\rho)$ with respect to weak $*$ L^∞ convergence $\rho_n \rightharpoonup \tilde{\rho}$. This is easy to prove. Let u_n denote the sequence of solutions to problem (20) corresponding to the coefficients ρ_n. By Lemma 3.1, $\|u_n\|_{H^1}$ is uniformly bounded; hence each subsequence of $\{u_n\}$ has a further subsequence $\{u_{n'}\}$ which converges weakly in H^1 to some $u \in H^1$. We show that the weak limit u of each such subsequence is the same, thus proving that the original sequence $\{u_n\}$ converges weakly to u.

Holding $v \in H^1$ fixed, observe that

$$B_{\rho_{n'}}(u,v) - B_{\rho_{n'}}(u_{n'},v) = \int_\Omega (\nabla + i\underline{\alpha})(u-u_{n'}) \cdot \overline{(\nabla + i\underline{\alpha})v} - \int_\Omega \rho_{n'}(u-u_{n'})\overline{v} - \int_{\Gamma_a} (T_a(u-u_{n'}))\overline{v} + \int_{\Gamma_s} (T_s(u-u_{n'}))\overline{v}.$$

Since $u_{n'} \rightharpoonup u$ in H^1 and the operators T_j are bounded maps from $H^{1/2}(\Gamma_j)$ into $H^{-1/2}(\Gamma_j)$, the first integral and the last two integrals vanish as $n' \to \infty$. Further, the weak convergence of $u_{n'}$ in H^1 implies strong convergence in L^2 so that

$$\left| \int_\Omega \rho_{n'}(u-u_{n'})\overline{v} \right| \le \|\rho_{n'}\|_{L^\infty} \|u-u_{n'}\|_{L^2} \|v\|_{L^2} \to 0.$$

Thus $B_{\rho_{n'}}(u,v) \to B_{\rho_{n'}}(u_{n'},v) = g(v)$. The convergence $\rho_{n'} \overset{*}{\rightharpoonup} \tilde{\rho}$ implies $B_{\rho_{n'}}(u,v) \to B_{\tilde{\rho}}(u,v)$. Hence $B_{\tilde{\rho}}(u,v) = g(v)$ for all v; i.e., u solves (20) for $\tilde{\rho}$. Since the solution u is unique by Lemma 3.1, we conclude that the original sequence $u_n \rightharpoonup u$ weakly in H^1.

Since the traces $u_n|_{\Gamma_a} \rightharpoonup u|_{\Gamma_a}$ weakly in $H^{1/2}$, it follows by the definition (21) that the corresponding reflection coefficients, and hence the reflectances, converge. ∎

Acknowledgments. The authors wish to thank Jeff Rauch for raising the question which led to this work. They are also very grateful to the referees and the associate editor for their helpful comments, which improved the paper greatly.

REFERENCES

[1] M. Auslender, D. Levy, and S. Hava, *Design and analysis of antireflection grating structure for solar energy absorber*, in Optical Materials Technology for Energy Efficiency and Solar Energy Conversion XV, C.M. Lampert, C.-G. Granqvist, M. Graetzel, and S. Deb, eds., Proc. SPIE 3138, SPIE, Bellingham, WA, 1997, pp. 160–165.

[2] G. Bao and E. Bonnetier, *Optimal design of periodic diffractive structures in TM polarization*, Appl. Math. Optim., 43 (2001), pp. 103–116.

[3] G. Bao, D. C. Dobson, and J. A. Cox, *Mathematical studies of rigorous grating theory*, J. Opt. Soc. Amer. A, 12 (1995), pp. 1029–1042.

[4] M. Born and E. Wolf, *Principles of Optics, Electromagnetic Theory of Propagation Interference and Diffraction of Light*, 6th ed., Pergamon Press, Oxford, UK, 1980.

[5] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[6] D. C. Dobson, *Optimal design of periodic antireflective structures for the Helmholtz equation*, European. J. Appl. Math., 4 (1993), pp. 321–340.

[7] D. C. Dobson, *Optimal shape design of blazed diffraction gratings*, Appl. Math. Optim., 40 (1999), pp. 61–78.

[8] A. B. Harker and J. F. Denatale, *Diamond gradient index "moth-eye" antireflection surfaces for LWIR windows,* in Window and Dome Technologies and Materials III, Paul Klocek, ed., Proc. SPIE 1760, Bellingham, WA, 1992, pp. 261–267.

[9] L. Hörmander, *Linear Partial Differential Operators*, Springer-Verlag, Berlin, 1969.

[10] L. Nirenberg, *Uniqueness in Cauchy problems for differential operators with constant leading coefficients*, Comm. Pure Appl. Math., 10 (1957), pp. 89–105.

[11] R. Petit, ed., *Electromagnetic Theory of Gratings*, Topics in Current Physics 22, Springer-Verlag, Berlin, New York, 1980.

[12] R. Petit and G. Bouchitté, *Replacement of a very fine grating by a stratified layer: Homogenization techniques, and the multiple scale method*, in Proceedings of the International Conference on the Application and Theory of Periodic Structures, Diffraction Gratings, and Moire Phenomena III, Jeremy M. Lerner, ed., Proc. SPIE 815, Bellingham, WA, 1988, pp. 25–31.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 6, pp. 1867–1887

EXISTENCE CONDITIONS AND PROPERTIES OF THE FREQUENCY RESPONSE OPERATORS OF CONTINUOUS-TIME PERIODIC SYSTEMS*

JUN ZHOU† AND TOMOMICHI HAGIWARA†

Abstract. The definition of the frequency response operator via the steady-state analysis in finite-dimensional linear continuous-time periodic (FDLCP) systems is revisited. It is shown that the frequency response operator is guaranteed to be well defined only densely on the linear space l_2, which is different from the usual understanding. Fortunately, however, it turns out that this frequency response operator can have an extension onto l_2 so that the equivalence between the time-domain H_2 norm (respectively, the L_2-induced norm) and the frequency-domain H_2 norm (respectively, the H_∞ norm of the frequency response operator) is recovered. Under some stronger assumptions, it is also shown that the frequency response operator can be viewed as a bounded operator from l_1 to l_1, which can also be established via the steady-state analysis.

Key words. continuous-time periodic system, frequency response operator, existence conditions, H_2 and H_∞ norms

AMS subject classifications. 42A20, 47A05, 47B99, 47N70

PII. S0363012900382357

1. Introduction.

In general there are two ways to establish the frequency response relations in finite-dimensional linear continuous-time periodic (FDLCP) systems: the lifting technique [2] and the steady-state input-output analysis [26], [27]. The former technique [3], [30] has been widely used in the sampled-data systems analysis [5], [15], [28], while the latter is the standard technique of the frequency response analysis in linear continuous-time systems, and there are also works in which the latter is utilized in establishing the frequency response operator, or FR-operator, in sampled-data systems [1], [8], [9], [11], [14], [16]. There are also some works [29], [32] in which these two approaches are compared and/or combined.

In this paper, we reconsider the problem of establishing the frequency response relation of FDLCP systems through the steady-state input-output analysis. The general idea of such a treatment has been discussed in [27], but the rigorous existence conditions of the frequency response operators are not given explicitly. To clarify these conditions, we concentrate our attention on the validity of the Fourier series expansions and the Toeplitz transformation employed in the arguments. Our study reveals that the frequency response operator defined with the steady-state input-output analysis is guaranteed to be well defined only on the subset l_1 of the linear space l_2 rather than on the whole l_2 even though the latter is an implicit assumption in [26], [27]. This subtle deviation of the domains of the frequency response operator leads us to put up such a question: if it still makes sense to define the H_2 and H_∞ norms on such a "deficient" frequency response operator for FDLCP systems. Fortunately, our study shows that the domain of the frequency response operator can actually be extended to l_2 apart from the steady-state analysis interpretation, and hence the existing definitions

*Received by the editors December 12, 2000; accepted for publication (in revised form) October 1, 2001; published electronically March 5, 2002. A preliminary version of this paper appeared as Tech report 00-05, Automatic Control Engineering Group, Department of Electrical Engineering, Kyoto University, 2000.

http://www.siam.org/journals/sicon/40-6/38235.html

†Department of Electrical Engineering, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan (zhouj@jaguar.kuee.kyoto-u.ac.jp).

of these norms suffer no problems. In fact, the equivalence between the time-domain and frequency-domain H_2 norms and that between the L_2-induced norm and the H_∞ norm (i.e., the maximum of the l_2-induced norm over a certain frequency interval) of the frequency response operator are proved under some stronger assumptions on the system matrices. Another important point of our study is to show that the frequency response operator can also be regarded as a bounded mapping defined on the linear space l_1 (with its range also contained in l_1) under some stronger assumptions.

The paper is organized as follows. Section 2 gives some mathematical preliminaries about FDLCP systems, the Fourier series expansions, and the Toeplitz transformation of h-periodic time-varying matrix functions. The definition of the frequency response operator of FDLCP systems is reconsidered in section 3. In section 4, we prove that the frequency response operator via the steady-state analysis is also a bounded mapping on l_1 under some strengthened conditions. The equivalences of the various norms are dealt with in section 5.

The notation of this paper is standard. The Euclidean norm of a vector and the norm of a matrix induced by this vector norm are denoted by $||\cdot||$. l_1 is the set of all infinite-dimensional vectors $\underline{x}$ such that $||\underline{x}||_{l_1} := \sum_{m=-\infty}^{+\infty} ||[\underline{x}]_m|| < \infty$, where $[\underline{x}]_m$ is the mth (block) entry of $\underline{x}$. l_2 is the set of all infinite-dimensional vectors $\underline{x}$ such that $||\underline{x}||_{l_2} := (\underline{x}^*\underline{x})^{1/2} < \infty$, where $*$ denotes the complex conjugate transpose. $L_2[0,h]$ is the linear space of all vector measurable functions x defined on the interval $[0,h]$ such that $||x(\cdot)||_{L_2[0,h]} := [\int_0^h ||x(t)||^2 dt]^{1/2} < \infty$. L_1 is the set of all vector measurable functions x defined on $[0,\infty)$ such that $||x(\cdot)||_{L_1} := \int_0^\infty ||x(t)||dt < \infty$, while L_2 is the set of all such defined vector functions x satisfying $||x(\cdot)||_{L_2} := [\int_0^\infty ||x(t)||^2 dt]^{1/2} < \infty$. $||\cdot||_{Y/X}$ denotes the induced norm from X to Y. In particular, $||\cdot||_{l_2/l_2}$ is the l_2-induced norm. With a little abuse of notation, we say $F(t) \in L_2[0,h]$ means that F is a matrix function, each element of which is h-periodic and belongs to $L_2[0,h]$ when its domain is restricted to the interval $[0,h]$. The same is true for other function sets defined over $[0,h]$. $\mathcal{Z}$ is the set of all integers.

2. Preliminaries. Consider the FDLCP system

$$G : \begin{cases} \dot{x} = A(t)x + B(t)u, \\ y = C(t)x + D(t)u, \end{cases} \tag{1}$$

where $A(t), B(t), C(t)$, and $D(t)$ are h-periodic time-varying matrix functions belonging to $L_2[0,h]$. The transition matrix of the system (1) is denoted by $\Phi(t,t_0)$ when the initial time is t_0. The system is said to be strictly proper if $D(t) \equiv 0$ for all $t \in [0,h]$.

PROPOSITION 2.1 (Floquet theorem [10], [20], [22], [23]). *Let $A(t)$ be defined as in the system* (1). *Then the transition matrix $\Phi(t,t_0)$ is continuous with respect to t and can be expressed as $\Phi(t,t_0) = P(t,t_0)e^{Q(t-t_0)}$, where $P(t,t_0)$ is a nonsingular h-periodic matrix and Q is a constant matrix. Moreover, the system is asymptotically stable if and only if the eigenvalues of the monodromy matrix, $\Phi(h+t_0,t_0)$, are in the open unit disk or, equivalently, the eigenvalues of Q lie in the open left-half plane.*

Now expand $A(t)$ to its Fourier series $A(t) = \sum_{m=-\infty}^{+\infty} A_m e^{jm\omega_h t}$ with $\omega_h = \frac{2\pi}{h}$, which is well defined in the sense that $||A(\cdot) - \sum_{m=-\infty}^{+\infty} A_m e^{jm\omega_h(\cdot)}||_{L_2[0,h]} = 0$. The Toeplitz transformation on $A(t)$ [27], denoted by $\mathcal{T}\{A(t)\}$, maps $A(t)$ into a doubly infinite-dimensional block Toeplitz operator [26] (or to be more precise, block Laurent operator [13]) of the form

$$\mathcal{T}\{A(t)\} := \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \iddots \\ \cdots & A_0 & A_{-1} & A_{-2} & \cdots \\ \cdots & A_1 & A_0 & A_{-1} & \cdots \\ \cdots & A_2 & A_1 & A_0 & \cdots \\ \iddots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} =: \underline{A}. \tag{2}$$

It is straightforward to show that $\mathcal{T}\{X(t)+Y(t)\} = \mathcal{T}\{X(t)\} + \mathcal{T}\{Y(t)\}$ when $X(t)$ and $Y(t)$ are h-periodic and belong to $L_2[0,h]$. However, the situation is different for the Toeplitz transformation of the product of two matrix functions. Now consider the two conformable h-periodic matrix functions $X(t)$ and $Y(t)$ with the Fourier series expansions $X(t) = \sum_{m=-\infty}^{+\infty} X_m e^{jm\omega_h t}$ and $Y(t) = \sum_{m=-\infty}^{+\infty} Y_m e^{jm\omega_h t}$, respectively. To facilitate our statements, we introduce the sets of h-periodic functions, each of which is a subset of $L_2[0,h]$.

$$L_{\mathrm{PCD}}[0,h] := \left\{ f(t) : \begin{array}{l} f(t) \text{ is piecewise continuous and} \\ \text{differentiable at a.e. } t \in [0,h] \end{array} \right\},$$

$$L_{\mathrm{PCC}}[0,h] := \left\{ f(t) : \begin{array}{l} f(t) \text{ is piecewise continuous and the Fourier series} \\ \text{expansion of } f(t) \text{ is convergent to } f(t_0) \text{ for a.e. } t_0 \in [0,h] \end{array} \right\},$$

$$L_{\mathrm{CAC}}[0,h] := \left\{ f(t) : \begin{array}{l} f(t) \text{ is continuous and the Fourier series} \\ \text{expansion of } f(t) \text{ is absolutely convergent} \end{array} \right\} \subset L_{\mathrm{PCC}}[0,h],$$

$$L_{\mathrm{CPCD}}[0,h] := \left\{ f(t) : \begin{array}{l} f(t) \text{ is continuous and the derivative of} \\ f(t) \text{ is piecewise continuous in } [0,h] \end{array} \right\} \subset L_{\mathrm{PCD}}[0,h],$$

where PCD stands for piecewise continuous and differentiable and PCC is short for piecewise continuous and convergent, while CAC and CPCD are abbreviated from continuous and absolute convergent and continuous and piecewise continuously (first-order) differentiable, respectively. By Theorem 10′ of [18, p. 173], $L_{\mathrm{CPCD}}[0,h] \subset L_{\mathrm{PCD}}[0,h] \subset L_{\mathrm{PCC}}[0,h]$. It is also clear that $L_{\mathrm{CAC}}[0,h] \subset L_{\mathrm{PCC}}[0,h]$. The following results are helpful in our subsequent arguments. The proofs for these lemmas are given in Appendix B.

LEMMA 2.1. *Suppose that the Fourier series expansion of $X(t)$ converges to $X(t_0)$ for almost every (a.e.) $t_0 \in [0,h]$. Also suppose that $Y(t) \in L_{\mathrm{CAC}}[0,h]$. Then $\mathcal{T}\{X(t)Y(t)\} = \mathcal{T}\{X(t)\}\mathcal{T}\{Y(t)\}$ and $\mathcal{T}\{Y(t)X(t)\} = \mathcal{T}\{Y(t)\}\mathcal{T}\{X(t)\}$.*

Remark 2.1. In [27, p. 36], the absolute convergence of the Fourier series expansions of both $X(t)$ and $Y(t)$ is required to ensure the validity of the relations of Lemma 2.1. However, for some reasons unknown from this thesis [27], this constraint was not taken into consideration when the frequency response relation is established with the Toeplitz operator expressions. Hence, the frequency response operator has been introduced to a larger class of FDLCP systems than a precise class for which it can be introduced in a rigorous manner. Our study in the following will pay attention to such a precise class while trying to make the class as large as possible, which leads us to a little different but rigorous interpretation about the frequency response operator of an FDLCP system. In the development of such an argument, it is quite important to note that the conditions of Lemma 2.1 are much weaker than the absolute convergence condition of both $X(t)$ and $Y(t)$.

LEMMA 2.2. *$L_{\mathrm{CAC}}[0,h]$ is dense in $L_2[0,h]$.*

LEMMA 2.3. *If $X(t) \in L_{\mathrm{PCC}}[0,h]$, then $||\underline{X}||_{l_2/l_2} = \sup_{t\in[0,h]} ||X(t)||$ and $\underline{X}$ is bounded on l_2.*

Now we are in a position to derive Propositions 2.2 and 2.3, which describe the basic properties of the FDLCP system (1) in the Toeplitz operator sense and guarantee the existence of the frequency response operator we will introduce.

PROPOSITION 2.2. *Assume that in the system* (1) *the state matrix* $A(t)$ *is piecewise continuous in* $[0,h]$*, and let* $\mathcal{T}\{P(t,0)\} =: \underline{P}$*. Then the Fourier series expansions of* $P(t,0)$ *and* $P^{-1}(t,0)$ *are absolutely convergent, and* $\mathcal{T}\{P^{-1}(t,0)\} = \underline{P}^{-1}$*.*

Proof. By Theorem 6.3.2 of [20] and the Floquet theorem, it follows that

$$(3)\quad \begin{cases} P(t,0) = \Phi(t,0)e^{-Qt}, & \dfrac{d}{dt}P(t,0) = [A(t)\Phi(t,0) - \Phi(t,0)Q]e^{-Qt} \ \text{(a.e.)}, \\ P^{-1}(t,0) = e^{Qt}\Phi(0,t), & \dfrac{d}{dt}P^{-1}(t,0) = e^{Qt}[Q\Phi(0,t) - \Phi(0,t)A(t)] \ \text{(a.e.)}, \end{cases}$$

which clearly says that $P(t,0)$ and $P^{-1}(t,0)$ are continuous and their first-order derivatives are piecewise continuous by the assumption on $A(t)$. Hence, by Theorem 2 of [6, p. 104], the Fourier series expansions of $P(t,0)$ and $P^{-1}(t,0)$ are absolutely convergent. On the other hand,

$$P(t,0)P^{-1}(t,0) = I \quad \forall t \in [0,h].$$

Hence, from Lemma 2.1, applying the Toeplitz transformation on the above equation gives

$$\underline{I} = \mathcal{T}\{P(t,0)P^{-1}(t,0)\} = \mathcal{T}\{P(t,0)\}\mathcal{T}\{P^{-1}(t,0)\}.$$

Similarly, $\underline{I} = \mathcal{T}\{P^{-1}(t,0)\}\mathcal{T}\{P(t,0)\}$. Hence we have $\mathcal{T}\{P^{-1}(t,0)\} = \underline{P}^{-1}$ by the uniqueness of the inverse operator [7]. □

Next let us define $l_E := \{\underline{x} \in l_2 : \underline{E}(j0)\underline{x} \in l_2\} \subset l_2$ with

$$\underline{E}(j0) = \mathrm{diag}[\ldots, -j2\omega_h I, -j\omega_h I, 0, j\omega_h I, j2\omega_h I, \ldots]$$

and establish the following results.

PROPOSITION 2.3. *Assume that in* (1) $A(t) \in L_{\mathrm{PCD}}[0,h]$ *and* $B(t), C(t) \in L_{\mathrm{PCC}}[0,h]$*. Then* l_E *is* $\underline{P}$*-,* $\underline{P}^{-1}$*-,* $\underline{P}^*$*-, and* $\underline{P}^{-*}$*-invariant, where* $\underline{P}^{-*} := [\underline{P}^{-1}]^*$*.* $\underline{P}$ *is invertible on* l_E*, and the unique inverse of* $\underline{P}$ *on* l_E *is* $\underline{P}^{-1}$ *restricted to* l_E*. It holds on* $l_E \subset l_2$ *that*

$$(4)\qquad \underline{P}(\underline{E}(j0) - \underline{Q})\underline{P}^{-1} = \underline{E}(j0) - \underline{A},$$

where $\underline{Q} = \mathcal{T}\{Q\}$*. Moreover, let* $\hat{\underline{B}} := \mathcal{T}\{P^{-1}(t,0)B(t)\}$ *and* $\hat{\underline{C}} := \mathcal{T}\{C(t)P(t,0)\}$*. Then it holds on the whole* l_2 *that* $\hat{\underline{B}} = \underline{P}^{-1}\underline{B}$ *and* $\hat{\underline{C}} = \underline{C}\,\underline{P}$*.*

Proof. By the Floquet theorem and Theorem 6.3.2 of [20], we obtain

$$(5)\qquad P(t,0)Q = A(t)P(t,0) - \dot{P}(t,0) \quad \text{(a.e.)}.$$

By Proposition 2.2, the Fourier series expansion of $P(t,0)$ is absolutely convergent. Note also that by Theorem $10'$ of [18, p. 173], the Fourier series expansion of $A(t)$ converges to $A(t_0)$ for a.e. $t_0 \in [0,h]$ from the assumption. Hence, by Lemma 2.1, we have

$$(6)\qquad \mathcal{T}\{A(t)P(t,0)\} = \mathcal{T}\{A(t)\}\mathcal{T}\{P(t,0)\}.$$

Again by (3) and the assumption about $A(t)$, the first-order derivative of $P(t,0)$ is piecewise continuous, and its second-order derivative exists a.e. in $[0,h]$. Thus, by Theorem 3 of [6, p. 106],

$$\dot{P}(t,0) = \sum_{m=-\infty}^{+\infty} jm\omega_h P_m e^{jm\omega_h t} \quad \text{(a.e.)} \tag{7}$$

through the termwise differentiation, where $\{P_m\}_{m=-\infty}^{+\infty}$ is the Fourier coefficients sequence of $P(t,0)$. In other words, $\{jm\omega_h P_m\}_{m=-\infty}^{+\infty}$ is the Fourier coefficients sequence of $\dot{P}(t,0)$ so that by some trivial algebra [27] we are led to

$$\mathcal{T}\{\dot{P}(t,0)\} = \underline{E}(j0)\underline{P} - \underline{P}\,\underline{E}(j0). \tag{8}$$

Note that $\mathcal{T}\{\dot{P}(t,0)\}$ is bounded on l_2 (which follows from Lemma 2.3 since $\dot{P}(t,0)$ belongs to $L_{\mathrm{PCC}}[0,h]$ by the assumption on $A(t)$, again from Theorem 10′ of [18, p. 173]) but that the two operators on the right-hand side of the above equation are unbounded since $\underline{E}(j0)$ is. This means that we are not allowed to use the operators $\underline{E}(j0)\underline{P}$ and $\underline{P}\,\underline{E}(j0)$ separately if the underlying space is l_2. To get around the problem, we have to restrict the domain of these operators to $l_E \subset l_2$. Now take $\underline{x} \in l_E \subset l_2$. Then $\mathcal{T}\{\dot{P}(t,0)\}\underline{x} \in l_2$. Also, $\underline{P}\,\underline{E}(j0)\underline{x} \in l_2$ since $\underline{E}(j0)\underline{x} \in l_2$ and $\underline{P}$ is bounded on l_2 (which follows again from Lemma 2.3 by the fact that the Fourier series expansion of $P(t,0)$ is absolutely convergent). It follows that $\underline{E}(j0)\underline{P}\,\underline{x} \in l_2$, which clearly says that l_E is $\underline{P}$-invariant.

Similarly, by repeating the arguments about $\dot{P}(t,0)$ on $\dot{P}^{-1}(t,0)$, it readily follows that l_E is also $\underline{P}^{-1}$-invariant. Hence $\underline{P}$ and $\underline{P}^{-1}$ are mappings on l_E. From this, it can be asserted that $\underline{P}$ is invertible on l_E, and the unique inverse of $\underline{P}$ on l_E is nothing but $\underline{P}^{-1}$ restricted to $l_E \subset l_2$.

On the other hand, (6) and (8) actually say that the Toeplitz transformation applies to each term of (5) under the given assumptions so that we obtain

$$\underline{P}\,\underline{Q} = \underline{A}\,\underline{P} - \underline{E}(j0)\underline{P} + \underline{P}\,\underline{E}(j0).$$

Therefore, if we work on l_E instead of l_2, the operators involved are well defined from l_E to l_2; i.e., the above equation can be rewritten as

$$\underline{P}(\underline{E}(j0) - \underline{Q}) = (\underline{E}(j0) - \underline{A})\underline{P}, \tag{9}$$

which, together with the fact that $\underline{P}$ is invertible on l_E, gives (4).

To see that l_E is $\underline{P}^*$-invariant, we note that $\underline{P}^* = \mathcal{T}\{P^*(t,0)\}$. It is evident from the assumption about $A(t)$ that the first-order derivative of $P^*(t,0)$ is piecewise continuous in $[0,h]$ and the second-order derivation of $P^*(t,0)$ exists a.e. in $[0,h]$. Then from Theorem 3 of [6, p. 106], it is true that

$$\mathcal{T}\{\dot{P}^*(t,0)\} = \underline{E}(j0)\underline{P}^* - \underline{P}^*\underline{E}(j0),$$

which gives the assertion immediately. Similarly, one can show that l_E is $\underline{P}^{-*}$-invariant.

Recall that the Fourier series expansion of $P^{-1}(t,0)$ is absolutely convergent. This, by the assumption on $B(t)$ and Lemma 2.1, implies that $\mathcal{T}\{P^{-1}(t,0)B(t)\} = \mathcal{T}\{P^{-1}(t,0)\}\mathcal{T}\{B(t)\}$. Combining this with Proposition 2.2, the last assertion follows. □

Now we answer the question that asks under what conditions the operator $\underline{E}(j0)-\underline{A}$ is invertible. This is another problem remaining unsolved in the works of [26], [27]. It is evident that $\underline{E}(j0)-\underline{A}$ is invertible if and only if $\underline{E}(j0)-\underline{Q}$ is and that if such an inverse exists, the inverse operator is a mapping from l_2 to l_E. The following proposition gives the answer to this question.

PROPOSITION 2.4. *Assume that $A(t) \in L_{\mathrm{PCD}}[0,h]$ and that the system (1) is asymptotically stable in the Floquet theorem sense. Then $\underline{E}(j\varphi)-\underline{A}$ is invertible for all $\varphi \in [-\frac{\omega_h}{2}, \frac{\omega_h}{2}) =: \mathcal{I}_0$ and*

$$(\underline{E}(j\varphi)-\underline{A})^{-1} = \underline{P}(\underline{E}(j\varphi)-\underline{Q})^{-1}\underline{P}^{-1}, \tag{10}$$

where $\underline{E}(j\varphi) = \underline{E}(j0) + j\varphi\underline{I}$ and

$$\begin{aligned}&(\underline{E}(j\varphi)-\underline{Q})^{-1}\\ &= \mathrm{diag}[\ldots,(j\varphi_{-1}I-Q)^{-1},(j\varphi_0 I-Q)^{-1},(j\varphi_1 I-Q)^{-1},\ldots]\end{aligned} \tag{11}$$

with $\varphi_m := \varphi + m\omega_h, m \in \mathcal{Z}$. Moreover, $(\underline{E}(j\varphi)-\underline{A})^{-1}$ is compact and uniformly bounded on l_2 over $\varphi \in \mathcal{I}_0$.

Proof. By the assumption on $A(t)$, we have (4) so that for any $\varphi \in \mathcal{I}_0$

$$\underline{P}(\underline{E}(j\varphi)-\underline{Q})\underline{P}^{-1} = \underline{E}(j\varphi)-\underline{A}.$$

Also, by the stability assumption, all of the eigenvalues of $Q - j\varphi_m I$ for all $m \in \mathcal{Z}$ have negative real parts. Thus the operator on the right-hand side of (11), which is denoted by $\underline{D}(Q,\varphi)$, is well defined and bounded on l_2. To see this, we note that there exists $K > 0$ such that

$$||(j\varphi_m I - Q)^{-1}|| \le Kf(m) \quad (m \in \mathcal{Z}), \tag{12}$$

where f is defined in Appendix A. Noting that $\underline{D}(Q,\varphi)$ is block-diagonal, it follows that

$$||\underline{D}(Q,\varphi)||_{l_2/l_2} = \sup_{m\in\mathcal{Z}} ||(j\varphi_m I - Q)^{-1}|| \le K. \tag{13}$$

Some simple computations show that $\underline{D}(Q,\varphi)(\underline{E}(j\varphi)-\underline{Q}) = (\underline{E}(j\varphi)-\underline{Q})\underline{D}(Q,\varphi) = \underline{I}$. This, together with the fact that $\underline{P}$ and $\underline{P}^{-1}$ are invertible on l_2 and l_E, respectively, establishes (10). Noting that $(\underline{E}(j\varphi)-\underline{Q})^{-1}$ is uniformly bounded on l_2 by (13) and that $\underline{P}$ and $\underline{P}^{-1}$ are bound on l_2, then the uniform boundedness of $(\underline{E}(j\varphi)-\underline{A})^{-1}$ on l_2 over $\varphi \in \mathcal{I}_0$ follows from (10).

To see the compactness of $(\underline{E}(j\varphi)-\underline{Q})^{-1}$, we define

$$\begin{aligned}&[(\underline{E}(j\varphi)-\underline{Q})^{-1}]_N\\ &= \mathrm{diag}[\ldots,0,(j\varphi_{-N}I-Q)^{-1},\ldots,(j\varphi_0 I-Q)^{-1},\ldots,(j\varphi_N I-Q)^{-1},0,\ldots].\end{aligned}$$

It is clear that for any fixed N, the operator $[(\underline{E}(j\varphi)-\underline{Q})^{-1}]_N$ is bounded on l_2 by (12) and is a compact operator. Furthermore, it is easy to see from (12) that for any $\varphi \in \mathcal{I}_0$, $\lim_{N\to\infty}[(\underline{E}(j\varphi)-\underline{Q})^{-1}]_N = (\underline{E}(j\varphi)-\underline{Q})^{-1}$ in the l_2-induced norm sense, which implies that $(\underline{E}(j\varphi)-\underline{Q})^{-1}$ is a compact mapping on l_2. Noting that $\underline{P}$ and $\underline{P}^{-1}$ are bounded on l_2, it follows by (10) that $(\underline{E}(j\varphi)-\underline{A})^{-1}$ are also compact. ∎

The following result describes basic properties of the set l_E (see Appendix B for the proof).

LEMMA 2.4. *l_E is dense in l_2. Also, l_E is a proper and dense subset of l_1 in the l_2-norm sense.*

Lemma 2.4 reveals that (4) can be seen as an operator-valued relation densely defined on l_2 (i.e., from the dense subset $l_E \subset l_2$ to l_2).

Finally, for our later use, we denote the Fourier series expansion operator from $L_2[0,h]$ to l_2 by $\mathcal{F}$. Then, it is easy to see that if $\mathcal{F}\{x(\cdot)\} \in l_1$ for $x(t) \in L_2[0,h]$, then the Fourier series expansion of $x(t)$ is absolutely convergent.

3. Frequency response operators viewed on l_2. In this section, we construct the frequency response relation of the FDLCP system (1) by the steady-state input-output analysis. This is first proposed in [26], [27]. The basic idea can be described as follows. First, impose an l_2-EMP signal u (where EMP stands for exponentially modulated periodic) to the system of (1), that is,

$$u(t) = \sum_{m=-\infty}^{+\infty} u_m e^{j(\varphi+m\omega_h)t} = \sum_{m=-\infty}^{+\infty} u_m e^{j\varphi_m t} \quad (t \geq 0, \varphi \in \mathcal{I}_0),$$

where the infinite-dimensional vector $\underline{u} := [\ldots, u_{-1}^T, u_0^T, u_1^T, \ldots]^T$ belongs to l_2. Second, measure the steady-state output y of the system, which is (assumed to be) also an l_2-EMP signal under the asymptotic stability assumption of the system and represent the signal y by the infinite-dimensional vector $\underline{y} := [\ldots, y_{-1}^T, y_0^T, y_1^T, \ldots]^T \in l_2$ according to the definition of l_2-EMP signals. Finally, the input-output response relation observed in the above is expressed as a mapping $\underline{G}(j\varphi) : \underline{u} \mapsto \underline{y}$.

In the above arguments, the Fourier series expansions of $A(t), B(t), C(t)$, and $D(t)$ as well as the Toeplitz operators expressions of these h-periodic matrix functions are used repeatedly. It should be pointed out that the validity of such use has not been verified rigorously in [26], [27]. In the following, we will reconsider the above-mentioned arguments and concentrate our attention on the convergence problems of the Fourier series expansions and the Toeplitz transformations involved. To this purpose, we note from the Floquet theorem that

$$\begin{aligned} y(t) &= C(t)P(t,0)e^{Q(t-t_0)}P^{-1}(t_0,0)x_0 \\ &\quad + C(t)P(t,0)\int_{t_0}^{t} e^{Q(t-\tau)}P^{-1}(\tau,0)B(\tau)u(\tau)d\tau + D(t)u(t) \\ &= \hat{C}(t)\left[e^{Q(t-t_0)}q_0 + \int_{t_0}^{t} e^{Q(t-\tau)}\hat{B}(\tau)u(\tau)d\tau\right] + D(t)u(t) \end{aligned} \tag{14}$$

with $\hat{B}(t) := P^{-1}(t,0)B(t), \hat{C}(t) := P(t,0)C(t)$, and $q_0 := P^{-1}(t_0,0)x_0$. The second relation of (14) says that if we introduce the initial value transformation $q_0 = P^{-1}(t_0,0)x_0$, the system (1) can be represented equivalently in the input-output sense by the system configuration shown in Figure 1.

Now we are in a position to establish the frequency response relation in the system of Figure 1 by imposing an l_2-EMP sinusoid input u to the system and measuring the steady-state output y. From Figure 1, this can be completed by showing that under certain assumptions given below, the steady-state responses at the points p, q, and y are also l_2-EMP signals so that the input-output response relation $u \mapsto y$ can be written as a mapping of $\underline{u} \mapsto \underline{y}$. We complete this in three steps.

Step 1. Take an h-periodic continuous signal $\hat{u}(t) \in L_2[0,h]$ such that $\mathcal{F}\{\hat{u}(\cdot)\} =: \underline{\hat{u}} \in l_1 \subset l_2$. Then the Fourier series expansion of $\hat{u}(t)$ is absolutely convergent.

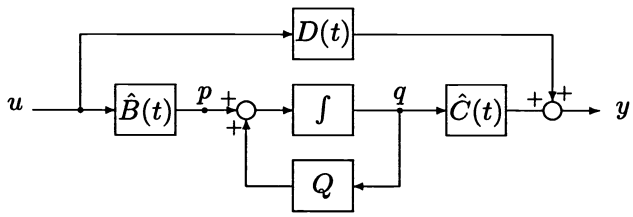


FIG. 1. *Equivalent system configuration.*

Constructing the input l_2-EMP signal as $u(t) = \hat{u}(t)e^{j\varphi t}, \varphi \in \mathcal{I}_0$ (where $\mathcal{I}_0$ is defined in Proposition 2.4), it follows that the corresponding output of $\hat{B}(t)$ to this l_2-EMP signal is

$$p(t) = \hat{p}(t)e^{j\varphi t} \quad (t \geq 0, \varphi \in \mathcal{I}_0),$$

where $\hat{p}(t) = \hat{B}(t)\hat{u}(t) = P^{-1}(t,0)B(t)\hat{u}(t)$.

Now assume that $B(t) \in L_{\mathrm{PCC}}[0,h]$. Then, from the choice of $\hat{u}$ and Lemma 2.1 (the assertion is expressed in terms of the operator $\mathcal{F}$ by taking the central column), we obtain

$$\mathcal{F}\{B(\cdot)\hat{u}(\cdot)\} = \underline{B}\,\mathcal{F}\{\hat{u}(\cdot)\} = \underline{B}\,\underline{\hat{u}},$$

and the Fourier series expansion of $B(t)\hat{u}(t)$ is convergent to $B(t_0)\hat{u}(t_0)$ for a.e. $t_0 \in [0,h]$.

Furthermore, let us assume that $A(t) \in L_{\mathrm{PCD}}[0,h]$. Then, from Proposition 2.2, $P^{-1}(t,0)$ is continuous, and the Fourier series expansion of $P^{-1}(t,0)$ is absolutely convergent. Again by Lemma 2.1 and Proposition 2.2, we obtain

$$\mathcal{F}\{P^{-1}(\cdot,0)B(\cdot)\hat{u}(\cdot)\} = \mathcal{T}\{P^{-1}(\cdot,0)\}\mathcal{F}\{B(\cdot)\hat{u}(\cdot)\} = \underline{P}^{-1}\underline{B}\,\underline{\hat{u}},$$

which can be interpreted as

$$\mathcal{F}\{\hat{p}(\cdot)\} =: \underline{\hat{p}} = \underline{P}^{-1}\underline{B}\,\underline{\hat{u}} \in l_2. \tag{15}$$

The assertion that $\underline{\hat{p}} \in l_2$ follows from the facts that $\underline{P}^{-1}$ and $\underline{B}$ are bounded on l_2 under the given assumptions on $A(t)$ and $B(t)$. From the above arguments, it follows that $\hat{p}(t)e^{j\varphi t} = p(t)$ is also l_2-EMP. In other words, one can conclude that $p(t) = \sum_{m=-\infty}^{+\infty} p_m e^{j(\varphi + m\omega_h)t}$ with $p_m := [\underline{\hat{p}}]_m$.

Remark 3.1. The reason we constrain the domain of $\hat{u}$ is that if we work on a general $\hat{u}(t) \in L_2[0,h]$, we may not arrive at the above conclusions for some $\underline{\hat{u}} \in l_2$ because of the convergence problems in the Fourier series expansions and the Toeplitz transformations.

Step 2. Now impose the signal p to the linear time-invariant (LTI) subsystem of Figure 1. We suppose that this subsystem is asymptotically stable (i.e., all the eigenvalues of Q have negative real parts). Then, by the superposition principle of linear systems [21, Theorem 5.6.2, p. 237], the output q of the LTI subsystem to p is

$$
\begin{aligned}
q(t) &= e^{Qt}q_0 + \int_0^t e^{Q(t-\tau)} \sum_{m=-\infty}^{+\infty} p_m e^{j(\varphi+m\omega_h)\tau} d\tau \\
&= e^{Qt}\left(q_0 + \sum_{m=-\infty}^{+\infty} \int_0^t e^{(j\varphi_m I - Q)\tau} d\tau p_m\right) \\
&= e^{Qt}\left(q_0 + \sum_{m=-\infty}^{+\infty} (Q - j\varphi_m I)^{-1} p_m\right) + \sum_{m=-\infty}^{+\infty} (j\varphi_m I - Q)^{-1} p_m e^{j\varphi_m t}. \qquad (16)
\end{aligned}
$$

On the other hand, by the stability assumption of Q, (12) is true for all $\varphi \in \mathcal{I}_0$. Therefore, we observe by the Cauchy–Schwarz inequality and Appendix A that

$$
\begin{aligned}
\sum_{m=-\infty}^{+\infty} ||(Q - j\varphi_m I)^{-1} p_m|| &\le \sum_{m=-\infty}^{+\infty} ||(Q - j\varphi_m I)^{-1}|| \cdot ||p_m|| \\
&\le \left(\sum_{m=-\infty}^{+\infty} ||(Q - j\varphi_m I)^{-1}||^2\right)^{\frac{1}{2}} \left(\sum_{m=-\infty}^{+\infty} ||p_m||^2\right)^{\frac{1}{2}} \\
&\le K\left(\sum_{m=-\infty}^{+\infty} f(m)^2\right)^{\frac{1}{2}} ||\underline{p}||_{l_2} \le \sqrt{5}K\,||\underline{p}||_{l_2}, \qquad (17)
\end{aligned}
$$

where $\underline{p} := [\ldots, p_{-1}^T, p_0^T, p_1^T, \ldots] = \underline{\hat{p}} \in l_2$. The above inequality implies that the summation $\sum_{m=-\infty}^{+\infty}(Q - j\varphi_m I)^{-1} p_m$ is absolutely convergent for any $\varphi \in \mathcal{I}_0$. Combining this fact with (16), it follows that as $t \to \infty$, the steady-state response is

$$\left(\sum_{m=-\infty}^{+\infty} (j\varphi_m I - Q)^{-1} p_m e^{jm\omega_h t}\right) e^{j\varphi t}$$

since $e^{Qt} \to 0$. This steady-state output q of the LTI subsystem can be expressed as

$$q(t) := \hat{q}(t) e^{j\varphi t} \quad (t \ge 0),$$

where $\hat{q}(t) = \sum_{m=-\infty}^{+\infty} \hat{q}_m e^{jm\omega_h t}$ with $\hat{q}_m := (j\varphi_m I - Q)^{-1} p_m$. The arguments in (17) also indicate that $\mathcal{F}\{\hat{q}(\cdot)\} =: \underline{\hat{q}} \in l_1 \subset l_2$. More precisely, since $(\underline{E}(j\varphi) - \underline{Q})^{-1}$ is a mapping from l_2 to $l_E \subset l_1$ by Proposition 2.4 and Lemma 2.4, it also follows that $\underline{\hat{q}} \in l_E \subset l_1$. Consequently, $\hat{q}(t) \in L_2[0, h]$ and the Fourier series expansion of $\hat{q}(t)$ is absolutely convergent. Obviously, $q(t)$ is an l_2-EMP signal.

Step 3. Since the Fourier series expansion of $P(t, 0)$ is absolutely convergent, the assertion in the last paragraph of Step 2 actually says that the Fourier series expansion of $P(t, 0)\hat{q}(t)$ is also absolutely convergent by the fact [19] that the Fourier series expansion of the product of two matrix functions is absolutely convergent if the Fourier series expansions of these two matrix functions are absolutely convergent. Now, repeating the arguments in Step 1 on the matrix function $C(t)P(t, 0)\hat{q}(t)$, the relation

$$\mathcal{F}\{\hat{y}(\cdot)\} =: \underline{\hat{y}} = \underline{C}\,\underline{P}\,\underline{\hat{q}} \qquad (18)$$

can be asserted if $C(t) \in L_{\mathrm{PCC}}[0, h]$, where $\hat{y}$ is the output of $C(t)P(t, 0)$ to the input $\hat{q}(t)$. It is clear that $\underline{\hat{y}} \in l_2$, and thus the output y of $C(t)P(t, 0)$ to the input $q(t) = \hat{q}(t)e^{j\varphi t}$ is l_2-EMP.

Finally, from (15) and (18) and by Proposition 2.4, we obtain

$$\underline{\hat{y}} = \underline{C}\,\underline{P}(\underline{E}(j\varphi) - \underline{Q})^{-1}\underline{P}^{-1}\underline{B}\,\underline{u} = \underline{C}(\underline{E}(j\varphi) - \underline{A})^{-1}\underline{B}\,\underline{u}$$

by setting $\underline{u} := \underline{\hat{u}}$. Summarizing the above discussions and taking $D(t)$ into consideration, we can state the following main result about the existence conditions of the frequency response operator.

THEOREM 3.1. *Assume that in the system* (1) $A(t)$ *belongs to* $L_{\mathrm{PCD}}[0,h]$, $B(t)$, $C(t)$, *and* $D(t)$ *belong to* $L_{\mathrm{PCC}}[0,h]$, *and that the system* (1) *is asymptotically stable in the Floquet theorem sense. Then the steady-state response of the system* (1) *to the* l_2-EMP *input* $u(t) = \sum_{m=-\infty}^{+\infty} u_m e^{j\varphi_m t}$ *with* $\underline{u} = [\ldots, u_{-1}^T, u_0^T, u_1^T, \ldots]^T \in l_1 \subset l_2$ *is also an* l_2-EMP *signal* $y(t) = \sum_{m=-\infty}^{+\infty} y_m e^{j\varphi_m t}$ *with* $\underline{y} = [\ldots, y_{-1}^T, y_0^T, y_1^T, \ldots]^T = \underline{G}(j\varphi)\underline{u} \in l_2$, *where*

$$\underline{G}(j\varphi) := \underline{C}(\underline{E}(j\varphi) - \underline{A})^{-1}\underline{B} + \underline{D}. \tag{19}$$

Hence the frequency response operator $\underline{G}(j\varphi)$ *is a densely defined mapping on* l_2 *for each* $\varphi \in \mathcal{I}_0$. *Furthermore, it is uniformly bounded over* $\varphi \in \mathcal{I}_0$ *in the sense that* $||\underline{G}(j\varphi)||_{l_2/l_1(l_2)} \le K < \infty$ *for all* $\varphi \in \mathcal{I}_0$ *for some* $K > 0$, *where*

$$||\underline{G}(j\varphi)||_{l_2/l_1(l_2)} := \sup_{0 \neq \underline{x} \in l_1} \left\{ \frac{||\underline{G}(j\varphi)\underline{x}||_{l_2}}{||\underline{x}||_{l_2}} \right\}.$$

Proof. The first assertion follows directly from the preceding arguments. It is clear from the above arguments that $\underline{G}(j\varphi)$ is a mapping from l_1 into l_2. However, l_1 is dense in l_2 so that the frequency response operator established via the steady-state input-output analysis is a densely defined operator on l_2 [21, p. 486]. To see the uniform boundedness of $\underline{G}(j\varphi)$ over the interval $\mathcal{I}_0$, we note that $\underline{B}, \underline{C}$, and $\underline{D}$ are bounded on l_2 by the assumptions on $B(t), C(t)$, and $D(t)$ from Lemma 2.3. Then the uniform boundedness assertion of $\underline{G}(j\varphi)$ follows from Proposition 2.4. ∎

Remark 3.2. Note that we have used the l_2 norm on l_1 in Theorem 3.1. Accordingly, $||\underline{G}(j\varphi)||_{l_2/l_1(l_2)}$ is the l_2-induced norm of $\underline{G}(j\varphi)$ on the subset l_1 of l_2.

By the mathematical expression of the frequency response operator, this operator can have two interpretations. The first one is to view it as a mapping from l_1 into l_2, which has a clear steady-state analysis interpretation as we discussed above; the second is to treat it as a mapping on l_2. The second viewpoint makes sense because it can be seen as a mapping with the extended domain l_2 instead of the original domain l_1, and this mapping itself is bounded on l_2 (since all the operators in $\underline{G}(j\varphi)$ are bounded on l_2, and this fact is used in the uniform boundedness proof of $\underline{G}(j\varphi)$ in Theorem 3.1). Here, to distinguish these two operators, the frequency response operator in the first interpretation is given a new notation $\underline{\tilde{G}}(j\varphi)$, while the original notation $\underline{G}(j\varphi)$ is taken over by the second interpretation. Note that $\underline{\tilde{G}}(j\varphi)$ and $\underline{G}(j\varphi)$ have the same mathematical expression but are defined on different domains.

Compared with $\underline{G}(j\varphi)$, the frequency response operator $\underline{\tilde{G}}(j\varphi)$ defined via the steady-state analysis is "deficient" in the sense that the domain of $\underline{\tilde{G}}(j\varphi)$ is a dense subset of l_2. However, it is straightforward to establish the following corollary, which shows that the l_2-induced norm of $\underline{\tilde{G}}(j\varphi)$ from l_1 to l_2 coincides with the l_2-induced norm of $\underline{G}(j\varphi)$ on l_2. This validates the existing studies on the definition and computation of the H_∞ norm of the FDLCP system (1) based on the frequency response operator $\underline{G}(j\varphi)$.

COROLLARY 3.1. *Suppose that in the system* (1) $A(t)$ *belongs to* $L_{\mathrm{PCD}}[0,h]$, $B(t), C(t)$, *and* $D(t)$ *belong to* $L_{\mathrm{PCC}}[0,h]$, *and that the system* (1) *is asymptotically stable. Then for all* $\varphi \in \mathcal{I}_0$

$$||\tilde{\underline{G}}(j\varphi)||_{l_2/l_1(l_2)} = ||\underline{G}(j\varphi)||_{l_2/l_2}.$$

Hence $\underline{G}(j\varphi)$ *is bounded on* l_2 *uniformly over* $\varphi \in \mathcal{I}_0$.

As final words of this section, we make a few remarks about the H_2 norm of FDLCP systems. In [27], the H_2 norm of FDLCP systems has been defined by

$$||\mathcal{G}||_2 := \left\{ \frac{1}{2\pi} \int_{-\frac{\omega_h}{2}}^{\frac{\omega_h}{2}} \mathrm{trace}(\underline{G}(j\varphi)^* \underline{G}(j\varphi)) d\varphi \right\}^{\frac{1}{2}}. \tag{20}$$

In (20), we have implicitly assumed that the system (1) is strictly proper. Thus $\underline{G}(j\varphi)$ is compact by Proposition 2.4 and the fact that $\underline{B}$ and $\underline{C}$ are bounded on l_2. Even though, rigorously speaking, the steady-state input-output analysis of [27] leads only to $\tilde{\underline{G}}(j\varphi)$ rather than $\underline{G}(j\varphi)$ as mentioned above, it is indeed reasonable to adopt the above definition in terms of $\underline{G}(j\varphi)$ rather than $\tilde{\underline{G}}(j\varphi)$. This is because, as is well known (see, e.g., [12, p. 105], [21, p. 389], [31, p. 347]), trace is defined only for trace class operators, which are a class of (compact) operators defined on Hilbert spaces. Since the domain of $\tilde{\underline{G}}(j\varphi)$ is only a subset of the Hilbert space l_2, it is not adequate to define trace on $\tilde{\underline{G}}(j\varphi)^* \tilde{\underline{G}}(j\varphi)$. On the other hand, it is not hard to show that $\underline{G}(j\varphi)^* \underline{G}(j\varphi)$ is indeed a trace class operator on l_2, which follows from (12). Some further discussions are given in section 5.

4. Frequency response operators viewed on l_1. In this section, we show that under certain conditions, the frequency response operator can also be established via the steady-state analysis of l_1-EMP input signals as a bounded mapping on l_1 (i.e., from l_1 into l_1). Now let us define the set $l_e = \{\underline{x} \in l_1 : \underline{E}(j0)\underline{x} \in l_1\}$ and state the following lemma, which can be shown in a similar way to Lemma 2.4.

LEMMA 4.1. l_e *is a proper dense subset of* l_1 *and* $l_e \subset l_E$.

Lemma 4.1 says that the role of the subset l_e of l_1 is similar to that of the subset l_E of l_2 so that the inverse of $\underline{E}(j0) - \underline{A}$ can be derived as a mapping from l_1 to l_e from (9). For brevity, the details are omitted, but the assertions are stated in the following proposition, which is helpful in establishing the frequency response relation of the FDLCP system in terms of a mapping on l_1.

PROPOSITION 4.1. *Suppose that in* (1) $A(t) \in L_{\mathrm{CPCD}}[0,h], B(t), C(t) \in L_{\mathrm{CAC}}[0,h]$. *Then* $\underline{P}$ *and* $\underline{P}^{-1}$ *are bounded on* l_1. *In particular,* l_e *is* $\underline{P}$- *and* $\underline{P}^{-1}$-*invariant, and hence* $\underline{P}$ *is invertible on* l_e. *The unique inverse of* $\underline{P}$ *on* l_e *is* $\underline{P}^{-1}$ *restricted to* l_e. *It is also true that on* $l_e \subset l_1$

$$\underline{P}(\underline{E}(j0) - \underline{Q})\underline{P}^{-1} = \underline{E}(j0) - \underline{A}. \tag{21}$$

Also, it holds on the whole l_1 *that* $\hat{\underline{B}} = \underline{P}^{-1}\underline{B}$ *and* $\hat{\underline{C}} = \underline{C}\,\underline{P}$. *Furthermore, if the system* (1) *is asymptotically stable in the Floquet theorem sense, then* $\underline{E}(j\varphi) - \underline{A}$ *is invertible for all* $\varphi \in \mathcal{I}_0$, *and*

$$\underline{P}(\underline{E}(j\varphi) - \underline{Q})^{-1}\underline{P}^{-1} = (\underline{E}(j\varphi) - \underline{A})^{-1}, \tag{22}$$

which is a mapping from l_1 *to* l_e. *Also,* $(\underline{E}(j\varphi) - \underline{A})^{-1}$ *is compact and uniformly bounded on* l_1 *over* $\varphi \in \mathcal{I}_0$.

Proof. The proof can be given by some similar steps to those in Propositions 2.3 and 2.4. Here it remains only to show that the operators $\mathcal{T}\{\dot{P}(t,0)\}$, $\underline{P}, \underline{P}^{-1}, \underline{B}$, and $\underline{C}$ are bounded on l_1 and that $(\underline{E}(j\varphi) - \underline{Q})^{-1}$ is uniformly bounded on l_1 over $\varphi \in \mathcal{I}_0$. By (3) and the assumption on $A(t)$, $\dot{P}(t,0)$ is continuous and the first-order derivative of $\dot{P}(t,0)$ is piecewise continuous in $[0,h]$. Hence, by Theorem 2 of [6, p. 104], the Fourier series expansion of $\dot{P}(t,0)$ is absolutely convergent. Now we denote the Fourier coefficients sequence of $\dot{P}(t,0)$ by $\{\hat{P}_m\}_{m=-\infty}^{+\infty}$. Obviously, if $\underline{x} \in l_1$, then

$$\begin{aligned} ||\mathcal{T}\{\dot{P}(t,0)\}\underline{x}||_{l_1} &= \sum_{m=-\infty}^{+\infty} \left|\left| \sum_{k=-\infty}^{+\infty} \hat{P}_{m-k} x_k \right|\right| \\ &\le \sum_{m=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} ||\hat{P}_{m-k}|| \cdot ||x_k|| = \left(\sum_{m=-\infty}^{+\infty} ||\hat{P}_m|| \right) ||\underline{x}||_{l_1}, \end{aligned}$$

where $\sum_{m=-\infty}^{+\infty} ||\hat{P}_m|| < \infty$ by the absolute convergence mentioned above. From this, it follows readily that the operator $\mathcal{T}\{\dot{P}(t,0)\}$ is bounded on l_1. Similarly, since the Fourier series expansions of $P(t,0)$ and $P^{-1}(t,0)$ are absolutely convergent, $\underline{P}$ and $\underline{P}^{-1}$ are bounded on l_1. The boundedness of $\underline{B}$ and $\underline{C}$ on l_1 follows directly from the assumption that $B(t)$ and $C(t)$ belong to $L_{\mathrm{CAC}}[0,h]$. The last assertion follows from the above discussions, (12), and (22). ∎

By Proposition 4.1, one can establish the frequency response operator on l_1 by the steady-state analysis but with the l_1-EMP signals. That is, the l_1-EMP input is $u(t) = \sum_{m=-\infty}^{+\infty} u_m e^{j\varphi_m t}, \varphi \in \mathcal{I}_0$, with $\underline{u} := [\ldots, u_{-1}^T, u_0^T, u_1^T, \ldots]^T \in l_1$. The following theorem summarizes such discussions.

THEOREM 4.1. *Assume that in the system* (1) $A(t) \in L_{\mathrm{CPCD}}[0,h]$, $B(t), C(t)$, *and* $D(t)$ *belong to* $L_{\mathrm{CAC}}[0,h]$, *and that the system* (1) *is asymptotically stable. Then the steady-state response of* (1) *to the* l_1-EMP *input* $u(t) = \sum_{m=-\infty}^{+\infty} u_m e^{j\varphi_m t}$ *with* $\underline{u} = [\ldots, u_{-1}^T, u_0^T, u_1^T, \ldots]^T \in l_1$ *is also an* l_1-EMP *signal* $y(t) = \sum_{m=-\infty}^{+\infty} y_m e^{j\varphi_m t}$ *with* $\underline{y} = [\ldots, y_{-1}^T, y_0^T, y_1^T, \ldots]^T = \underline{G}(j\varphi)\underline{u} \in l_1$, *where* $\underline{G}(j\varphi)$ *is given in* (19). *Hence the frequency response operator* $\underline{G}(j\varphi)$ *is well defined on* l_1 *for each* $\varphi \in \mathcal{I}_0$. *Also, it is uniformly bounded over* $\varphi \in \mathcal{I}_0$ *in the sense that* $||\underline{G}(j\varphi)||_{l_1/l_1} \le K < \infty$ *for all* $\varphi \in \mathcal{I}_0$ *for some* $K > 0$.

Remark 4.1. We mention that Theorem 4.1 is not a special case of Theorem 3.1. To see this, the following facts are mentioned. (i) In Theorem 3.1, the EMP signal $u(t)$ is viewed as an l_2-EMP signal even though $u(t)$ itself is l_1-EMP; (ii) Theorems 3.1 and 4.1 are proved by using Propositions 2.4 and 4.1, respectively, which hold on different linear spaces; (iii) The uniform boundedness of the frequency response operator $\underline{G}(j\varphi)$ in Theorem 3.1 is stated in the l_2-induced norm sense from $l_1 \subset l_2$ to l_2, while that of Theorem 4.1 is in the l_1-induced norm sense.

5. Relations to the time-domain definitions of H_2 and H_∞ norms. In this section, we show that the H_2 and H_∞ norms of FDLCP systems defined via the frequency response operator are equal to the time-domain counterparts under some conditions. To this end, let us first define the formal frequency response operator of the system in Figure 1 by

$$\hat{\underline{G}}(j\varphi) = \hat{\underline{C}}(\underline{E}(j\varphi) - \underline{Q})^{-1}\hat{\underline{B}} + \underline{D},$$

where $\hat{\underline{B}} = \mathcal{T}\{\hat{B}(t)\}$ and $\hat{\underline{C}} = \mathcal{T}\{\hat{C}(t)\}$. For $\hat{\underline{G}}(j\varphi)$ to make sense, we assume that $A(t) \in L_2[0,h]$ and the system is asymptotically stable so that $\underline{E}(j\varphi) - \underline{Q}$ is invertible

for all $\varphi \in \mathcal{I}_0$. We also assume that $\underline{\hat{B}}, \underline{\hat{C}}$, and $\underline{D}$ are bounded on l_2. The purpose of introducing this frequency response operator is that the time-domain H_2 norm and the L_2-induced norm are more naturally connected with $\underline{\hat{G}}(j\varphi)$ rather than $\underline{G}(j\varphi)$, although, under stronger assumptions, the matrix representations of these two operators eventually coincide with each other. This coincidence will help to recover the equivalences between these two norms and the counterparts in terms of the frequency response operator.

First, we consider the H_2 norm of FDLCP systems. In what follows, we assume that the FDLCP system (1) is strictly proper when the H_2 norm problem is considered. Denoting the impulse response of the system (1) by $g(\cdot,\cdot)$, the time-domain H_2 norm [5], [27], [32] of the FDLCP system (1) is

$$||\mathcal{G}||_2 = \left\{ \frac{1}{h} \int_0^h \int_{-\infty}^{+\infty} \mathrm{trace}(g(t,\tau)^* g(t,\tau)) dt d\tau \right\}^{\frac{1}{2}}.$$

PROPOSITION 5.1. *Suppose that in the system* (1) $A(t) \in L_2[0,h]$*, and the system is asymptotically stable. Also assume that* $\hat{B}(t)$ *and* $\hat{C}(t)$ *belong to* $L_{\mathrm{CAC}}[0,h]$*. Then it holds that*

$$||\mathcal{G}||_2 = \left\{ \frac{1}{2\pi} \int_{-\frac{\omega_h}{2}}^{\frac{\omega_h}{2}} \mathrm{trace}(\underline{\hat{G}}(j\varphi)^* \underline{\hat{G}}(j\varphi)) d\varphi \right\}^{\frac{1}{2}}.$$

Proof. See Appendix B. □

We are in a position to show the following result, which says that the time-domain definition of the H_2 norm is equivalent to the frequency-domain definition given in [27]. Even though this fact seems to be regarded as well known, we stress here that no direct and explicit proof via the frequency response operator is available in the literature to the best of the knowledge of the authors.

THEOREM 5.1. *Suppose that in the system* (1) $A(t)$ *belongs to* $L_{\mathrm{PCD}}[0,h]$*,* $B(t)$ *and* $C(t)$ *belong to* $L_{\mathrm{CAC}}[0,h]$*, and that the system is asymptotically stable. Then*

$$||\mathcal{G}||_2 = \left\{ \frac{1}{2\pi} \int_{-\frac{\omega_h}{2}}^{\frac{\omega_h}{2}} \mathrm{trace}(\underline{G}(j\varphi)^* \underline{G}(j\varphi)) d\varphi \right\}^{\frac{1}{2}}.$$

Proof. Under the given conditions, it is clear from Propositions 2.3 and 2.4 that $\underline{\hat{G}}(j\varphi) = \underline{G}(j\varphi)$. Hence, by the result in Proposition 5.1, it remains to show that the Fourier series expansions of $\hat{B}(t) = P^{-1}(t,0)B(t)$ and $\hat{C}(t) = C(t)P(t,0)$ are absolutely convergent. To see this, it is enough to note that the Fourier series expansions of $P^{-1}(t,0)$ and $P(t,0)$ are absolutely convergent from Proposition 2.2 by the assumption on $A(t)$. □

Now we discuss the relation between the L_2-induced norm and the H_∞ norm of FDLCP systems. The L_2-induced norm of the FDLCP system (1) is

$$||\mathcal{G}||_{L_2/L_2} = \sup_{0 \neq u \in L_2} \frac{||y(\cdot)||_{L_2}}{||u(\cdot)||_{L_2}}.$$

To establish the relation between the L_2-induced norm and the H_∞ norm of FDLCP systems, we first prove the following proposition. To this purpose, we introduce the so-called SD-Fourier transform [1]. For a signal $x \in L_2$, its SD-Fourier

transform is defined as

$$\underline{X}_{\mathrm{SD}}(j\varphi) := [\ldots, X(j\varphi_{-1})^T, X(j\varphi_0)^T, X(j\varphi_1)^T, \ldots]^T, \tag{23}$$

where $X(j\omega)$ is the Fourier transform of x, and $X(j\varphi_n) = X(j(\varphi + n\omega_h)), n \in \mathcal{Z}, \varphi \in \mathcal{I}_0$. It can also be said that $\underline{X}_{\mathrm{SD}}(j\varphi)$ is the lifted version of $X(j\omega)$ in the frequency domain. This kind of frequency-domain lifting technique has been frequently used in sampled-data system sensitivity analysis [4] and signal processing [24].

PROPOSITION 5.2. *Suppose that in the system* (1) $A(t)$ *belongs to* $L_2[0,h]$, $\hat{B}(t), \hat{C}(t)$, *and* $D(t)$ *belong to* $L_{\mathrm{CAC}}[0,h]$, *and that the system is asymptotically stable in the Floquet theorem sense. Then the system* (1) *is* L_2*-stable and*

(1) $\underline{Y}_{\mathrm{SD}}(j\varphi) = \underline{\hat{G}}(j\varphi)\underline{U}_{\mathrm{SD}}(j\varphi)$ *for all* $\varphi \in \mathcal{I}_0$ *for any* $u(t) \in C_0^1$, *where* C_0^1 *denotes the space of continuously differentiable functions with compact support;*

(2) $||y(\cdot)||_{L_2}^2 = \frac{1}{2\pi}\int_{\mathcal{I}_0} \underline{U}_{\mathrm{SD}}^*(j\varphi)\underline{\hat{G}}^*(j\varphi)\underline{\hat{G}}(j\varphi)\underline{U}_{\mathrm{SD}}(j\varphi)d\varphi$ *for any* $u(t) \in C_0^1$;

where $\underline{U}_{\mathrm{SD}}(j\varphi)$ *and* $\underline{Y}_{\mathrm{SD}}(j\varphi)$ *are the* SD*-Fourier transforms of* $u(t)$ *and* $y(t)$, *respectively.*

Proof. See Appendix B. ∎

Based on Proposition 5.2, we establish the equivalence between the L_2-induced norm of the system (1) and the maximum of the l_2-induced norm of $\underline{G}(j\varphi)$ over $\varphi \in \mathcal{I}_0$, which is called the H_∞ norm of the frequency response operator $\underline{G}(j\varphi)$ of the FDLCP system (1) [17], [27]. Again, no direct and explicit proof for this equivalence is available in the literature.

THEOREM 5.2. *Suppose that in the system* (1) $A(t)$ *belongs to* $L_{\mathrm{PCD}}[0,h]$, $B(t)$, $C(t)$, *and* $D(t)$ *belong to* $L_{\mathrm{CAC}}[0,h]$, *and that the system is asymptotically stable. Then*

$$||\mathcal{G}||_{L_2/L_2} = \max_{\varphi\in\mathcal{I}_0} ||\underline{G}(j\varphi)||_{l_2/l_2}.$$

Proof. By the assumptions on $A(t), B(t)$, and $C(t)$, together with Proposition 2.2, it follows that the Fourier series expansions of $\hat{B}(t)$ and $\hat{C}(t)$ are also absolutely convergent. This implies that Proposition 5.2 applies to the system of Figure 1. In view of this, we show that

$$||\mathcal{G}||_{L_2/C_0^1(L_2)} := \sup_{u\in C_0^1} \frac{||y(\cdot)||_{L_2}}{||u(\cdot)||_{L_2}} = \max_{\varphi\in\mathcal{I}_0} ||\underline{\hat{G}}(j\varphi)||_{l_2/l_2}. \tag{24}$$

This can be accomplished by some similar arguments to those in the proof of Theorem 5 of [1]. On the other hand, since C_0^1 is dense in L_2, it follows that $||\mathcal{G}||_{L_2/C_0^1(L_2)} = ||\mathcal{G}||_{L_2/L_2}$. Furthermore, under the given assumptions, it is obvious that $\underline{\hat{G}}(j\varphi) = \underline{G}(j\varphi)$ by Propositions 2.3 and 2.4. This, together with (24), completes the proof. ∎

6. Conclusion. The frequency response relation of FDLCP systems via the steady-state input-output analysis is reconsidered in this paper. We found that because of the various convergence problems related to the Fourier analysis and the Toeplitz transformation, the frequency response operator defined in [27] via such an analysis is actually guaranteed to be defined only densely on l_2 but that the frequency response operator can be extended to have the whole l_2 as its domain so that we still can define and compute the H_2 and H_∞ norms of FDLCP systems based on the frequency response operator [33]. It is also proved that, under some strengthened conditions, the time-domain H_2 norm (respectively, the L_2-induced norm) is equal to

the frequency-domain H_2 norm (respectively, the H_∞ norm of the frequency response operator). Thus the well-known equivalences about the H_2 and H_∞ norms in LTI continuous-time systems are recovered in a class of FDLCP systems. As another contribution of this study, it is verified that the frequency response operator defined by the steady-state input-output analysis can also be established as a mapping on l_1 under some strengthened assumptions on the system matrices $\{A(t), B(t), C(t), D(t)\}$. How to exploit this frequency response operator on l_1 remains one of our future research topics. The implication of this work is that the frequency response operator defined via the steady-state input-output analysis is well defined in most practical FDLCP systems, and the mathematical expression of these operators are similar to the well-known results in LTI systems. In addition, it is worth mentioning that this study reveals that the frequency response operators via the steady-state analysis may contain much more system structural information of FDLCP systems than we have understood in the usual ways before. For example, we believe that through the steady-state input-output analysis to l_p-EMP signals, $2 < p < \infty$, the frequency response operators of FDLCP systems can be introduced as a mapping (densely defined) on l_p under possibly weaker assumptions than those in the l_2 case. This is left for our further study.

Appendix A.

LEMMA A.1 (see [1]). *If the function $f(n)$ of an integer n is defined by*

$$f(n) = \begin{cases} 1, & n = 0, \\ |n|^{-1}, & n \neq 0, \end{cases}$$

then we have $\sum_{n=N+1}^{\infty} f(n)^2 < \frac{1}{N}$ $(N \geq 1)$ and $\sum_{n=-\infty}^{\infty} f(n)^2 < 5$.

Appendix B.

Proof of Lemma 2.1. Since $Y(t) \in L_{\mathrm{CAC}}[0, h]$, it follows that its Fourier series expansion is uniformly convergent with respect to t over $[0, h]$. Thus the Fourier series expansion of $Y(t)$ defines a continuous function over $[0, h]$, which is nothing but $Y(t)$ by the continuity of $Y(t)$. In other words, for every $t_0 \in [0, h]$, the Fourier series expansion of $Y(t)$ converges to $Y(t_0)$. Hence, for each $t_0 \in [0, h]$ at which the Fourier series expansion of $X(t)$ converges to $X(t_0)$, we have

$$\begin{aligned} X(t_0)Y(t_0) &= \left(\sum_{m=-\infty}^{+\infty} X_m e^{jm\omega_h t_0} \right) \left(\sum_{m=-\infty}^{+\infty} Y_m e^{jm\omega_h t_0} \right) \\ &= \sum_{m=-\infty}^{+\infty} \left(\sum_{k=-\infty}^{+\infty} X_{m-k} Y_k \right) e^{jm\omega_h t_0}, \end{aligned} \tag{25}$$

where we used the Mertens theorem to compute the product of two infinite sums, which also ensures that the right-hand side of (25) is convergent at t_0. Noting that it is in the form of the Fourier series expansion, it immediately follows that $\{\sum_{k=-\infty}^{+\infty} X_{m-k} Y_k\}_{m=-\infty}^{+\infty}$ is indeed the Fourier coefficients sequence of $X(t)Y(t)$ because we readily have

$$\left\| X(t)Y(t) - \sum_{m=-\infty}^{+\infty} \left(\sum_{k=-\infty}^{+\infty} X_{m-k} Y_k \right) e^{jm\omega_h t} \right\|_{L_2[0,h]} = 0 \tag{26}$$

(since (25) holds for a.e. $t_0 \in [0, h]$ by the assumption), and the Fourier series expansion is unique. This gives $\mathcal{T}\{X(t)Y(t)\} = \mathcal{T}\{X(t)\}\mathcal{T}\{Y(t)\}$. The same is true for $\mathcal{T}\{Y(t)X(t)\} = \mathcal{T}\{Y(t)\}\mathcal{T}\{X(t)\}$. ∎

Proof of Lemma 2.2. Take an arbitrary $f(t) \in L_2[0,h]$, and expand it into the Fourier series expansion $f(t) = \sum_{n=-\infty}^{+\infty} f_n e^{jn\omega_h t}$ with $\underline{f} := [\ldots, f_{-1}^T, f_0^T, f_1^T, \ldots]^T \in l_2$. Note that, for any $\epsilon > 0$, we can find $\underline{d} \in l_1$ such that $||\underline{f} - \underline{d}||_{l_2} < \epsilon$ since l_1 is dense in l_2. Now construct $d(t) := \sum_{n=-\infty}^{+\infty} d_n e^{jn\omega_h t}$. It is easy to see that $d(t) \in L_{\mathrm{CAC}}[0,h]$ and

$$||f(\cdot) - d(\cdot)||_{L_2[0,h]} = \left|\left| \sum_{n=-\infty}^{+\infty} (f_n - d_n) e^{jn\omega_h t} \right|\right|_{L_2[0,h]} = ||\underline{f} - \underline{d}||_{l_2} < \epsilon$$

by the Parseval theorem. This completes the proof. □

Proof of Lemma 2.3. Taking $f(t) \in L_{\mathrm{CAC}}[0,h]$ and expanding it into the Fourier series expansion $f(t) = \sum_{n=-\infty}^{+\infty} f_n e^{jn\omega_h t}$, it follows that $\underline{f} := [\ldots, f_{-1}^T, f_0^T, f_1^T, \ldots]^T \in l_1$. The converse is also true. On the other hand, by the assumption on $X(t)$, it follows from Lemma 2.1 that $\underline{X}\,\underline{f} = \underline{y}$, where $\underline{y}$ is similarly defined to $\underline{f}$ but in terms of the Fourier coefficients of $X(t)f(t)$. Thus it follows that

$$||\underline{X}||_{l_2/l_1(l_2)} := \sup_{\underline{f} \in l_1} \left\{ \frac{||\underline{X}\,\underline{f}||_{l_2}}{||\underline{f}||_{l_2}} \right\} = \sup_{f(t) \in L_{\mathrm{CAC}}[0,h]} \left\{ \frac{||X(\cdot)f(\cdot)||_{L_2[0,h]}}{||f(\cdot)||_{L_2[0,h]}} \right\} =: ||X(\cdot)||_*.$$

Obviously, $||\underline{X}||_{l_2/l_1(l_2)} = ||\underline{X}||_{l_2/l_2}$ since l_1 is dense in l_2. Similarly, from Lemma 2.2, $||X(\cdot)||_* = ||X(\cdot)||_{L_2[0,h]/L_2[0,h]}$. Hence we obtain $||\underline{X}||_{l_2/l_2} = ||X(\cdot)||_{L_2[0,h]/L_2[0,h]} = \sup_{t \in [0,h]} ||X(t)||$. Since $X(t) \in L_{\mathrm{PCC}}[0,h]$ by the assumption, the boundedness assertion follows. □

Proof of Lemma 2.4. Let $\underline{x} \in l_2$. For any $\epsilon > 0$, there exists $\underline{x}' \in l_2$ with only finite nonzero entries such that $||\underline{x} - \underline{x}'||_{l_2} < \epsilon$. It is obvious that $\underline{E}(j0)\underline{x}' \in l_2$. Hence $\underline{x}' \in l_E$, which implies that l_E is dense in l_2.

Furthermore, it is clear that $\underline{x} \in l_E$ if and only if $\sum_{\substack{m=-\infty \\ m \neq 0}}^{+\infty} m^2 \omega_h^2 ||[\underline{x}]_m||^2 < \infty$. It follows from the Cauchy–Schwarz inequality that if $\underline{x} \in l_E$, then

$$\sum_{\substack{m=-\infty \\ m \neq 0}}^{+\infty} ||[\underline{x}]_m|| = \sum_{\substack{m=-\infty \\ m \neq 0}}^{+\infty} m \cdot \frac{1}{m} ||[\underline{x}]_m|| \leq \left(\sum_{\substack{m=-\infty \\ m \neq 0}}^{+\infty} m^2 ||[\underline{x}]_m||^2 \right)^{\frac{1}{2}} \left(\sum_{\substack{m=-\infty \\ m \neq 0}}^{+\infty} \frac{1}{m^2} \right)^{\frac{1}{2}} \leq M < \infty$$

for some $M > 0$. This implies that if $\underline{x} \in l_E$, then $\underline{x} \in l_1$. The fact that l_E is dense in l_1 can be shown in exactly the same way as in the proof for the first assertion (i.e., via truncation). Now take

$$[\underline{x}]_m = \begin{cases} \frac{1}{|m|^{\frac{5}{4}}} [1, 0, \ldots, 0]^T & (m \neq 0), \\ 0 & (m = 0). \end{cases}$$

Then it can be shown that $\underline{x} \in l_1$ but $\underline{E}(j0)\underline{x} \notin l_2$, which says that l_E is a proper subset of l_1. □

Proof of Proposition 5.1. A complete proof can be found in [34]. Here we give only an outline of the proof. By the Floquet theorem, the transition matrix of (1) can be written as $\Phi(t,0) = P(t,0)e^{Qt}$ when the initial time $t_0 = 0$. Thus the impulse response of the system to the input $e_i \delta(t-\tau)$ ($\delta(t-\tau)$ is the delta function imposed at $t = \tau \geq 0$, e_i is the ith natural basis of $\mathbf{R}^m$) is given by

$$(Ge_i\delta_\tau)(t) = \begin{cases} C(t)P(t,0)e^{Q(t-\tau)}P^{-1}(\tau,0)B(\tau)e_i & (t \geq \tau), \\ 0 & (t < \tau). \end{cases} \tag{27}$$

Here we further define $\tilde{B}^T := [\ldots, \hat{B}_{-1}^T, \hat{B}_0^T, \hat{B}_1^T, \ldots]^T$, $\tilde{C} := [\ldots, \hat{C}_1, \hat{C}_0, \hat{C}_{-1}, \ldots]$, and $\Lambda(t) := [\ldots, e^{j\omega_h t} I, I, e^{-j\omega_h t} I, \ldots]^T$ with $\{\hat{B}\}_{m=-\infty}^{+\infty}$ and $\{\hat{C}\}_{m=-\infty}^{+\infty}$ being the Fourier coefficients sequence of $\hat{B}(t)$ and $\hat{C}(t)$, respectively. Then, by the assumptions on $\hat{B}(t)$ and $\hat{C}(t)$, $\hat{B}(t) = \Lambda(\tau)^* \tilde{B}$ and $\hat{C}(t) = \tilde{C}\Lambda(t)$ hold. Therefore, taking the Fourier transformation on (27) about t, we obtain

$$\begin{aligned} F[(Ge_i\delta_\tau)(t)](j\omega) &= \tilde{C} \int_\tau^{+\infty} \Lambda(t) e^{Q(t-\tau)} e^{-j\omega t} dt \Lambda(\tau)^* \tilde{B} e_i \\ &= \tilde{C}(\underline{E}(j\omega) - \underline{Q})^{-1} \Lambda(\tau)\Lambda(\tau)^* \tilde{B} e_i e^{-j\omega\tau}. \end{aligned} \tag{28}$$

In (28), the order of the integral and the infinite summation caused by $\hat{C}(t)\Lambda(t)$ is interchanged. This is validated by the stability assumption, $\hat{C}(t) \in L_{CAC}[0, h]$, and by the Levi theorem [21, p. 577].

Hence, by the time-domain definition of the H_2 norm, we have

$$\begin{aligned} ||\mathcal{G}||_2^2 &= \frac{1}{h} \int_0^h \sum_{i=1}^m \operatorname{trace} \left(\int_0^{+\infty} (Ge_i\delta_\tau)(t)(Ge_i\delta_\tau)^*(t) dt \right) d\tau \\ &= \frac{1}{2\pi h} \int_0^h \sum_{i=1}^m \operatorname{trace} \left(\int_{-\infty}^{+\infty} \tilde{C}(\underline{E}(j\omega) - \underline{Q})^{-1} \Lambda(\tau)\Lambda(\tau)^* \tilde{B} e_i \right. \\ &\qquad \left. \cdot e_i^* \tilde{B}^* \Lambda(\tau)\Lambda(\tau)^* (\underline{E}(j\omega) - \underline{Q})^{-*} \tilde{C}^* d\omega \right) d\tau \\ &= \frac{1}{2\pi h} \int_0^h \operatorname{trace} \left(\int_{-\infty}^{+\infty} \tilde{C}(\underline{E}(j\omega) - \underline{Q})^{-1} \Lambda(\tau)\Lambda(\tau)^* \tilde{B} \right. \\ &\qquad \left. \cdot \tilde{B}^* \Lambda(\tau)\Lambda(\tau)^* (\underline{E}(j\omega) - \underline{Q})^{-*} \tilde{C}^* d\omega \right) d\tau \\ &= \frac{1}{2\pi h} \int_{-\infty}^{+\infty} \operatorname{trace} \left(\tilde{C}(\underline{E}(j\omega) - \underline{Q})^{-1} \right. \\ &\qquad \left. \cdot \left[\int_0^h \Lambda(\tau)\Lambda(\tau)^* \tilde{B} \tilde{B}^* \Lambda(\tau)\Lambda(\tau)^* d\tau \right] (\underline{E}(j\omega) - \underline{Q})^{-*} \tilde{C}^* \right) d\omega \end{aligned} \tag{29}$$

by the Parseval theorem and (28). In (29), the orders of the integrals and the infinite summations are interchanged repeatedly. These can be validated by the Fubini theorem [21, p. 598] and the Levi theorem. Noting also that the trace computations are done on finite-dimensional matrices, the trace computations here are actually only finite summations.

Furthermore, it is easy to see that

$$\Lambda(\tau)\Lambda(\tau)^* \tilde{B}\, \tilde{B}^* \Lambda(\tau)\Lambda(\tau)^* = R(\tau)\Lambda(\tau)\Lambda(\tau)^* R(\tau)^*, \tag{30}$$

where $R(\tau) := \operatorname{diag}[\ldots, \hat{B}(\tau), \hat{B}(\tau), \hat{B}(\tau), \ldots]$. By the assumption, $\hat{B}(\tau)$ has the absolutely convergent Fourier series expansion $\hat{B}(\tau) = \sum_{q=-\infty}^{+\infty} \hat{B}_q e^{jq\omega_h\tau}$. Hence, it follows from Cauchy's rule that $\hat{B}(\tau)\hat{B}(\tau)^* = \sum_{m=-\infty}^{+\infty} (\sum_{n=-\infty}^{+\infty} \hat{B}_{m-n} \hat{B}_n^*) e^{jm\omega_h\tau}$, which implies that

$$\frac{1}{h} \int_0^h R(\tau)\Lambda(\tau)\Lambda(\tau)^* R(\tau)^* d\tau = \underline{\hat{B}}\, \underline{\hat{B}}^* \tag{31}$$

with $\hat{\underline{B}}$ similarly defined to $\underline{A}$ but in terms of $\{\hat{B}_q\}_{q=-\infty}^{+\infty}$. Finally, using (31) in (29) yields

$$\begin{aligned}
||\mathcal{G}||_2^2 &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} \mathrm{trace}(\tilde{C}(\underline{E}(j\omega)-\underline{Q})^{-1}\hat{\underline{B}}\,\hat{\underline{B}}^*(\underline{E}(j\omega)-\underline{Q})^{-*}\tilde{C}^*)d\omega \\
&= \frac{1}{2\pi}\sum_{m=-\infty}^{+\infty}\int_{-\frac{\omega_h}{2}}^{\frac{\omega_h}{2}} \mathrm{trace}(\tilde{C}(\underline{E}(j\varphi_m)-\underline{Q})^{-1}\hat{\underline{B}}\,\hat{\underline{B}}^*(\underline{E}(j\varphi_m)-\underline{Q})^{-*}\tilde{C}^*)d\varphi \\
&= \frac{1}{2\pi}\int_{-\frac{\omega_h}{2}}^{\frac{\omega_h}{2}}\sum_{m=-\infty}^{+\infty} \mathrm{trace}(\tilde{C}(\underline{E}(j\varphi_m)-\underline{Q})^{-1}\hat{\underline{B}}\,\hat{\underline{B}}^*(\underline{E}(j\varphi_m)-\underline{Q})^{-*}\tilde{C}^*)d\varphi \\
&= \frac{1}{2\pi}\int_{-\frac{\omega_h}{2}}^{\frac{\omega_h}{2}} \mathrm{trace}(\hat{\underline{C}}(\underline{E}(j\varphi)-\underline{Q})^{-1}\hat{\underline{B}}\,\hat{\underline{B}}^*(\underline{E}(j\varphi)-\underline{Q})^{-*}\hat{\underline{C}}^*)d\varphi
\end{aligned}$$

as claimed. In the above, we have interchanged the order of the integral and the summation. To validate this, it suffices to show that the convergence of

$$\sum_{|m|\le M} \mathrm{trace}(\tilde{C}(\underline{E}(j\varphi_m)-\underline{Q})^{-1}\hat{\underline{B}}\,\hat{\underline{B}}^*(\underline{E}(j\varphi_m)-\underline{Q})^{-*}\tilde{C}^*) \to \mathrm{trace}(\hat{\underline{G}}(j\varphi)^*\hat{\underline{G}}(j\varphi))$$

is uniform over $\varphi \in \mathcal{I}_0$ as $M \to \infty$. This can be completed by using the trace formula on an orthonormal basis of l_2 [21]. A full explanation is given in [34]. □

Proof of Proposition 5.2. By the Floquet theorem, the L_2-stability is obvious. Then, for any $u(t) \in L_2$, the output $y(t)$ belongs to L_2. Also, C_0^1 is a dense subset of L_2 [21, Exercise D.13.3, p. 593]. Therefore, it makes sense to define the Fourier transforms $U(j\omega)$ and $Y(j\omega)$ for the input $u(t) \in C_0^1$ and the output $y(t)$. We compute $Y(j\omega)$ in four steps.

Step 1. The Fourier transform of the signal p (see Figure 1) is given by

$$P(j\omega) = \int_{-\infty}^{+\infty}\left(\sum_{m=-\infty}^{+\infty}\hat{B}_m e^{jm\omega_h t}\right)u(t)e^{-j\omega t}dt = \sum_{m=-\infty}^{+\infty}\hat{B}_m U(j(\omega - m\omega_h)), \tag{32}$$

which is well defined since $\hat{B}(t)$ is L_2-stable (by the boundedness of $\hat{B}(t)$ on $[0,h]$). Here, the order of infinite integral and infinite summation is interchanged. This is valid by the absolute convergence of the Fourier series expansion of $\hat{B}(t)$ and the fact that $u(t)$ has compact support.

Step 2. Imposing the signal p to the LTI subsystem of Figure 1, the Fourier transform of q is

$$Q(j\omega) = (j\omega I - Q)^{-1}\sum_{m=-\infty}^{+\infty}\hat{B}_m U(j(\omega - m\omega_h)). \tag{33}$$

Since $u(t) \in C_0^1$, it is clear that $\hat{B}(t)u(t) \in L_1$. Also, by the stability assumption, the LTI subsystem of Figure 1 is L_1-stable [25]. Hence $q(t) \in L_1$. Now truncate $q(t)$ as follows. Obviously, $q_{_T}(t) \in L_1$.

$$q_{_T}(t) = \begin{cases} q(t) & (0 \le t \le T), \\ 0 & (t > T). \end{cases}$$

Based on the facts that $q(t)$ and $q_{_T}(t)$ belong to L_1 for all $T > 0$, we have

$$\lim_{T\to\infty} Q_T(j\omega) = Q(j\omega) \tag{34}$$

uniformly over $\omega \in (-\infty, +\infty)$, where $Q_T(j\omega)$ is the Fourier transform of the signal $q_{_T}(t)$ since

$$||Q_T(j\omega) - Q(j\omega)|| = \left|\left|\int_0^\infty (q_{_T}(t) - q(t))e^{-j\omega t}dt\right|\right| \le \int_0^\infty ||q_{_T}(t) - q(t)||dt \to 0 \quad (T \to \infty).$$

Step 3. Let $\hat{y}(t)$ be the output of $\hat{C}(t)$ to the input $q(t)$, and let $\hat{y}_{_T}(t)$ be that corresponding to the truncated signal $q_{_T}(t)$, which has compact support. Then we clearly have $\hat{y}_{_T}(t) = \hat{C}(t)q_{_T}(t)$ so that by repeating the arguments about (32) on $\hat{C}(t)$, the Fourier transform of $y_{_T}(t)$ is given by

$$\hat{Y}_T(j\omega) = \sum_{n=-\infty}^{+\infty} \hat{C}_n Q_T(j(\omega - n\omega_h)). \tag{35}$$

It is obvious that $\hat{y}(t)$ and $\hat{y}_{_T}(t)$ belong to L_1 since $\hat{C}(t)$ is bounded on $t \ge 0$. Based on this fact, repeating the arguments about $q(t)$ and $q_{_T}(t)$ on $y(t)$ and $y_{_T}(t)$, it follows that $\lim_{T\to\infty} Y_T(j\omega) = Y(j\omega)$ uniformly over $\omega \in (-\infty, +\infty)$. This further gives the relation

$$\hat{Y}(j\omega) = \sum_{n=-\infty}^{+\infty} \hat{C}_n Q(j(\omega - n\omega_h)) \tag{36}$$

since it is evident that

$$\left|\left| \sum_{n=-\infty}^{+\infty} \hat{C}_n Q_T(j(\omega - n\omega_h)) - \sum_{n=-\infty}^{+\infty} \hat{C}_n Q(j(\omega - n\omega_h)) \right|\right| \le \sum_{n=-\infty}^{+\infty} ||\hat{C}_n|| \cdot ||Q_T(j(\omega - n\omega_h)) - Q(j(\omega - n\omega_h))|| \to 0 \quad (T \to \infty)$$

uniformly over $\omega \in (-\infty, +\infty)$ by (34) and $\sum_{n=-\infty}^{+\infty} ||\hat{C}_n|| < \infty$, which follows from the absolute convergence of the Fourier series expansion of $\hat{C}(t)$.

Step 4. Taking the feedforward term $D(t)$ into consideration and lifting the Fourier transform $Y(j\omega)$ of the whole output to its SD-Fourier transform $\underline{Y}_{\rm SD}(j\varphi)$ leads to the assertion (1).

To show the assertion (2), by the well-known Parseval theorem, we note that

$$||y(\cdot)||_{L_2}^2 = \frac{1}{2\pi}\int_{-\infty}^{+\infty} Y(j\omega)^* Y(j\omega)d\omega = \frac{1}{2\pi}\sum_{m=-\infty}^{+\infty}\int_{\mathcal{I}_0} Y(j\varphi_m)^* Y(j\varphi_m)d\varphi$$
$$= \frac{1}{2\pi}\int_{\mathcal{I}_0} \underline{Y}_{\rm SD}(j\varphi)^* \underline{Y}_{\rm SD}(j\varphi)d\varphi = \frac{1}{2\pi}\int_{\mathcal{I}_0} \underline{U}_{\rm SD}(j\varphi)^* \underline{\hat{G}}(j\varphi)^* \underline{\hat{G}}(j\varphi)\underline{U}_{\rm SD}(j\varphi)d\varphi.$$

To complete the proof, it remains to show that the interchange of the integral and summation is valid. To this end, it is enough to show that the convergence of $\sum_{m=-M}^{M} Y(j\varphi_m)^* Y(j\varphi_m) \to \underline{Y}_{\rm SD}(j\varphi)^* \underline{Y}_{\rm SD}(j\varphi)$ as $M \to \infty$ is uniform over $\varphi \in \mathcal{I}_0$. A complete proof can be found in [34]. □

REFERENCES

[1] M. Araki, Y. Ito, and T. Hagiwara, *Frequency response of sampled-data systems*, Automatica J. IFAC, 32 (1996), pp. 483–497.

[2] B. Bamieh and J. B. Pearson, *A general framework for linear periodic systems with applications to H^∞ sampled-data control*, IEEE Trans. Automat. Control, 37 (1992), pp. 418–435.

[3] B. Bamieh and J. B. Pearson, *The H^2 problem for sampled-data systems*, Systems Control Lett., 19 (1992), pp. 1–12.

[4] J. H. Braslavsky, R. H. Middleton, and J. S. Freudenberg, *L_2-induced norms and frequency-gains of sampled-data sensitivity operators*, IEEE Trans. Automat. Control, 43 (1998), pp. 252–258.

[5] T. Chen and B. A. Francis, *Optimal Sampled-Data Control Systems*, Springer-Verlag, London, 1995.

[6] R. V. Churchill, *Fourier Series and Boundary Value Problems*, 2nd ed., McGraw-Hill, New York, 1963.

[7] C. L. Devito, *Functional Analysis and Linear Operator Theory*, Addison-Wesley, Reading, MA, 1990.

[8] G. E. Dullerud and K. Glover, *Robust stabilization of sampled-data systems to structured LTI perturbations*, IEEE Trans. Automat. Control, 38 (1993), pp. 1497–1508.

[9] G. E. Dullerud, *Control of Uncertain Sampled-Data Systems*, Birkhäuser Boston, Boston, 1996.

[10] M. Farkas, *Periodic Motions*, Springer-Verlag, New York, 1994.

[11] J. S. Freudenberg, R. H. Middleton, and J. H. Braslavsky, *Inherent design limitations for linear sampled-data feedback systems*, Internat. J. Control, 61 (1995), pp. 1387–1421.

[12] I. Gohberg, S. Goldberg, and M. A. Kaashoek, *Classes of Linear Operators. Vol.* I, Birkhäuser Verlag, Basel, 1990.

[13] I Gohberg, S. Goldberg, and M. A. Kaashoek, *Classes of Linear Operators. Vol.* II, Birkhäuser Verlag, Basel, 1993.

[14] G. C. Goodwin and M. Salgado, *Frequency domain sensitivity functions for continuous-time systems under sampled data control*, Automatica J. IFAC, 30 (1994), pp. 1263–1270.

[15] T. Hagiwara, *Nyquist stability criterion and positive realness of sampled-data systems*, in Proceedings of the American Control Conference, Chicago, IL, 2000, pp. 958–962.

[16] T. Hagiwara and M. Araki, *FR-operator approach to the H_2 analysis and synthesis of sampled-data systems*, IEEE Trans. Automat. Control, 40 (1995), pp. 1411–1421.

[17] T. Hagiwara and M. Araki, *Robust stability of sampled-data systems under possibly unstable additive/multiplicative perturbations*, IEEE Trans. Automat. Control, 43 (1998), pp. 1340–1346.

[18] H. Fukawa, *Mathematics in Control and Vibration*, Corona, 1974 (in Japanese).

[19] G. A. Korn and T. M. Korn, *Mathematical Handbook for Scientists and Engineers*, McGraw-Hill, New York, 1968.

[20] D. L. Lukes, *Differential Equations: Classical to Controlled*, Academic Press, New York, 1982.

[21] A. W. Naylor and G. R. Sell, *Linear Operator Theory in Engineering and Science,* Springer-Verlag, New York, 1982.

[22] G. D. Nicolao, G. F. Trecate, and S. Pinzon, *Zeros of continuous-time linear periodic systems*, Automatica J. IFAC, 34 (1998), pp. 1651–1655.

[23] W. J. Rugh, *Linear System Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[24] R. G. Shenoy, D. Burnside, and T. W. Parks, *Linear periodic systems and multirate filter design*, IEEE Trans. Signal Processing, 42 (1994), pp. 2242–2256.

[25] M. Vidyasagar, *Nonlinear Systems Analysis*, 2nd ed., Prentice Hall, Upper Saddle River, NJ, 1998.

[26] N. M. Wereley and S. R. Hall, *Frequency response of linear time periodic systems*, Proceedings of the IEEE Conference on Decision and Control, Honolulu, HI, 1990, pp. 3650–3655.

[27] N. M. Wereley, *Analysis and Control of Linear Periodically Time Varying Systems*, Ph.D. Thesis, Dept. of Aeronautics and Astronautics, M.I.T., Cambridge, MA, 1990.

[28] Y. Yamamoto and P. Khargonekar, *Frequency response of sampled-data systems,* IEEE Trans. Automat. Control, 41 (1996), pp. 166–176.

[29] Y. Yamamoto and M. Araki, *Frequency response of sampled-data systems—their equivalence and relationships*, Linear Algebra Appl., 205/206 (1994), pp. 1319–1339.

[30] Y. Yamamoto, *A function space approach to sampled-data control systems and tracking problems*, IEEE Trans. Automat. Control, 39 (1994), pp. 703–713.

[31] E. Zeidler, *Applied Functional Analysis—Applications to Mathematical Physics*, Springer-Verlag, New York, 1995.

[32] C. Zhang and J. Zhang, *H_2 performance of continuous time periodically time varying controllers*, Systems Control Lett., 32 (1997), pp. 209–221.

[33] J. Zhou and T. Hagiwara, *H_2 and H_∞ Norm Computations of Linear Continuous-Time Periodic Systems via the Skew Analysis of Frequency Response Operators*, Tech. rep. 00-04, Automatic Control Engineering Group, Dept. of Electrical Engineering, Kyoto University, Kyoto, Japan, 2000. Automatica J. IFAC, to appear.

[34] J. Zhou and T. Hagiwara, *Existence Conditions and Properties of the Frequency Response Operators of Continuous-Time Periodic Systems*, Tech. rep. 00-05, Automatic Control Engineering Group, Dept. of Electrical Engineering, Kyoto University, Kyoto, Japan, 2000.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 6, pp. 1888–1904

A UNIFYING INTEGRAL ISS FRAMEWORK FOR STABILITY OF NONLINEAR CASCADES*

MURAT ARCAK†, DAVID ANGELI‡, AND EDUARDO SONTAG§

Abstract. We analyze nonlinear cascades in which the driven subsystem is integral input-to-state stable (ISS), and we characterize the admissible integral ISS gains for stability. This characterization makes use of the convergence speed of the driving subsystem and allows a larger class of gain functions when the convergence is faster. We show that our integral ISS gain characterization unifies different approaches in the literature which restrict the nonlinear growth of the driven subsystem and the convergence speed of the driving subsystem. The result is used to develop a new observer-based backstepping design in which the growth of the nonlinear damping terms is reduced.

Key words. nonlinear cascades, stabilization, integral input-to-state stability

AMS subject classifications. 93C10, 93D05, 93D15, 93D25

PII. S0363012901387987

1. Introduction.

Studies on the stabilization of cascade systems have paved the road to major advances in nonlinear control theory. Among these advances are several constructive design methods such as *backstepping* and *forwarding*, which are based on recursive applications of cascade designs (see, e.g., Sepulchre, Janković, and Kokotović [17]), and the discovery of structural obstacles to stabilization such as the *peaking phenomenon* (see Sussmann and Kokotović [23]).

One of the main motivations for the stabilization of cascades came from the linear-nonlinear cascade

$$\dot{x} = f(x, z), \tag{1}$$

$$\dot{z} = Az + Bu \tag{2}$$

resulting from input-output linearization. Because global asymptotic stability (GAS) of the x-subsystem $\dot{x} = f(x, 0)$ is not sufficient to achieve GAS of the whole cascade with z-feedback $u = Kz$, alternative designs which employ x-feedback were developed, such as the *feedback passivation* design of Kokotović and Sussmann [8]. To achieve GAS by z-feedback, Sontag [18], Seibert and Suarez [16], Mazenc and Praly [11], Janković, Sepulchre, and Kokotović [6], and Panteley and Loría [12, 13] studied general cascades in the which the z-subsystem is nonlinear and derived conditions for the x- and z-subsystems that ensure stability of the cascade. Among these results, a particularly useful one is the input-to-state stability (ISS) condition in [19], which states that if the x-subsystem is ISS with input z and the z-subsystem is GAS, then the cascade is GAS. This result has been widely used for nonlinear designs based on the *normal form* (1)–(2), in which the *zero dynamics* (1) is ISS. Other results, such

*Received by the editors April 12, 2001; accepted for publication (in revised form) October 24, 2001; published electronically March 5, 2002.
http://www.siam.org/journals/sicon/40-6/38798.html

†Department of Electrical, Computer and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180-3590 (arcak@ecse.rpi.edu).

‡Dipartimento Sistemi e Informatica, University of Florence, 50139 Firenze, Italy (angeli@dsi.unifi.it).

§Department of Mathematics, Rutgers University, New Brunswick, NJ 08903 (sontag@hilbert.rutgers.edu).

as [11] and [6], make less restrictive assumptions than ISS for the x-subsystem but restrict the z-subsystem to be locally exponentially stable (LES).

The integral version of ISS (iISS), recently introduced in [20], requires that an energy norm of the input be bounded to ensure boundedness of the states. As shown in [20], iISS is less restrictive than ISS because, in an iISS system, a bounded input may lead to unbounded solutions if its energy norm is infinite.

In this paper, we analyze the stability of nonlinear cascades in which the x-subsystem is iISS and the z-subsystem is GAS. The admissible iISS gains for stability are characterized from the speed of convergence of the z-subsystem. When the convergence is fast, the iISS gain function of the x-subsystem is allowed to be "steep" at zero. We show that this trade-off between slower convergence and steeper iISS gain encompasses and unifies several results in the literature. In particular, if the x-subsystem is ISS, then the slope of its iISS gain function is very gentle at zero and tolerates every GAS z-subsystem no matter how slow its convergence is. On the other hand, if the convergence is exponential, that is, if the z-subsystem is LES, then the cascade is stable for a large class of iISS gains. This class includes all iISS x-subsystems that are affine in the input z. Thus, for systems like (1)–(2), where a control law can be designed to render the z-subsystem GAS and LES, the iISS of the x-subsystem ensures GAS of the cascade.

In section 2, we define the new concepts used in the paper and present lemmas which are preliminary to our main results. In section 3, we present our main result, Theorem 1, which characterizes the admissible iISS gains from the speed of convergence of the z-subsystem. We show that several results in the literature are special cases of Theorem 1, including those that restrict the x-subsystem to be ISS (Corollary 1) and those that restrict the z-subsystem to be LES (Corollary 2). In section 4, we show that Corollary 2 restricts the nonlinear growth of the interconnection term and illustrate with an example that violating this growth condition leads to instability of the cascade.

The second main contribution of the paper is an output-feedback application of our cascade result. Due to the absence of a separation principle, it is necessary to design control laws that guarantee robustness against the observer error. We present a design which renders the system iISS with respect to the observer error and hence ensures robustness when the error is exponentially decaying. The advantage of our design over the observer-based backstepping scheme of Kanellakopoulos, Kokotović, and Morse [7] is that we employ "weak" nonlinear damping terms which grow slower than those in [7] and result in a "softer" control law. The main features of the design are discussed and illustrated in an example in section 5. The general design procedure and its stability proof are given in section 6.

2. Definitions and preliminary lemmas. In this section, we give definitions and present lemmas that will be used in the rest of the paper. The proofs are given in section 7.

We first recall standard definitions: $\mathcal{K}$ is the class of functions $\mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ which are zero at zero, strictly increasing, and continuous. $\mathcal{K}_\infty$ is the subset of class-$\mathcal{K}$ functions that are unbounded. $\mathcal{L}$ is the set of functions $\mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ which are continuous, decreasing, and converging to zero as their argument tends to $+\infty$. $\mathcal{KL}$ is the class of functions $\mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ which are class-$\mathcal{K}$ in the first argument and class-$\mathcal{L}$ in the second argument.

DEFINITION 1. *We say that the function $\mu(\cdot) : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is class-$\mathcal{K}_o$ if it is class-$\mathcal{K}$ and $\mathcal{O}(s)$ near $s = 0$; that is, for all $s \in [0, 1]$, $\mu(s) \leq ks$ for some $k > 0$.*

DEFINITION 2 (see [20]). *The system*

$$\dot{x} = f(x, z) \tag{3}$$

is said to be iISS with input z if there exist a class-$\mathcal{K}_\infty$ function $\omega(\cdot)$, a class-$\mathcal{KL}$ function $\beta(\cdot,\cdot)$, and a class-$\mathcal{K}$ iISS gain $\mu(\cdot)$ such that, for all $t \geq 0$,

$$\omega(|x(t)|) \leq \beta(|x(0)|, t) + \int_0^t \mu(|z(\tau)|)d\tau. \tag{4}$$

Lemma 1 computes an iISS gain $\mu(\cdot)$ from the derivative of an *iISS Lyapunov function.*

LEMMA 1.

(i) *If there exists a C^1, positive definite, radially unbounded function $V(x)$ satisfying*

$$\frac{\partial V}{\partial x} f(x, z) \leq -\rho(|x|) + \mu(|z|) \tag{5}$$

for some positive definite function $\rho(\cdot)$, then the system (3) *with input z is iISS with gain $\mu(\cdot)$.*

(ii) *If there exists a C^1, positive definite, radially unbounded function $V(x)$ satisfying*

$$\frac{\partial V}{\partial x} f(x, z) \leq \sigma(|z|) \tag{6}$$

for some class-$\mathcal{K}$ function $\sigma(\cdot)$, and if the system (3) *is GAS when $z \equiv 0$, then there exists a class-$\mathcal{K}_o$ function $\theta_o(\cdot)$ such that* (3) *is iISS with gain $\mu(\cdot) = \sigma(\cdot) + \theta_o(\cdot)$.*

The following lemma proves that if (3) is affine in the input z, then $\sigma(\cdot)$ in (6) is class-$\mathcal{K}_o$, which means that the iISS gain $\mu(\cdot) = \sigma(\cdot) + \theta_o(\cdot)$ is also class-$\mathcal{K}_o$.

LEMMA 2. *If the input-affine system*

$$\dot{x} = f(x) + g(x)z \tag{7}$$

is iISS, then it is also iISS with a class-$\mathcal{K}_o$ gain $\mu(\cdot)$.

It is known [20] that ISS implies iISS. We further show that ISS allows us to select the iISS gain $\mu(\cdot)$ in (4) to match any desired class-$\mathcal{K}$ function $\tilde{\mu}(\cdot)$ locally.

LEMMA 3. *Suppose the system* (3) *is ISS. Then, for any class-$\mathcal{K}$ function $\tilde{\mu}(\cdot)$, it is iISS with a gain $\mu(\cdot)$ satisfying $\mu(s) = \tilde{\mu}(s)$ for all $s \in [0, 1]$.*

It is proved in [20, Proposition 7] that, for a GAS system $\dot{z} = q(z)$, the solutions $z(t)$ satisfy

$$|z(t)| \leq \alpha(e^{-kt}\gamma(|z(0)|)) \tag{8}$$

for some constant $k > 0$ and class-$\mathcal{K}_\infty$ functions $\alpha(\cdot)$ and $\gamma(\cdot)$. The following definition classifies GAS systems using the function $\alpha(\cdot)$ as an index of their speed of convergence to zero.

DEFINITION 3. *Given a class-$\mathcal{K}_\infty$ function $\alpha(\cdot)$, we say that the system $\dot{z} = q(z)$ is GAS(α) if there exist a class-$\mathcal{K}_\infty$ function $\gamma(\cdot)$ and a positive constant $k > 0$ such that* (8) *holds for all $z(0)$.*

Thus, for the identity function $\alpha(\cdot) = I(\cdot)$, GAS(I) consists of systems in which the convergence is exponential. We next show that $\alpha(\cdot)$ is determined by the local speed of convergence.

LEMMA 4. *If the equilibrium $z = 0$ of $\dot{z} = q(z)$ is GAS and if there exist a constant $\epsilon > 0$ and a $\mathcal{K}_\infty$ function $\tilde{\gamma}(\cdot)$ such that*

$$|z(0)| \le \epsilon \quad \Rightarrow \quad |z(t)| \le \alpha(e^{-kt}\tilde{\gamma}(|z(0)|)), \tag{9}$$

then there exists a class-$\mathcal{K}_\infty$ function $\gamma(s) = \mathcal{O}(\tilde{\gamma}(s))$ near $s = 0$ such that (8) holds for all $z(0)$; that is, $\dot{z} = q(z)$ is GAS(α).

The following definition will be used in our cascade result to characterize the admissible iISS gains from the speed of convergence $\alpha(\cdot)$ of the GAS(α) driving subsystem.

DEFINITION 4. *Given a class-$\mathcal{K}$ $\alpha(\cdot)$, we say that the function $\mu(\cdot)$ is class-$\mathcal{H}_\alpha$ if it is class-$\mathcal{K}$ and satisfies*

$$\int_0^1 \frac{(\mu \circ \alpha)(s)}{s} ds < \infty. \tag{10}$$

In particular, for the identity function $\alpha(\cdot) = I(\cdot)$, class-$\mathcal{H}_I$ is defined by

$$\int_0^1 \frac{\mu(s)}{s} ds < \infty. \tag{11}$$

Thus $\mu(s)$ is class-$\mathcal{H}_I$ if it is class-$\mathcal{K}_o$ or if $\mu(s) \le s^p$ for some $p > 0$, such as $\mu(s) = \sqrt{s}$.

3. Main results. We consider the cascade

$$\dot{x} = f(x, z), \tag{12}$$

$$\dot{z} = q(z), \tag{13}$$

where $x \in \mathbb{R}^{n_x}$, $z \in \mathbb{R}^{n_z}$, and $f(\cdot,\cdot)$ and $q(\cdot)$ are locally Lipschitz and satisfy $f(0,0) = 0$, $q(0) = 0$. The stability properties to be analyzed are with respect to the origin $(x, z) = (0, 0)$, which is an equilibrium for (12)–(13).

Our main stability result characterizes the admissible iISS gains for the x-subsystem from the speed of convergence of the z-subsystem.

THEOREM 1. *If the z-subsystem (13) is GAS(α) as in (8) and the x-subsystem with input z is iISS with a class-$\mathcal{H}_\alpha$ iISS gain $\mu(\cdot)$ as in (10), then the cascade (12)–(13) is GAS.*

Proof. We note from (8) that

$$\int_0^\infty \mu(|z(\tau)|)d\tau \le \int_0^\infty (\mu \circ \alpha)(\gamma(|z(0)|)e^{-k\tau})d\tau = \frac{1}{k}\int_0^{\gamma(|z(0)|)} \frac{(\mu \circ \alpha)(s)}{s} ds, \tag{14}$$

where $s := \gamma(|z(0)|)e^{-k\tau}$. From (10),

$$\lambda(s') := \frac{1}{k}\int_0^{s'} \frac{(\mu \circ \alpha)(s)}{s} ds \tag{15}$$

exists for all $s' \ge 0$, and it is class-$\mathcal{K}$ because $\lambda(0) = 0$ and $\frac{(\mu\circ\alpha)(s)}{s} > 0$ for all $s > 0$. Thus, from (4),

$$\omega(|x(t)|) \le \beta(|x(0)|, 0) + \lambda(\gamma(|z(0)|)), \tag{16}$$

which proves stability of the cascade (12)–(13). Because $\int_0^\infty \mu(|z(\tau)|)d\tau$ is bounded, (4) implies that $x(t) \to 0$ as $t \to \infty$, as proved in [20, Proposition 6]. ∎

If the x-subsystem with input z is ISS as in [19], then, from Lemma 3, it is also iISS with a class-$\mathcal{H}_\alpha$ gain. This means that no matter what the speed of convergence $\alpha(\cdot)$ is for the z-subsystem, the ISS x-subsystem satisfies the corresponding $\mathcal{H}_\alpha$ iISS gain condition of Theorem 1. Thus Theorem 1 encompasses the following well-known result.

COROLLARY 1. *If the z-subsystem* (13) *is GAS and the x-subsystem is ISS, then the cascade* (12)–(13) *is GAS.*

Another particular case of interest is when the z-subsystem is LES, that is, when (9) holds with $\alpha(\cdot) = I(\cdot)$ and $\tilde{\gamma}(s) = cs$ for some $c \geq 1$. From Lemma 4, there exist a class-$\mathcal{K}_o$ function $\gamma(\cdot)$ and a constant $k > 0$ such that, for all $z(0)$,

$$|z(t)| \leq e^{-kt}\gamma(|z(0)|)\,. \tag{17}$$

This means that the z-subsystem is GAS(I), and hence Theorem 1 requires that the iISS gain be class-$\mathcal{H}_I$.

COROLLARY 2. *If the z-subsystem* (13) *is GAS and LES and the x-subsystem is iISS with a class-$\mathcal{H}_I$ gain $\mu(\cdot)$ as in* (11)*, then the cascade* (12)–(13) *is GAS.*

We note from Lemma 2 that the class-$\mathcal{H}_I$ restriction of Corollary 2 is satisfied when the iISS x-subsystem is affine in z.

COROLLARY 3. *If the z-subsystem* (13) *is GAS and LES and the x-subsystem is iISS and affine in z, then the cascade* (12)–(13) *is GAS.*

Examples 1 and 2 illustrate that the LES and the class-$\mathcal{H}_I$ gain restrictions cannot be removed from Corollary 2.

Example 1. For the cascade system

$$\dot{x} = -\mathrm{sat}(x) + xz, \tag{18}$$

$$\dot{z} = -z^3\,, \tag{19}$$

where $\mathrm{sat}(x) := \mathrm{sgn}(x)\min\{1, |x|\}$, the x-subsystem with input z is iISS with a class-$\mathcal{K}_o$ (hence class-$\mathcal{H}_I$) gain, as verified from $V(x) = \frac{1}{2}\ln(1+x^2)$, which satisfies $\dot{V} \leq |z|$ as in Lemma 1. However, the cascade (18)–(19) has unbounded solutions because the z-subsystem is not LES. To prove this, we let $z(0) = 1$ so that $z(t) = \frac{1}{\sqrt{1+2t}}$, and we let $x(0) > 1$ so that, as long as $x(t) \geq 1$,

$$\dot{x} = \frac{1}{\sqrt{1+2t}}\,x - 1 \quad \Rightarrow \quad x(t) = e^{(\sqrt{1+2t}-1)}\left[x(0) - \int_0^t e^{(1-\sqrt{1+2\tau})}d\tau\right]. \tag{20}$$

Using the change of variables $s = -1 + \sqrt{1+2\tau}$, we obtain

$$\int_0^t e^{(1-\sqrt{1+2\tau})}d\tau \leq \int_0^\infty e^{(1-\sqrt{1+2\tau})}d\tau = \int_0^\infty e^{-s}(s+1)\,ds = 2; \tag{21}$$

thus, if $x(0) \geq 3$, then (20) implies $x(t) \geq e^{(\sqrt{1+2t}-1)}$. This means that $x(t) \geq 1$ for all $t \geq 0$ and $x(t) \to \infty$ as $t \to \infty$.

Example 2. In this example, we show that the class-$\mathcal{H}_I$ gain restriction cannot be removed from Corollary 2. Consider the locally Lipschitz system

$$\dot{x} = -m(x) + g^{-1}\left(g\left(e^{e^{\frac{x-1}{b}}-a}z^2\right)\mathrm{sat}(z^2)/2\right), \tag{22}$$

$$\dot{z} = -z/2, \tag{23}$$

where $g : [0, +\infty) \to [0, 1]$ is defined as $g(x) := e^{-\frac{1}{x}}$ for $x > 0$ and $g(0) = 0$, $m(x)$ is a locally Lipschitz function satisfying $m(x)x > 0$ for all $x \neq 0$,

$$m(x) = \frac{1}{2e^{\frac{|x|-1}{b}}} \qquad \forall\, x \geq 1 + b\ln(a)\,, \tag{24}$$

and $a, b > 0$ are constants to be specified. Using $V(x) = \ln(1 + x^2)$ and $|g(r)| \leq 1$ for all $r \in \mathbb{R}_{\geq 0}$, we obtain

$$\begin{aligned} \dot{V} &= -\frac{2m(x)x}{1+x^2} + \frac{2x}{1+x^2} g^{-1}\left(g\left(e^{e^{\frac{x-1}{b}} - a z^2}\right)\mathrm{sat}(z^2)/2\right) \\ &\leq -\frac{2m(x)x}{1+x^2} + g^{-1}(\mathrm{sat}(z^2)/2)\,, \end{aligned} \tag{25}$$

which, from Lemma 1, proves that the x-subsystem is iISS with input z. Moreover, the z-subsystem is exponentially stable as in Corollary 2. However, because the iISS gain is not class-$\mathcal{H}_I$, the cascade (22)–(23) admits the unbounded solution

$$x(t) = 1 + b\ln(a + t) \tag{26}$$

when $z(0) \in (0, 1)$, $a = \ln(2/z^2(0)g(z^2(0)))$, and $b = 1/2$. To see that (26) satisfies (22), note that the time derivative of $x(t)$ is $\dot{x}(t) = b/(t+a)$ and that the substitution in (22) of

$$m(x(t)) = \frac{1}{2(a+t)} = \frac{b}{a+t}, \tag{27}$$

$$\begin{aligned} g^{-1}\left(g\left(e^{e^{\frac{x(t)-1}{b}} - a z^2(t)}\right)\mathrm{sat}(z^2(t))/2\right) &= g^{-1}(g(e^t z^2(0)e^{-t})z^2(0)e^{-t}/2) \\ = g^{-1}(g(z^2(0))z^2(0)e^{-t}/2) &= \frac{1}{t + \ln(2/z^2(0)g(z^2(0)))} = \frac{2b}{a+t} \end{aligned} \tag{28}$$

indeed yields $\dot{x}(t) = b/(t + a)$.

Theorem 1 characterized the class of admissible iISS gains from the speed of convergence of the z-subsystem. It may appear that this class can be enlarged by a change of coordinates in which the z-subsystem converges faster as in Grüne, Sontag, and Wirth [5]. The following example illustrates that such an attempt fails because in the new coordinates the iISS gain of the x-subsystem becomes steeper.

Example 3. For the system (18)–(19), the change of coordinates

$$\tilde{z} = \Phi(z) := z e^{-\frac{1}{2z^2}} \tag{29}$$

ensures exponential convergence for the $\tilde{z}$-subsystem:

$$\dot{x} = -\mathrm{sat}(x) + x\Phi^{-1}(\tilde{z}), \tag{30}$$

$$\dot{\tilde{z}} = -(1 + [\Phi^{-1}(z)]^2)\tilde{z}\,. \tag{31}$$

Using $V(x) = \frac{1}{2}\ln(1 + x^2)$, we obtain

$$\dot{V} \leq -\frac{x\mathrm{sat}(x)}{1+x^2} + \Phi^{-1}(|\tilde{z}|)\,,$$

which, from Lemma 1, implies that the x-subsystem with input $\tilde{z}$ is iISS with gain $\Phi^{-1}(\cdot)$. The cascade (30)–(31) has unbounded solutions as proved in Example 1 because all derivatives of $\Phi(z)$ vanish at $z = 0$, and hence the inverse function $\Phi^{-1}(\cdot)$ is too steep at zero to satisfy the class-$\mathcal{H}_I$ condition of Corollary 2.

4. Growth restrictions on the interconnection term. It is well known that the nonlinear growth of the interconnection term $h(x,z) := f(x,z) - f(x,0)$ plays an important role for the stability of the cascade (12)–(13), rewritten here as

$$\dot{x} = f(x,0) + h(x,z), \tag{32}$$

$$\dot{z} = q(z)\,. \tag{33}$$

In this section, we show that the class-$\mathcal{H}_I$ gain condition of Corollary 2 imposes a restriction on the nonlinear growth of $h(x,z)$ in x. To this end, we consider the cascade (32)–(33) with $x \in \mathbb{R}$ and with $h(x,z)$ bounded by

$$|h(x,z)| \le \gamma_1(|z|) + \gamma_2(|z|)\varphi(|x|). \tag{34}$$

We characterize the class-$\mathcal{K}$ functions $\varphi(\cdot)$ for which the x-subsystem satisfies the class-$\mathcal{H}_I$ gain condition of Corollary 2.

PROPOSITION 1. *Consider the cascade* (32)–(33)*, where* $x \in \mathbb{R}$*. If* $\dot{x} = f(x,0)$ *is GAS,* $\dot{z} = q(z)$ *is GAS and LES, and* $h(x,z)$ *satisfies* (34) *for some class-*$\mathcal{H}_I$ *functions* $\gamma_1(\cdot)$*,* $\gamma_2(\cdot)$*, and a class-*$\mathcal{K}$ *function* $\varphi(\cdot)$ *satisfying*

$$\int_1^\infty \frac{1}{\varphi(s)} ds = \infty\,, \tag{35}$$

then the origin is GAS.

Proof. To prove that the x-subsystem is iISS with a class-$\mathcal{H}_I$ gain, we let $V(x)$ be a smooth, positive definite function such that $V(x) = V(-x)$, and

$$x \ge 1 \quad \Rightarrow \quad V(x) = V(1) + \int_1^x \frac{1}{\varphi(s)} ds\,. \tag{36}$$

Because of (35), $V(x)$ is radially unbounded and satisfies

$$|x| \ge 1 \quad \Rightarrow \quad \left|\frac{\partial V}{\partial x}\right| = \frac{1}{\varphi(|x|)}\,. \tag{37}$$

Thus, if $|x| \ge 1$, (34) and (37) yield

$$\frac{\partial V}{\partial x}[f(x,0) + h(x,z)] \le \frac{1}{\varphi(1)}\gamma_1(|z|) + \gamma_2(|z|) =: \gamma_5(|z|)\,. \tag{38}$$

If $|x| \le 1$, then $\left|\frac{\partial V}{\partial x}\right| \le b$ for some positive constant b, and

$$\frac{\partial V}{\partial x}[f(x,0) + h(x,z)] \le b\gamma_1(|z|) + b\gamma_2(|z|)\varphi(1) := \gamma_6(|z|)\,. \tag{39}$$

Because $\dot{V} \le \max\{\gamma_5(|z|), \gamma_6(|z|)\}$, the x-subsystem is iISS with a class-$\mathcal{H}_I$ gain as in Corollary 2, and hence the cascade (32)–(33) is GAS. ∎

It is important to note that the growth condition (35) encompasses functions that grow faster than linear, such as $\varphi(|x|) = |x|\ln(|x|)$. On the other hand, (35) disallows $\varphi(|x|) = |x|^2$, $\varphi(|x|) = |x|^3$, etc. Growth conditions similar to (35) have been derived by Mazenc and Praly [11] and, more recently, by Panteley and Loría [13]. Proposition 1 gives a simple iISS interpretation of their more involved Lyapunov arguments. We finally show that (35) is tight and cannot be relaxed.

Example 4. The cascade

$$\dot{x} = -\mathrm{sat}(x) + \varphi(x)z, \tag{40}$$

$$\dot{z} = -z \tag{41}$$

exhibits finite escape time when the class-$\mathcal{K}$ function $\varphi(x)$ fails to satisfy (35), that is,

$$\int_1^\infty \frac{1}{\varphi(s)} ds < \infty\,. \tag{42}$$

To prove this, we let $V(x)$ be as in (36) and note from (42) that there exists a constant $V_\infty > 0$ such that $V(x) < V_\infty$ for all $x \in \mathbb{R}$, and $V(x) \to V_\infty$ as $x \to \infty$. Moreover, from (36),

$$x(t) \geq 1 \quad \Rightarrow \quad \dot{V} = -\frac{1}{\varphi(x(t))} + z(t) \geq -\frac{1}{\varphi(1)} + z(0)e^{-t}\,. \tag{43}$$

If $z(0) > 1/\varphi(1)$, then we can find $T > 0$ such that

$$\phi(t) := \int_0^t \left(-\frac{1}{\varphi(1)} + z(0)e^{-t}\right) dt = -\frac{t}{\varphi(1)} + z(0)(1 - e^{-t}) \tag{44}$$

satisfies $\phi(t) > 0$ for all $t \in (0, T]$. Thus, if $x(0)$ is such that $x(0) \geq 1$ and $V_\infty - \phi(T) \leq V(x(0)) < V_\infty$, then it follows from (43) that

$$V(x(T)) \geq V(x(0)) + \phi(T) \geq V_\infty\,,$$

which proves that $x(T)$ is not defined.

5. Application to output-feedback design. One of the major difficulties in nonlinear output-feedback design is the absence of a separation principle. Even when a nonlinear observer is available, it may be necessary to redesign the control law to make it robust against the observer error. One such design is the observer-based backstepping scheme of Kanellakopoulos, Kokotović, and Morse [7], further extended by Praly and Jiang [14], which makes use of *nonlinear damping* terms to render the system ISS with respect to the observer error. A shortcoming of this design, pointed out by several authors, is the rapid growth of the nonlinearities in the control law due to nonlinear damping terms. Such nonlinearities in the control law make the implementation difficult and are harmful in the presence of unmodeled actuator dynamics, saturation, etc. Efforts to reduce the growth of nonlinear damping terms are restricted to a result for Euler–Lagrange systems, Aamo et al. [1], and an adaptive backstepping design in Krstić, Kanellakopoulos, and Kokotović [10, section 5.8], where stability is achieved with the help of a strengthened parameter identifier.

We now give a systematic procedure to reduce the growth of nonlinear damping terms. Our main idea is to render the system iISS against the observer error. Because iISS is less restrictive than ISS, it is achieved with a "weak" form of nonlinear damping. Closed-loop stability is then established using Corollary 2 because the observer error is exponentially decaying. This exponential decay condition is satisfied by most observers used in backstepping, including Krener and Isidori [9], Arcak and Kokotović [4, 3], and Praly and Kanellakopoulos [15]. We wish to emphasize that our iISS design is not of "certainty-equivalence" type because, as in the ISS design of [7], the design of the controller makes use of the observer equations.

To make the main features of our design more apparent, we first illustrate it in an example. The general design procedure and its stability proof are given in the next section.

Example 5. For the system

$$\begin{aligned} \dot{x}_1 &= x_2 + x_1^3, \\ \dot{x}_2 &= u + x_2 - x_2^3, \\ y &= x_1, \end{aligned} \tag{45}$$

the problem is to stabilize the origin $x = 0$ by output-feedback. The following observer, designed as in [3], ensures exponential convergence of the estimates $\hat{x}_1$ and $\hat{x}_2$ to the true states:

$$\begin{aligned} \dot{\hat{x}}_1 &= \hat{x}_2 + G_1(y, \hat{x}_1) := \hat{x}_2 + y^3 - 2(\hat{x}_1 - y), \\ \dot{\hat{x}}_2 &= u + G_2(y, \hat{x}_1, \hat{x}_2) := u + \hat{x}_2 - (\hat{x}_2 - 1.5(\hat{x}_1 - y))^3 - 3(\hat{x}_1 - y). \end{aligned} \tag{46}$$

To incorporate this observer in feedback control we employ the observer-based backstepping procedure of Kanellakopoulos, Kokotović, and Morse [7]. Defining the observer error $z_2 := x_2 - \hat{x}_2$ and letting $\chi_1 := x_1$, we rewrite the first equation of (45) as

$$\dot{\chi}_1 = \hat{x}_2 + \chi_1^3 + z_2\,. \tag{47}$$

For $\hat{x}_2$ we design the virtual control law

$$\alpha_1(\chi_1) = -c_1\chi_1 - \chi_1^3\,, \quad c_1 > 0, \tag{48}$$

which results in

$$\dot{\chi}_1 = -c_1\chi_1 + \chi_2 + z_2\,, \tag{49}$$

where $\chi_2 = \hat{x}_2 - \alpha_1(\chi_1)$. Differentiating χ_2 with respect to time, we obtain

$$\dot{\chi}_2 = u + G_2(y, \hat{x}_1, \hat{x}_2) - \frac{\partial \alpha_1}{\partial \chi_1}(\chi_1^3 + \hat{x}_2) - \frac{\partial \alpha_1}{\partial \chi_1} z_2 \tag{50}$$

and note that the control law

$$u = -(c_2 + \delta(\chi_1))\chi_2 - \chi_1 - G_2(y, \hat{x}_1, \hat{x}_2) + \frac{\partial \alpha_1}{\partial \chi_1}(\chi_1^3 + \hat{x}_2) \tag{51}$$

yields

$$\dot{\chi}_2 = -(c_2 + \delta(\chi_1))\chi_2 - \chi_1 - \frac{\partial \alpha_1}{\partial \chi_1} z_2\,. \tag{52}$$

In the ISS design of Kanellakopoulos, Kokotović, and Morse [7], the *nonlinear damping* term is

$$\delta(\chi_1) = \delta_{\text{ISS}}(\chi_1) = \left(\frac{\partial \alpha_1}{\partial \chi_1}\right)^2 = (c_1 + 3\chi_1^2)^2, \tag{53}$$

whose growth in χ_1 is quartic.

To design a "softer" $\delta(\chi_1)$, let us pursue an iISS design with the help of the Lyapunov function $U(\chi_1,\chi_2)=\frac{1}{2}\chi_1^2+\frac{1}{2}\chi_2^2$. From (47) and (50),

$$\dot U=-c_1\chi_1^2+\chi_1z_2-c_2\chi_2^2-\delta(\chi_1)\chi_2^2+\chi_2(c_1+3\chi_1^2)z_2, \tag{54}$$

and hence, using the inequalities

$$3\chi_2\chi_1^2z_2\le\chi_2^2\chi_1^2+\frac{9}{4}\chi_1^2z_2^2,\quad \chi_1z_2\le\frac{c_1}{2}\chi_1^2+\frac{1}{2c_1}z_2^2,\quad c_1\chi_2z_2\le\frac{c_2}{2}\chi_2^2+\frac{c_1^2}{2c_2}z_2^2\,, \tag{55}$$

we can find positive constants k_1, k_2, k_3 such that

$$\dot U\le-k_1U-\delta(\chi_1)\chi_2^2+\chi_1^2\chi_2^2+k_2Uz_2^2+k_3z_2^2\,. \tag{56}$$

Unlike the iISS Lyapunov function (5) in Lemma 1, the inequality (56) contains the product of U and the disturbance z_2^2, which means that $U(\chi)$ cannot be an iISS Lyapunov function. However, the new C^1 function $V(\chi):=\ln(1+U(\chi))$ results in

$$\dot V=\frac{\dot U}{1+U}\,, \tag{57}$$

which means that the product $k_2Uz_2^2$ in (56) is now

$$k_2\frac{U}{1+U}z_2^2\le k_2z_2^2, \tag{58}$$

and hence

$$\dot V\le-k_1\frac{U}{1+U}+(k_2+k_3)z_2^2+(\chi_1^2-\delta(\chi_1))\frac{\chi_2^2}{1+U}\,. \tag{59}$$

Because the first two terms on the right-hand side are as in the iISS Lyapunov function (5), the choice

$$\delta(\chi_1)=\delta_{\mathrm{iISS}}(\chi_1)=\chi_1^2 \tag{60}$$

eliminates the third term and ensures iISS as in Lemma 1. It follows from Corollary 3 that our iISS design ensures GAS of the closed-loop system (45)–(46) because the observer error z is exponentially decaying and the χ-subsystem is affine in z_2.

Unlike the quartic $\delta_{\mathrm{ISS}}(\chi_1)$, the "weak" nonlinear damping term $\delta_{\mathrm{iISS}}(\chi_1)$ is only quadratic. This reduction in the nonlinear growth is more pronounced for higher relative degree systems studied in the next section, which require several steps of observer-based backstepping.

6. Observer-based backstepping with weak nonlinear damping. We now generalize the above design to the system

$$\begin{aligned}
y&=x_1,\\
\dot x_1&=x_2+g_1(x_1),\\
\dot x_2&=x_3+g_2(x_1,x_2),\\
&\cdots\\
\dot x_r&=p(\xi,u)+g_r(x),\\
\dot\xi&=q(\xi,x,u)\,,
\end{aligned} \tag{61}$$

where $x := (x_1, \ldots, x_r)^T$ and the functions $g_1, \ldots, g_r, p$ and q are smooth and vanish when their arguments are zero; that is, the origin $(x, \xi) = (0, 0)$ is an equilibrium when $u = 0$.

We assume the availability of a global observer of the form

$$\begin{aligned} \dot{\hat{x}}_1 &= \hat{x}_2 + G_1(y, \hat{x}_1), \\ \dot{\hat{x}}_2 &= \hat{x}_3 + G_2(y, \hat{x}_1, \hat{x}_2), \\ &\cdots \\ \dot{\hat{x}}_r &= p(\hat{\xi}, u) + G_r(y, \hat{x}), \\ \dot{\hat{\xi}} &= Q(\hat{\xi}, \hat{x}, u, y)\,. \end{aligned} \tag{62}$$

Assumption 1. *The observer* (62) *guarantees exponential convergence of the state estimates to the true states; that is, there exist a constant $k > 0$ and a class-$\mathcal{K}$ function $\gamma(\cdot)$ such that, for every input u and for every initial condition $x(0)$, $\xi(0)$, $\hat{x}(0)$, $\hat{\xi}(0)$, the observer error $z = (x^T, \xi^T)^T - (\hat{x}^T, \hat{\xi}^T)^T$ satisfies*

$$|z(t)| \leq e^{-kt}\gamma(|z(0)|) \tag{63}$$

for all t in the maximal interval of existence $[0, t_f)$ of (61)–(62).

Observer designs satisfying Assumption 1 are being reported at an increasing rate [15, 4, 3]. Our next assumption is that the function $p(\xi, u)$ is invertible in u.

Assumption 2. *There exists a function $\pi(\cdot, \cdot)$ satisfying*

$$v = p(\xi, u) \quad \Leftrightarrow \quad u = \pi(\xi, v). \tag{64}$$

Finally, the zero dynamics of the system (61),

$$\dot{\xi} = q(\xi, 0, \pi(\xi, 0)), \tag{65}$$

satisfy the following robust stability assumption.

Assumption 3. *The zero dynamics* (65)*, perturbed by v_0, v_1, and v_2,*

$$\dot{\xi} = q(\xi, v_0, \pi(\xi - v_1, v_2)), \tag{66}$$

are ISS with input (v_0, v_1, v_2).

The class of systems defined by (61) and Assumptions 1–3 encompasses the one studied in [7] and [10] for observer-based backstepping. We now present our new design, which renders the system (61) iISS with respect to the observer error z.

Step 1. Defining

$$\chi_1 := y \tag{67}$$

and using $x_2 = \hat{x}_2 + z_2$, we rewrite the first equation of (61) as

$$\dot{\chi}_1 = \hat{x}_2 + g_1(\chi_1) + z_2\,. \tag{68}$$

For $\hat{x}_2$, we design the *virtual* control law

$$\alpha_1(\chi_1) = -c_1\chi_1 - g_1(\chi_1), \quad c_1 > 0\,, \tag{69}$$

and obtain

$$\dot{\chi}_1 = -c_1\chi_1 + \chi_2 + z_2, \tag{70}$$

where

$$\chi_2 := \hat{x}_2 - \alpha_1(\chi_1)\,. \tag{71}$$

Step 2. From (62) and (68), χ_2 is governed by

$$\dot{\chi}_2 = \hat{x}_3 + G_2(\chi_1, \hat{x}_1, \hat{x}_2) - \phi_1(\chi_1)(\hat{x}_2 + g_1(\chi_1)) - \phi_1(\chi_1)z_2\,, \tag{72}$$

where

$$\phi_1(\chi_1) := \frac{\partial \alpha_1}{\partial \chi_1}\,. \tag{73}$$

Because $\phi_1(\chi_1)$ is a smooth function, we can rewrite it as

$$\phi_1(\chi_1) = \phi_{10} + \chi_1\Phi_1(\chi_1), \tag{74}$$

where

$$\phi_{10} = \phi_1(0)\,, \quad \Phi_1(\chi_1) = \int_0^1 \left.\frac{\partial \phi_1(X_1)}{\partial X_1}\right|_{X_1 = s\chi_1} ds\,. \tag{75}$$

For $\hat{x}_3$, we design the virtual control law

$$\alpha_2(\chi_1, \hat{x}_1, \hat{x}_2) = -[c_2 + d_2\Phi_1^2(\chi_1)]\chi_2 - \chi_1 - G_2(\chi_1, \hat{x}_1, \hat{x}_2) + \phi_1(\chi_1)(\hat{x}_2 + g_1(\chi_1)), \tag{76}$$

which results in

$$\dot{\chi}_2 = -\chi_1 - [c_2 + d_2\Phi_1^2(\chi_1)]\chi_2 + \chi_3 - [\phi_{10} + \chi_1\Phi_1(\chi_1)]z_2\,, \tag{77}$$

where

$$\chi_3 := \hat{x}_3 - \alpha_2(\chi_1, \hat{x}_1, \hat{x}_2)\,. \tag{78}$$

Step i ($3 \le i \le r$). For

$$\chi_i := \hat{x}_i - \alpha_{i-1}(\chi_1, \hat{x}_1, \ldots, \hat{x}_{i-1})\,, \tag{79}$$

we obtain

$$\begin{aligned}\dot{\chi}_i = {} & \hat{x}_{i+1} + G_i(\chi_1, \hat{x}_1, \ldots, \hat{x}_i) - \phi_{i-1}(\chi_1, \hat{x}_1, \ldots, \hat{x}_{i-1})[\hat{x}_2 + g_1(\chi_1) + z_2] \\ & - \frac{\partial \alpha_{i-1}}{\partial \hat{x}_1}(\hat{x}_2 + G_1(y, \hat{x}_1)) - \cdots - \frac{\partial \alpha_{i-1}}{\partial \hat{x}_{i-1}}(\hat{x}_i + G_{i-1}(y, \ldots, \hat{x}_{i-1})),\end{aligned} \tag{80}$$

where $\hat{x}_{r+1} := p(\hat{\xi}, u)$ and

$$\phi_{i-1}(\chi_1, \hat{x}_1, \ldots, \hat{x}_{i-1}) := \frac{\partial \alpha_{i-1}}{\partial \chi_1}\,. \tag{81}$$

To factor ϕ_{i-1} as in (74) in Step 2, we first rewrite it as a function of $(\chi_1, \ldots, \chi_{i-1}, z_1)$, where $z_1 = x_1 - \hat{x}_1$:

$$\phi_{i-1}(\chi_1, \hat{x}_1, \ldots, \hat{x}_{i-1}) = \tilde{\phi}_{i-1}(\chi_1, \ldots, \chi_{i-1}, z_1)\,. \tag{82}$$

Next, defining $\phi_{i-1,\,0}(z_1) = \tilde{\phi}_{i-1}(0,\ldots,0,z_1)$ and

$$\Phi_{i-1}(\chi_1,\ldots,\chi_{i-1},z_1) = \int_0^1 \left(\frac{\partial\tilde{\phi}_{i-1}(X_1,\ldots,X_{i-1},z_1)}{\partial(X_1,\ldots,X_{i-1})}\right)^T\Bigg|_{(X_1,\ldots,X_{i-1})=s(\chi_1,\ldots,\chi_{i-1})} ds\,, \tag{83}$$

we get

$$\phi_{i-1}(\chi_1,\hat{x}_1,\ldots,\hat{x}_{i-1}) = \phi_{i-1,\,0}(z_1) + [\chi_1\cdots\chi_{i-1}]\,\Phi_{i-1}(\chi_1,\ldots,\chi_{i-1},z_1)\,. \tag{84}$$

The virtual control law for $\hat{x}_{i+1}$ is

$$\begin{aligned}\alpha_i(\chi_1,\hat{x}_1,\ldots,\hat{x}_i) = {}& -\chi_{i-1} - [c_i + d_i\Phi_{i-1}^T\Phi_{i-1}]\chi_i - G_i(\chi_1,\hat{x}_1,\ldots,\hat{x}_i)\\ & + \frac{\partial\alpha_{i-1}}{\partial\hat{x}_1}(\hat{x}_2 + G_1(y,\hat{x}_1)) + \cdots + \frac{\partial\alpha_{i-1}}{\partial\hat{x}_{i-1}}(\hat{x}_i + G_{i-1}(y,\ldots,\hat{x}_{i-1}))\\ & + \phi_{i-1}(\chi_1,\hat{x}_1,\ldots,\hat{x}_{i-1})[\hat{x}_2 + g_1(\chi_1)]\,.\end{aligned} \tag{85}$$

If $i < r$, we define

$$\chi_{i+1} = \hat{x}_{i+1} - \alpha_i(\chi_1,\hat{x}_1,\ldots,\hat{x}_i) \tag{86}$$

and proceed with step $i+1$. The control law obtained by r steps of backstepping is

$$u = \pi(\hat{\xi},\alpha_r(\chi_1,\hat{x}_1,\ldots,\hat{x}_r))\,, \tag{87}$$

where the function $\pi(\cdot,\cdot)$ is as in (64).

The resulting closed-loop system consists of the exponentially converging observer error z driving the (χ,ξ)-subsystem

$$\begin{aligned}\dot{\xi} &= q(\xi, v_0(\chi,z), \pi(\xi - v_1(z), v_2(\chi,z))),\\ \dot{\chi}_1 &= -c_1\chi_1 + \chi_2 + z_2,\\ &\cdots\\ \dot{\chi}_i &= -\chi_{i-1} - [c_i + d_i\Phi_{i-1}^T\Phi_{i-1}]\chi_i + \chi_{i+1} - (\phi_{i-1,0}(z_1) + [\chi_1\cdots\chi_{i-1}]\Phi_{i-1})z_2,\\ &\cdots\\ \dot{\chi}_r &= -\chi_{r-1} - [c_r + d_r\Phi_{r-1}^T\Phi_{r-1}]\chi_r - (\phi_{r-1,0}(z_1) + [\chi_1\cdots\chi_{r-1}]\Phi_{r-1})z_2\,,\end{aligned} \tag{88, 89}$$

where the functions $v_0(\chi,z) = x$, $v_2(\chi,z) = \alpha_r(\chi_1,\hat{x}_1,\ldots,\hat{x}_r)$, and $v_1(z) = \xi - \hat{\xi}$ vanish at $(\chi,z) = (0,0)$.

THEOREM 2. *If Assumptions* 1–3 *hold, then the control law* (87) *guarantees GAS of the closed-loop system* (61)–(62).

Proof. We first prove that the χ-subsystem (89) is iISS with input z. To this end, we note that the function $U = \frac{1}{2}\sum_{i=1}^r \chi_i^2$ satisfies

$$\dot{U} = \left(\sum_{i=1}^r -c_i\chi_i^2 - \chi_i\phi_{i-1,0}(z_1)z_2\right) + \left(\sum_{i=2}^r -d_i\Phi_{i-1}^T\Phi_{i-1}\chi_i^2 - z_2[\chi_1\cdots\chi_{i-1}]\Phi_{i-1}\chi_i\right)\,, \tag{90}$$

where $\phi_{00}(z_1) = -1$ and $\phi_{10}(z_1) = \phi_{10}$ as in (74). Using the inequalities

$$-\chi_i\phi_{i-1,0}(z_1)z_2 \le \frac{c_i}{2}\chi_i^2 + \frac{1}{2c_i}\phi_{i-1,0}^2(z_1)z_2^2, \tag{91}$$

$$-z_2[\chi_1\cdots\chi_{i-1}]\Phi_{i-1}\chi_i \le d_i\Phi_{i-1}^T\Phi_{i-1}\chi_i^2 + \frac{1}{4d_i}[\chi_1\cdots\chi_{i-1}]\,[\chi_1\cdots\chi_{i-1}]^T z_2^2\,, \tag{92}$$

we obtain

$$\dot{U} \le -\sum_{i=1}^{r} \frac{c_i}{2}\chi_i^2 + \sum_{i=1}^{r} \frac{1}{2c_i}\phi_{i-1,0}^2(z_1)z_2^2 + \sum_{i=2}^{r} \frac{1}{4d_i}(\chi_1^2 + \cdots + \chi_{i-1}^2)z_2^2 \tag{93}$$

$$\le -cU + \frac{1}{2c}\left(\sum_{i=1}^{r} \phi_{i-1,0}^2(z_1)\right) z_2^2 + \frac{r}{2d} U z_2^2\,, \tag{94}$$

where $c := \min_{1\le i\le r} c_i$, $d := \min_{2\le i\le r} d_i$. Thus

$$V(\chi) := \ln(1 + U(\chi)) \tag{95}$$

satisfies

$$\dot{V} \le -c\frac{U}{1+U} + \left(\frac{r}{2d} + \frac{1}{2c}\sum_{i=1}^{r} \phi_{i-1,0}^2(z_1)\right) z_2^2\,, \tag{96}$$

which, from Lemma 1, proves that the χ-subsystem (89) with input z is iISS with a class-$\mathcal{K}_o$ gain. Moreover, the observer error z satisfies the exponential decay condition (63) for all $t \in [0, t_f)$. Thus, letting $T < t_f$, we can show from Corollary 2 that there exists a class-$\mathcal{KL}$ function $\beta_1(\cdot,\cdot)$ such that, for all $t \in [0, T]$,

$$|(\chi(t), z(t))| \le \beta_2(|(\chi(0), z(0))|, t)\,. \tag{97}$$

Next, we note from Assumption 3 that the ξ-subsystem (88) is ISS with input (χ, z). In view of Corollary 1, this means that a class-$\mathcal{KL}$ function $\beta_3(\cdot,\cdot)$ exists such that, for all $t \in [0, T]$,

$$|(\xi(t), \chi(t), z(t))| \le \beta_3(|(\xi(0), \chi(0), z(0))|, t)\,. \tag{98}$$

Finally, we note that $|(\xi(T), \chi(T), z(T))|$ is bounded by $\beta_3(|(\xi(0), \chi(0), z(0))|, 0)$, which is independent of T. This means that $t_f = \infty$, and hence (98) holds for all $t \ge 0$, which proves GAS of the closed-loop system (61)–(62). □

7. Proofs of lemmas.

Proof of Lemma 1. Part (i) is proved in [2]. To prove part (ii), we modify [2, Proposition II.5], which is proved for a class-$\mathcal{K}$ function $\theta(\cdot)$, and show that it actually holds with a class-$\mathcal{K}_o$ function $\theta_o(\cdot)$.

LEMMA 5. *The system $\dot{x} = f(x, 0)$ is GAS iff there exist a smooth semiproper*[1] *function $W(x)$, a class-$\mathcal{K}_o$ function $\theta_o(\cdot)$, and a continuous positive definite function $\rho : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ such that*

$$\frac{\partial W}{\partial x} f(x, z) \le -\rho(|x|) + \theta_o(|z|). \tag{99}$$

The proof is given at the end of this section. In passing we emphasize that, in general, Lemma 5 does not hold with a *proper* (radially unbounded) $W(x)$.

To complete the proof of Lemma 1, we define $V_1(x) = V(x) + W(x)$, where $V(x)$ and $W(x)$ are as in (6) and (99), respectively, and obtain

$$\dot{V}_1 \le -\rho(|x|) + \sigma(|z|) + \theta_o(|z|)\,, \tag{100}$$

which is (5) with $\mu(|z|) := \sigma(|z|) + \theta_o(|z|)$. □

[1] A positive definite function $W(x)$ is called *semiproper* if there exist a class-$\mathcal{K}$ function $\pi(\cdot)$ and a radially unbounded positive definite function $W_0(x)$ such that $W(x) = \pi(W_0(x))$. Thus $W(x)$ may not be radially unbounded.

Proof of Lemma 2. Because (7) is iISS with input z, it follows from [2] that there exists an *iISS Lyapunov function* $V(x)$ satisfying

$$L_fV(x) + L_gV(x)z := \frac{\partial V}{\partial x}f(x) + \frac{\partial V}{\partial x}g(x)z \leq -\rho(|x|) + \tilde{\sigma}(|z|) \tag{101}$$

for some positive definite function $\rho(\cdot)$ and some class-$\mathcal{K}$ function $\tilde{\sigma}(\cdot)$. To prove that $\mu(\cdot)$ in (4) is class-$\mathcal{K}_o$, we first show that

$$L_fV(x) + |L_gV(x)| \leq \tilde{\sigma}(1)\,. \tag{102}$$

If $L_gV(x) \neq 0$, then (102) follows by evaluating (101) at $z = \frac{1}{|L_gV(x)|}L_gV(x)^T$. If $L_gV(x) = 0$, then (102) holds because $L_fV(x) \leq -\rho(|x|)$ from (101). Next, we note that $L_fV(x) \leq L_fV(x)|z|$ when $|z| \leq 1$ and obtain

$$|z| \leq 1 \quad \Rightarrow \quad L_fV(x)+L_gV(x)z \leq L_fV(x)|z|+|L_gV|\,|z| = (L_fV+|L_gV|)|z| \leq \tilde{\sigma}(1)|z|\,. \tag{103}$$

Inequalities (101) and (103) imply

$$L_fV(x) + L_gV(x)z \leq \sigma(|z|), \tag{104}$$

where

$$\sigma(|z|) := \begin{cases} \tilde{\sigma}(1)|z| & \text{if} \quad |z| \leq 1, \\ \tilde{\sigma}(|z|) & \text{if} \quad |z| > 1\,. \end{cases} \tag{105}$$

Because $\sigma(\cdot)$ is class-$\mathcal{K}_o$, it follows from Lemma 1 that the system (7) is iISS with a class-$\mathcal{K}_o$ gain $\mu(\cdot)$. □

Proof of Lemma 3. Because of ISS, it follows from [22] that there exists an *ISS Lyapunov function* $V(x)$ satisfying

$$\dot{V} \leq -\rho(|x|) + \sigma(|z|) \tag{106}$$

for some class-$\mathcal{K}_\infty$ functions $\rho(\cdot)$ and $\sigma(\cdot)$. Letting $\mu(\cdot)$ be a class-$\mathcal{K}$ function such that $\mu(s) = \tilde{\mu}(s)$ when $s \in [0,1]$ and $\mu(s) = \sigma(s)$ when $s \geq 2$ and applying the *changing supply functions lemma* [21], we can find another ISS Lyapunov function $\tilde{V}(x)$ satisfying

$$\dot{\tilde{V}} \leq -\tilde{\rho}(|x|) + \mu(|z|) \tag{107}$$

for some class-$\mathcal{K}$ function $\tilde{\rho}(\cdot)$. From Lemma 1, this implies iISS with gain $\mu(\cdot)$. □

Proof of Lemma 4. Because of GAS, there exists a $T^* > 0$ such that $|z(T^*)| \leq \epsilon$, and hence, for all $t \geq T^*$,

$$|z(t)| \leq \alpha\left(e^{-k(t-T^*)}\tilde{\gamma}(\epsilon)\right) = \alpha\left(e^{kT^*}e^{-kt}\tilde{\gamma}(\epsilon)\right). \tag{108}$$

Again, from GAS, there exists a class-$\mathcal{KL}$ function $\beta(\cdot,\cdot)$ such that

$$|z(t)| \leq \alpha(\beta(|z(0)|,t)); \tag{109}$$

thus, for all $t \in [0, T^*]$,

$$|z(t)| \leq \alpha\left((\beta(|z(0)|,t)e^{kt})e^{-kt}\right) \leq \alpha((\beta(|z(0)|,0)e^{kT^*})e^{-kt})\,. \tag{110}$$

Choosing $T^* = T(|z(0)|)$ such that $T(|z(0)|) = 0$ for $|z(0)| \in [0, \epsilon]$ and $T(|z(0)|)$ is a continuous, strictly increasing function for $|z(0)| \geq \epsilon$, we conclude from (9), (108), and (110) that (8) holds with

$$\gamma(s) = \begin{cases} \max\left\{1, \frac{\beta(\epsilon,0)}{\tilde{\gamma}(\epsilon)}\right\} \tilde{\gamma}(s) & \text{if} \quad s \leq \epsilon, \\ \max\{\tilde{\gamma}(\epsilon), \beta(s,0)\} e^{kT(s)} & \text{if} \quad s > \epsilon\,. \end{cases} \tag{111}$$

□

Proof of Lemma 5. The result follows by using Lemma 6 instead of [2, Corollary IV.5] and modifying the proof of [2, Proposition II.5] accordingly. For Lemma 6, we define a function $\sigma_- : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ to be class-$\mathcal{K}_-$ if it is continuous and strictly increasing but, unlike a class-$\mathcal{K}$ function, not necessarily zero at zero.

LEMMA 6. *If $\gamma : \mathbb{R}^2_{\geq 0} \to \mathbb{R}$ is such that $\gamma(\cdot, s)$ is class-$\mathcal{K}$ for each $s \in \mathbb{R}_{\geq 0}$ and $\gamma(r, \cdot)$ is class-$\mathcal{K}_o$ for each $r \in \mathbb{R}_{\geq 0}$, then there exist a class-$\mathcal{K}_-$ function $\sigma_-(\cdot)$ and a class-$\mathcal{K}_o$ function $\sigma_o(\cdot)$, such that*

$$\gamma(r, s) \leq \sigma_-(r)\sigma_o(s)\,. \tag{112}$$

Proof. Because both $\gamma(\cdot, s)$ and $\gamma(r, \cdot)$ are class-$\mathcal{K}$, it follows from [2, Corollary IV.5] that there exists a class-$\mathcal{K}$ function $\sigma_1(\cdot)$ such that

$$\gamma(r, s) \leq \sigma_1(r)\sigma_1(s)\,. \tag{113}$$

Because $\gamma(r, s) = \mathcal{O}(s)$ for all $r \geq 0$, we can find a class-$\mathcal{K}_-$ function $\sigma_2(\cdot)$ such that, for all $s \leq 1$,

$$\gamma(r, s) \leq \sigma_2(r)s\,. \tag{114}$$

The inequalities (113) and (114) imply $\gamma(r, s) \leq \sigma_-(r)\tilde{\sigma}_o(s)$, where $\sigma_-(r) := \max\{\sigma_1(r), \sigma_2(r)\}$ and

$$\tilde{\sigma}_o(s) := \begin{cases} s, & s \leq 1, \\ \sigma_1(s), & s > 1\,. \end{cases} \tag{115}$$

Thus (112) follows by finding a continuous upper-bound $\sigma_o(\cdot)$ on $\tilde{\sigma}_o(\cdot)$. □

8. Conclusion. We have studied the stability of nonlinear cascades and showed that a trade-off exists between slower convergence for the driving subsystem and steeper iISS gain for the driven subsystem. This approach unifies several results in the literature, obtained by restricting the speed of convergence of the driving subsystem and the nonlinear growth of the driven subsystem. We have studied the connection between these growth conditions and the iISS gain and have proved that our iISS gain restriction leads to a less restrictive condition than the linear growth assumption common in the literature. The cascade result has been used to develop a new observer-based backstepping design which reduces the growth of nonlinear damping terms. It would be of interest to extend our cascade result to feedback interconnections, where small-gain formulations of iISS can be pursued.

REFERENCES

[1] O. M. AAMO, M. ARCAK, T. I. FOSSEN, AND P. V. KOKOTOVIĆ, *Global output tracking control of a class of Euler-Lagrange systems with monotonic nonlinearities in the velocities*, Internat. J. Control, 74 (2001), pp. 649–658.

[2] D. Angeli, E. D. Sontag, and Y. Wang, *A characterization of integral input-to-state stability*, IEEE Trans. Automat. Control, 45 (2000), pp. 1082–1097.
[3] M. Arcak and P. Kokotović, *Nonlinear observers: A circle criterion design and robustness analysis*, Automatica J. IFAC, 37 (2001), pp. 1923–1930.
[4] M. Arcak and P. V. Kokotović, *Observer-based stabilization of systems with monotonic nonlinearities*, Asian J. Control, 1 (1999), pp. 42–48.
[5] L. Grüne, E. D. Sontag, and F. R. Wirth, *Asymptotic stability equals exponential stability, and ISS equals finite energy gain—if you twist your eyes*, Systems Control Lett., 38 (1999), pp. 127–134.
[6] M. Janković, R. Sepulchre, and P. V. Kokotović, *Constructive Lyapunov stabilization of nonlinear cascade systems*, IEEE Trans. Automat. Control, 41 (1996), pp. 1723–1736.
[7] I. Kanellakopoulos, P. V. Kokotović, and A. S. Morse, *A toolkit for nonlinear feedback design*, Systems Control Lett., 18 (1992), pp. 83–92.
[8] P. V. Kokotović and H. J. Sussmann, *A positive real condition for global stabilization of nonlinear systems*, Systems Control Lett., 19 (1989), pp. 177–185.
[9] A. J. Krener and A. Isidori, *Linearization by output injection and nonlinear observers*, Systems Control Lett., 3 (1983), pp. 47–52.
[10] M. Krstić, I. Kanellakopoulos, and P. Kokotović, *Nonlinear and Adaptive Control Design*, John Wiley and Sons, New York, 1995.
[11] F. Mazenc and L. Praly, *Adding integrations, saturated controls and stabilization for feedforward systems*, IEEE Trans. Automat. Control, 41 (1996), pp. 1559–1578.
[12] E. Panteley and A. Loría, *On global uniform asymptotic stability of nonlinear time-varying systems in cascade*, Systems Control Lett., 33 (1998), pp. 131–138.
[13] E. Panteley and A. Loría, *Growth rate conditions for uniform asymptotic stability of cascaded time-varying systems*, Automatica J. IFAC, 37 (2001), pp. 453–460.
[14] L. Praly and Z.-P. Jiang, *Stabilization by output-feedback for systems with ISS inverse dynamics*, Systems Control Lett., 21 (1993), pp. 19–33.
[15] L. Praly and I. Kanellakopoulos, *Output-feedback asymptotic stabilization for triangular systems linear in the unmeasured state components*, in Proceedings of the 39th IEEE Conference on Decision and Control, Sydney, Australia, 2000, pp. 2466–2471.
[16] P. Seibert and R. Suarez, *Global stabilization of nonlinear cascade systems*, Systems Control Lett., 14 (1990), pp. 347–352.
[17] R. Sepulchre, M. Janković, and P. Kokotović, *Constructive Nonlinear Control*, Springer-Verlag, New York, 1997.
[18] E. D. Sontag, *Remarks on stabilization and input-to-state stability*, in Proceedings of the 28th IEEE Conference on Decision and Control, Tampa, FL, 1989, pp. 1376–1378.
[19] E. D. Sontag, *Smooth stabilization implies coprime factorization*, IEEE Trans. Automat. Control, 34 (1989), pp. 435–443.
[20] E. D. Sontag, *Comments on integral variants of ISS*, Systems Control Lett., 34 (1998), pp. 93–100.
[21] E. D. Sontag and A. Teel, *Changing supply functions in input/state stable systems*, IEEE Trans. Automat. Control, 40 (1995), pp. 1476–1478.
[22] E. D. Sontag and Y. Wang, *On characterizations of the input-to-state-stability property*, Systems Control Lett., 24 (1995), pp. 351–359.
[23] H. J. Sussmann and P. V. Kokotović, *The peaking phenomenon and the global stabilization of nonlinear systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 424–439.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 6, pp. 1905–1923

RIESZ BASIS PROPERTY AND EXPONENTIAL STABILITY OF CONTROLLED EULER–BERNOULLI BEAM EQUATIONS WITH VARIABLE COEFFICIENTS*

BAO-ZHU GUO†

Abstract. This paper studies the basis property and the stability of a distributed system described by a nonuniform Euler–Bernoulli beam equation under linear boundary feedback control. It is shown that there is a sequence of generalized eigenfunctions of the system, which forms a Riesz basis for the state Hilbert space. The asymptotic distribution of eigenvalues, the spectrum-determined growth condition, and the exponential stability are concluded. The results are applied to a nonuniform beam equation with viscous damping of variable coefficient as a generalization of existing results for the uniform beam.

Key words. beam equation, variable coefficients, asymptotic analysis, Riesz basis, stability

AMS subject classifications. 93C20, 93D15, 35B35, 35P10

PII. S0363012900372519

1. Introduction.

The Riesz basis property, meaning that the generalized eigenvectors of the system form an unconditional basis for the state Hilbert space, is one of the fundamental properties of a linear vibrating system. The establishment of the basis property will naturally lead to solutions to such problems as the spectrum-determined growth condition and the exponential stability for infinite dimensional systems. Unfortunately, verification of the Riesz basis generation is challenging even for extensively studied systems such as Euler–Bernoulli beam equations. Recently, a new approach has been suggested [1] to obtain a complete solution to the basis property of the following uniform Euler–Bernoulli beam equation under linear boundary feedback control:

$$\begin{cases} y_{tt}(x,t) + y_{xxxx}(x,t) = 0, & 0 < x < 1, t > 0, \\ y(0,t) = y_x(0,t) = 0, & \\ y_{xx}(1,t) = -k_1 y_{xt}(1,t), & k_1 \geq 0, \\ y_{xxx}(1,t) = k_2 y_t(1,t), & k_2 \geq 0. \end{cases} \tag{1}$$

In this paper, we shall develop parallel results for the same system with variable coefficients. What makes it unique compared to the case of constant coefficients is that with variable coefficients both the characteristic equation and the analytic expression of the eigenfunctions have no explicit formulae. The asymptotic technique appears to be essential for the study.

There are two steps usually found in the study of linear systems with variable coefficients. The first is to transform the "dominant term" of the system under study into a uniform "dominant equation" by space scaling and state transformation where no variable coefficient is involved any longer, while the second is to approximate the eigenfunctions of the system by those of the uniform "dominant equation." This

*Received by the editors May 19, 2000; accepted for publication (in revised form) September 25, 2001; published electronically March 5, 2002. This research was supported by the National Key Project of China and the National Natural Science Foundation of China.
http://www.siam.org/journals/sicon/40-6/37251.html
†Institute of Systems Science, Academy of Mathematics and System Sciences, Academia Sinica, Beijing 100080, China (bzguo@iss03.iss.ac.cn).

fundamental idea comes essentially from Birkhoff's works on asymptotic estimation of the eigenpairs of the linear differential operators with generalized homogeneous boundary conditions done in the beginning of the last century [5]. A comprehensive review can be found in [4]. This approach has been used in dealing with the beam equations with low order perturbation of variable coefficients (see [6], [7], and [8]). A similar adoption can also be found in the study of the string equations with variable coefficients, for which we refer to [9], [10], and [11] as well as the references therein.

By considering a sequence of eigenfunctions rather than whole sequences in the state Hilbert space, the author recently presented a corollary of Bari's theorem on the Riesz basis property [1]. The result greatly simplifies the procedure in establishing the Riesz basis property for systems described by discrete operators in a Hilbert space since the result eliminates the requirement of estimation of low eigenfunctions, which is rather difficult by other methods found in all previous papers [10], [11], [12].

Following the approach used in [1], together with the asymptotic analysis, this paper presents the Riesz basis property for the Euler–Bernoulli beam equation with variable coefficients. Other major contributions include the exponential stability and asymptotic behavior of the systems under boundary feedback control.

In the next section, we shall present the main results of the paper. The proof of the results and some remarks are given in section 3.

2. Main results. Consider the following nonuniform Euler–Bernoulli beam equation with linear boundary feedback control:

$$\begin{cases} \rho(x)y_{tt}(x,t) + (EI(x)y_{xx}(x,t))_{xx} = 0, & 0 < x < 1, t > 0, \\ y(0,t) = y_x(0,t) = y_{xx}(1,t) = 0, \\ (EI(x)y_{xx})_x(1,t) = ky_t(1,t), \end{cases} \tag{2}$$

where x stands for the position and t the time. $EI(x)$ is the flexural rigidity of the beam, and $\rho(x)$ is the mass density at x. $k \geq 0$ is a constant feedback gain. Unlike system (1), here we impose only one end feedback control for simplicity of computation because, from Theorem 2.5 below, it is sufficient for the exponential stabilization of the system. Moreover, it does not make much difference from the methodology point of view. Actually, the analysis in this paper can be used to similarly treat the boundary conditions of (1) along the same lines as the analysis in [1].

The total energy of system (2) is

$$E(t) = \frac{1}{2}\int_0^1 [\rho(x)y_t^2(x,t) + EI(x)y_{xx}^2(x,t)]dx.$$

Formally,

$$\frac{dE(t)}{dt} = -ky_t^2(x,t) \leq 0.$$

That is, system (2) is dissipative. Throughout this paper, we always assume that

$$\rho(x), EI(x) \in C^4[0,1], EI, \quad \rho > 0. \tag{3}$$

System (2) will be considered in the energy Hilbert space $\mathbf{H} = H_E^2(0,1) \times L^2(0,1)$, $H_E^2(0,1) = \{f \in H^2(0,1) | f(0) = f'(0) = 0\}$, in which the inner product induced norm is defined by

$$\|(f,g)\|_{\mathbf{H}}^2 = \int_0^1 [\rho(x)|g(x)|^2 + EI(x)|f''(x)|^2]dx \qquad \forall (f,g) \in \mathbf{H}. \tag{4}$$

Define operator $\mathcal{A} : D(\mathcal{A})(\subset \mathbf{H}) \to \mathbf{H}$ as

$$\begin{cases} \mathcal{A}(f,g) = (g, -\frac{1}{\rho(x)}(EI(x)f''(x))''), \\ D(\mathcal{A}) = \{(f,g) \in (H_E^2 \cap H^4) \times H_E^2 | f''(1) = 0, (EIf'')'(1) = kg(1)\}. \end{cases} \tag{5}$$

With the operator $\mathcal{A}$ at hand, we can write (2) into an evolutionary equation in $\mathbf{H}$:

$$\frac{d}{dt}Y(t) = \mathcal{A}Y(t), Y(t) = (y(\cdot,t), y_t(\cdot,t)). \tag{6}$$

When we talk about system (2) later, we mean its abstract formulation (6). We are concerned with the Riesz basis property of (6) in $\mathbf{H}$; that is, we want to know whether the generalized eigenfunctions of $\mathcal{A}$ form an unconditional basis for $\mathbf{H}$. To do this, we need the following spectral property of $\mathcal{A}$.

LEMMA 2.1. *Let $\mathcal{A}$ be defined by (5). Then $\mathcal{A}^{-1}$ exists and is compact on $\mathbf{H}$. Hence $\sigma(\mathcal{A})$, the spectrum of $\mathcal{A}$, consists only of isolated eigenvalues, which distribute in conjugate pairs on the complex plane. Moreover, the eigenfunction corresponding to $\lambda \in \sigma(\mathcal{A})$ is of the form $(\lambda^{-1}\phi, \phi)$, where $\phi \neq 0$ satisfies*

$$\begin{cases} \lambda^2\rho(x)\phi(x) + (EI(x)\phi''(x))'' = 0, \qquad 0 < x < 1, \\ \phi(0) = \phi'(0) = \phi''(1) = 0, \\ (EI(x)\phi'')'(1) = \lambda k\phi(1). \end{cases} \tag{7}$$

To verify the basis property, we need the asymptotic properties of both eigenvalues and eigenfunctions, which are stated as the following propositions.

PROPOSITION 2.2. *Let $\mathcal{A}$ be defined by (5). Then the eigenvalues $\{\lambda_n, \overline{\lambda_n}\}$ of $\mathcal{A}$ have the following asymptotic property:*

$$\lambda_n = \frac{\rho_n^2}{h^2}, h = \int_0^1 \left(\frac{\rho(\tau)}{EI(\tau)}\right)^{1/4} d\tau, \rho_n = \frac{1}{\sqrt{2}}\left(n + \frac{1}{2}\right)\pi(1+i) + \mathcal{O}(n^{-1}) \text{ as } n \to \infty, \tag{8}$$

where n is a large positive integer and $\overline{\lambda_n}$ denotes the complex conjugate of λ_n. Moreover, λ_n is geometrically simple when n is large enough.

PROPOSITION 2.3. *Let λ_n be defined as in Proposition 2.2. Then there is a solution ϕ_n to (7) corresponding to λ_n having the following asymptotic expansion:*

$$\begin{aligned} &-\frac{\sqrt{2}}{4}(1+i)e^{\frac{1}{4}\int_0^z a(\tau)d\tau}\phi_n(x) = \sin(n+\pi/2)z - \cos(n+\pi/2)z \\ &+e^{-(n+1/2)\pi z} + (-1)^n e^{-(n+1/2)\pi(1-z)} + \mathcal{O}(n^{-1}), \end{aligned} \tag{9}$$

$$\begin{aligned} &-\frac{\sqrt{2}}{4}(1+i)e^{\frac{1}{4}\int_0^z a(\tau)d\tau}\lambda_n^{-1}\phi_n''(x) = i\left(\frac{\rho(x)}{EI(x)}\right)^{1/2}[\cos(n+\pi/2)z - \sin(n+\pi/2)z \\ &+e^{-(n+1/2)\pi z} + (-1)^n e^{-(n+1/2)\pi(1-z)}] + \mathcal{O}(n^{-1}), \end{aligned} \tag{10}$$

where $z = z(x)$ and $a(z)$ are defined by

$$\begin{cases} z = z(x) = \dfrac{1}{h}\displaystyle\int_0^x \left(\frac{\rho(\tau)}{EI(\tau)}\right)^{1/4} d\tau, h = \int_0^1 \left(\frac{\rho(\tau)}{EI(\tau)}\right)^{1/4} d\tau, \\ a(z) = \dfrac{3h}{2}\left(\dfrac{\rho(x)}{EI(x)}\right)^{-5/4} \dfrac{d}{dx}\left(\dfrac{\rho(x)}{EI(x)}\right) + h\dfrac{2EI'(x)}{EI(x)}\left(\dfrac{\rho(x)}{EI(x)}\right)^{-1/4}. \end{cases} \tag{11}$$

The main result is the following basis property for system (2).

THEOREM 2.4. *Let $\mathcal{A}$ be defined by* (5). *Then the following hold.*

(i) *There is a sequence of generalized eigenfunctions of $\mathcal{A}$ which forms a Riesz basis for the state space* $\mathbf{H}$.

(ii) *The eigenvalues $\{\lambda_n, \overline{\lambda_n}\}$ of $\mathcal{A}$ have the asymptotic expansion* (8).

(iii) *All $\lambda \in \sigma(\mathcal{A})$ with sufficiently large modulus are algebraically simple.*

Therefore, $\mathcal{A}$ generates a C_0-group, and, for the semigroup $e^{\mathcal{A}t}$ generated by $\mathcal{A}$, the spectrum-determined growth condition holds: $\omega(\mathcal{A}) = S(\mathcal{A})$, where $\omega(\mathcal{A}) = \lim_{t\to\infty} \frac{1}{t}\|e^{\mathcal{A}t}\|$ is the growth order of $e^{\mathcal{A}t}$ and $S(\mathcal{A}) = \sup\{\mathrm{Re}\lambda | \lambda \in \sigma(\mathcal{A})\}$ is the spectral bound of $\mathcal{A}$.

Remark 1. From Theorem 2.4 (iii), (9) and (10) are asymptotic expansions for all generalized eigenfunctions of $\mathcal{A}$.

Theorem 2.4 is the fundamental property of system (2). Many other important properties of system (2) can be concluded from Theorem 2.4. The exponential stability stated below is one such important property that has been studied extensively in the past two decades.

THEOREM 2.5. *System* (2) *is exponentially stable for any $k > 0$. That is, there are constants $M, \omega > 0$ such that the energy $E(t)$ of system* (2) *satisfies*

$$E(t) = \frac{1}{2}\int_0^1 [\rho(x)y_t^2(x,t) + EI(x)y_{xx}^2(x,t)]dx \le Me^{-\omega t}E(0) \qquad \forall t \ge 0,$$

for any initial condition $(y(x,0), y_t(x,0)) \in \mathbf{H}$.

Theorem 2.4 will also be applied to the following beam equation with variable viscous damping:

$$\begin{cases} \rho(x)y_{tt}(x,t) + b(x)y_t(x,t) + (EI(x)y_{xx}(x,t))_{xx} = 0, & 0 < x < 1, t > 0, \\ y(0,t) = y_x(0,t) = y_{xx}(1,t) = 0, \\ (EI(x)y_{xx})_x(1,t) = ky_t(1,t). \end{cases} \tag{12}$$

The uniform case of $EI = \rho = \text{const}, b \in C[0,1], k = 0$ was discussed in [8]. System (12) can be written as

$$\frac{d}{dt}Y(t) = (\mathcal{A} + \mathcal{B})Y(t), Y(t) = (y(\cdot,t), y_t(\cdot,t)), \tag{13}$$

where $\mathcal{A}$ is defined by (5) and $\mathcal{B}$ is a linear bounded operator on $\mathbf{H}$:

$$\mathcal{B}(f,g) = (0, -b\cdot g). \tag{14}$$

Equation (13) can be put into the generic framework of discrete-type operators perturbed by the linear bounded operator in the Hilbert spaces. First, we introduce the following definition.

DEFINITION 2.6. *A linear operator A in a Hilbert space H is called discrete-type, or $[D]$-class for short, if there are Riesz basis $\{\phi_n\}_1^\infty$ of H, complex series $\{\lambda_n\}_1^\infty$, and an integer $N > 0$ such that*

(i) $\lim_{n\to\infty} |\lambda_n| = \infty, \lambda_n \ne \lambda_m$ *as* $n, m > N$;

(ii) $A\phi_n = \lambda_n\phi_n,\ n > N$;

(iii) $A[\phi_1, \phi_2, \ldots, \phi_N] \subset [\phi_1, \phi_2, \ldots, \phi_N]$, *and A has spectrum $\{\lambda_i\}_1^N$ in $[\phi_1, \phi_2, \ldots, \phi_N]$, where $[\phi_1, \phi_2, \ldots, \phi_N]$ is the linear subspace spanned by $\{\phi_i\}_1^N$.*

Remark 2. Theorem 2.4 shows that $\mathcal{A}$ is of $[D]$-class.

It is known that any $[D]$-class operator A must be a discrete operator [14], and for the C_0-semigroup e^{At} generated by A, the spectrum-determined growth condition holds. The following basic result can be concluded from the proof of a more general result in [14] (see also [15], [16, section V.4]). A short proof will be given in the next section.

THEOREM 2.7. *Suppose that A is of $[D]$-class satisfying conditions of Definition* 2.6 *in a Hilbert space H. Let $d_n = \min_{n\neq m} |\lambda_n - \lambda_m|$. If*

$$\sum_{n>N}^{\infty} d_n^{-2} < \infty, \tag{15}$$

then, for any linear bounded operator B on H, there are constants $C, L > 0$, an integer $M > 0$, and eigenpairs $\{\mu_n, \psi_n\}_M^\infty$ of $A + B$ such that

(i) $|\mu_n - \lambda_n| \leq C$ *for all* $n \geq M$.

(ii) $\|\psi_n - \phi_n\| \leq L d_n^{-1}, n \geq M$. *Hence* $\sum_{n=M}^{\infty} \|\psi_n - \phi_n\|^2 < \infty$.

We can now consider (12). By Remark 1, $\mathcal{A}$ is of $[D]$-class. And the spectral separation of $\mathcal{A}$ satisfies $d_n^{-1} = \mathcal{O}(n^{-1})$. In Remark 4 of the next section, we shall show that d_n is never vanishing. Hence Theorem 2.7 can be applied to $(A, B) = (\mathcal{A}, \mathcal{B})$ to get the following parallel result of Theorem 2.4 for system (12).

THEOREM 2.8. *Suppose $EI, \rho \in C^4[0,1], EI, \rho > 0, b \in C[0,1]$. Then the following hold.*

(i) $\mathcal{A} + \mathcal{B}$ *is of $[D]$-class.*

(ii) *The eigenvalues $\{\mu_n, \overline{\mu_n}\}$ of $\mathcal{A} + \mathcal{B}$ have the asymptotic expansion*

$$\mu_n = \lambda_n + \mathcal{O}(1) \text{ as } n \to \infty, \tag{16}$$

where λ_n is defined by (8).

(iii) *The corresponding eigenfunctions $\{(\mu_n^{-1}\psi_n, \psi_n)\} \cup \{$ their conjugates $\}$ of $\mathcal{A}+\mathcal{B}$ have the asymptotic expansion*

$$(\mu_n^{-1}\psi_n, \psi_n) = (\lambda_n^{-1}\phi_n, \phi_n) + \epsilon_n \text{ as } n \to \infty, \tag{17}$$

where ϕ_n is defined by (9) *and*

$$\|\epsilon_n\|_{\mathbf{H}} = \mathcal{O}(n^{-1}). \tag{18}$$

The following result can be viewed as a consequence of Theorem 2.8.

COROLLARY 2.9. *Let $\{\mu_n\}$ be the eigenvalues of $\mathcal{A}+\mathcal{B}$ determined in Theorem* 2.8. *Then*

$$\lim_{n\to\infty} \mathrm{Re}\mu_n = -\frac{1}{2} \frac{\int_0^1 b(x) e^{-\frac{1}{2}\int_0^z a(\tau)d\tau} dx + 4k e^{-\frac{1}{2}\int_0^1 a(\tau)d\tau}}{\int_0^1 \rho(x) e^{-\frac{1}{2}\int_0^z a(\tau)d\tau} dx}, \tag{19}$$

where $z = z(x), a(z)$ are defined in (11).

Corollary 2.9 concludes some existing results for system (12). We give several examples below.

Example 1. Suppose that $\rho = 1, k = 0$, and EI is a constant. Then $a = 0$. Equation (19) becomes

$$\lim_{n\to\infty} \mathrm{Re}\mu_n = -\frac{1}{2}\int_0^1 b(x)dx, \tag{20}$$

which is a strengthened conjecture

$$\lim_{n\to\infty}\frac{\sum_{j\le n}\mathrm{Re}\mu_j}{n}=-\frac{1}{2}\int_0^1 b(x)dx \tag{21}$$

made in [13] for the same system with hinged boundary conditions and resolved later in [6] under the assumption that $b(x)\ge 0$. However, we do not impose any assumption on the symbol of b.

Example 2. Suppose that $\rho = EI = 1, k = 0$. Then (12) becomes

$$\begin{cases} y_{tt}(x,t)+b(x)y_t(x,t)+y_{xxxx}(x,t)=0, & 0<x<1, t>0,\\ y(0,t)=y_x(0,t)=y_{xx}(1,t)=y_{xxx}(1,t)=0, \end{cases} \tag{22}$$

which is just the system studied in [8]. Equation (20) holds for this system. However, our result shows that the main hypothesis (1.3) of [8] is nothing but $\int_0^1 b(x)dx>0$. Moreover, from our discussion, Theorem 2.8 is sufficient to derive (20). This is because spectral analysis for system (22) with $b=0$ is quite simple and does not necessarily need to rely on Theorem 2.4 [1].

Example 3. Suppose that $b(x)\ge 0$ for $x\in[0,1]$, and $b(x)>b_0>0$ for all x in some subset $(a,b)\subset[0,1]$ in (22). The system is then exponentially stable. When $k=0, b(x)\ge 0$ for $x\in[0,1]$ and $b(x)>b_0>0$ for all x in some subset $(a,b)\subset[0,1]$, system (12) is also exponentially stable. However, the method used in [18] appears to be unavailable for this case. We will give a short interpretation for Example 3 in section 3.

Finally, we present a high order approximation of the eigenvalues of system (12).

PROPOSITION 2.10. *Suppose* (3) *and*

$$b(x)\in C^1[0,1],\quad \int_0^1 b(x)e^{-\frac{1}{2}\int_0^z a(\tau)d\tau}dx+4ke^{-\frac{1}{2}\int_0^1 a(\tau)d\tau}>0. \tag{23}$$

Then the eigenvalues $\{\mu_n,\overline{\mu_n}\}$ *of* $\mathcal{A}+\mathcal{B}$ *have the asymptotic expansion*

$$\mu_n=-\frac{2\tilde{k}}{h^2}+i\left[(n+1/2)\frac{\pi}{h}\right]^2-\frac{i}{2h^2}\int_0^1 a_1(\tau)d\tau-\frac{1}{2h^2}\int_0^1\tilde{b}(\tau)d\tau+\mathcal{O}(n^{-1}), \tag{24}$$

where $\tilde{b}(z), a_1(z)$, *and* $\tilde{k}$ *are given by*

$$\tilde{k}=\frac{kh}{EI(1)}\left(\frac{\rho(1)}{EI(1)}\right)^{-3/4}, \tag{25}$$

$$\tilde{b}(z)=\frac{b(x)}{\rho(x)}, z=\frac{1}{h}\int_0^x\left(\frac{\rho(\tau)}{EI(\tau)}\right)^{1/4}d\tau, \tag{26}$$

$$a_1(z)=-\frac{3}{2}a'(z)-\frac{9}{16}a^2(z)-\frac{1}{4}a(z). \tag{27}$$

3. Proof of main results.

Proof of Lemma 2.1. A direct calculation shows that

$$\mathcal{A}^{-1}(f,g)=(\phi,\psi)\text{ for any }(f,g)\in \mathrm{H},$$

where

$$\begin{cases} \psi & = f, \\ \phi(x) & = kf(1)\displaystyle\int_0^x (x-\tau)\frac{\tau-1}{EI(\tau)}d\tau + \int_0^x \rho(\tau)g(\tau)d\tau \int_\tau^x d\vartheta \int_\tau^\vartheta \frac{s-\tau}{EI(s)}ds. \end{cases}$$

The compactness follows from the Sobolev embedding theorem. Other conclusions are obvious, and the details are omitted. □

In order to study the asymptotic behavior of the solution of (7), we rewrite (7) in a standard form of the eigenproblem of a linear differential operator with generalized homogeneous boundary conditions:

$$\begin{cases} \phi^{(4)}(x) + \dfrac{2EI'(x)}{EI(x)}\phi'''(x) + \dfrac{EI''(x)}{EI(x)}\phi''(x) + \lambda^2\dfrac{\rho(x)}{EI(x)}\phi(x) = 0, \\ \phi(0) = \phi'(0) = \phi''(1) = 0, \\ \phi'''(1) = \lambda\dfrac{k}{EI(1)}\phi(1). \end{cases} \tag{28}$$

Two basic transformations are essential. First, the "dominant term," $\phi^{(4)}(x) + \lambda^2\rho(x)/EI(x)\phi(x)$ of (28), is transformed to become a uniform form by space scaling. In fact, set

$$\phi(x) = f(z), z = z(x) = \frac{1}{h}\int_0^x \left(\frac{\rho(\tau)}{EI(\tau)}\right)^{1/4} d\tau, h = \int_0^1 \left(\frac{\rho(\tau)}{EI(\tau)}\right)^{1/4} d\tau. \tag{29}$$

Then f satisfies

$$\begin{cases} f^{(4)}(z) + a(z)f'''(z) + b_f(z)f''(z) + c(z)f'(z) + \lambda^2h^4f(z) = 0, \\ f(0) = f'(0) = 0, \\ f''(1) + a_0f'(1) = 0, \\ f'''(1) = b_0f'(1) + \lambda\dfrac{kh^3}{EI(1)}\left(\dfrac{\rho(1)}{EI(1)}\right)^{-3/4} f(1), \end{cases} \tag{30}$$

where a_0 and b_0 are constants depending on $h, \rho^{(i)}(1), EI^{(i)}(1), i = 0, 1, 2, b_f(z)$ and $c(z)$ are the smooth functions of $h, \rho^{(i)}(x), EI^{(i)}(x), i = 0, 1, 2, 3$ through $z = z(x)$ defined by (11), and $a(z)$ is the function given by (11).

Second, in order to cancel the term $a(z)f'''$ in (30) as was done in [4], we make the invertible state transformation

$$f(z) = e^{-\frac{1}{4}\int_0^z a(\tau)d\tau}g(z). \tag{31}$$

Then g satisfies

$$\begin{cases} g^{(4)}(z) + a_1(z)g''(z) + a_2(z)g'(z) + a_3(z)g(z) + \lambda^2h^4g(z) = 0, \\ g(0) = g'(0) = 0, \\ g''(1) = a_{11}g'(1) + a_{12}g(1), \\ g'''(1) = a_{21}g'(1) + \left[\lambda\dfrac{kh^3}{EI(1)}\left(\dfrac{\rho(1)}{EI(1)}\right)^{-3/4} + a_{22}\right] g(1), \end{cases} \tag{32}$$

where $a_{ij}, i, j = 1, 2$ are some real constants depending on $h, \rho^{(i)}(1), EI^{(i)}(1), i = 0, 1, 2, a_2(z)$ and $a_3(z)$ are the smooth functions of $h, \rho^{(i)}(x), EI^{(i)}(x), i = 0, 1, 2, 3$ through $z = z(x)$ defined by (29), and $a_1(z)$ is given by (27).

It can be seen that (7) and (32) are equivalent. Our next task is to use the eigenpairs of the uniform "dominant term," $g^{(4)}(z)+\lambda^2 h^4 g(z)=0$ of (32), to approximate those of the whole system. Note that when $k=0$, (32) is in the standard form of a linear differential operator with generalized homogeneous boundary conditions, which was studied in [4] in greater detail.

Now we proceed as in section 4, Chapter 2 of [4] to estimate asymptotically the solutions to (32). Since system (2) is dissipative, all eigenvalues are located on the left half complex plane. Due to the conjugate property of the eigenvalues, we may consider only those λ with $\pi/2 \le \arg\lambda \le \pi$.

Let $\lambda=\rho^2/h^2$. Then, as $\pi/2 \le \arg\lambda \le \pi$,

$$\pi/4 \le \arg\rho \le \pi/2. \tag{33}$$

Now set

$$\begin{cases} \omega_1 = e^{3/4\pi i}, \omega_2 = e^{\pi/4 i}, \omega_3 = -\omega_2, \omega_4 = -\omega_1, \\ S = \left\{\rho | \frac{\pi}{4} \le \arg\rho \le \frac{\pi}{2}\right\}. \end{cases} \tag{34}$$

In what follows, ρ is always assumed to be in S. Note that

$$\mathrm{Re}(\rho\omega_1) \le \mathrm{Re}(\rho\omega_2) \le \mathrm{Re}(\rho\omega_3) \le \mathrm{Re}(\rho\omega_4) \qquad \forall \rho \in S. \tag{35}$$

The following important facts are used frequently in what follows.

$$\begin{cases} \mathrm{Re}(\rho\omega_1) = -|\rho|\sin(\arg\rho + \frac{\pi}{4}) \le -\sqrt{2}/2|\rho| < 0, \\ \mathrm{Re}(\rho\omega_2) = |\rho|\cos(\arg\rho + \frac{\pi}{4}) \le 0. \end{cases} \tag{36}$$

Lemma 3.1 comes from Theorem 2.4 in section 4, Chapter 2 of [4].

LEMMA 3.1. *For $|\rho|$ large enough, $\rho \in S$, there are four linearly independent solutions $g_k(z), k=1,2,3,4$ to*

$$g^{(4)}(z)+a_1(z)g''(z)+a_2(z)g'(z)+a_3(z)g(z)+\rho^4 g(z)=0,$$

such that

$$\begin{cases} g_k(z) = e^{\rho\omega_k z}[1+\mathcal{O}(\frac{1}{\rho})], \\ g_k'(z) = \rho\omega_k e^{\rho\omega_k z}[1+\mathcal{O}(\frac{1}{\rho})], \\ g_k''(z) = (\rho\omega_k)^2 e^{\rho\omega_k z}[1+\mathcal{O}(\frac{1}{\rho})], \\ g_k'''(z) = (\rho\omega_k)^3 e^{\rho\omega_k z}[1+\mathcal{O}(\frac{1}{\rho})]. \end{cases} \tag{37}$$

With these preparations, we come to the proof of Proposition 2.2.

Proof of Proposition 2.2. Let $g(z)$ be a solution of (32). There are constants $c_i, i=1,2,3,4$, such that

$$g(z) = c_1 g_1(z) + c_2 g_2(z) + c_3 g_3(z) + c_4 g_4(z), \tag{38}$$

where $g_k(z), k=1,2,3,4$ are defined by (37). By boundary conditions, $c_i, i=1,2,3,4$ are solutions to the following system of linear algebraic equations:

$$\begin{cases} c_1 g_1(0) + c_2 g_2(0) + c_3 g_3(0) + c_4 g_4(0) = 0, \\ c_1 g_1'(0) + c_2 g_2'(0) + c_3 g_3'(0) + c_4 g_4'(0) = 0, \\ \left[g_1''(1) - a_{11}g_1'(1) - a_{12}g_1(1)\right]c_1 + \left[g_2''(1) - a_{11}g_2'(1) - a_{12}g_2(1)\right]c_2 \\ +[g_3''(1) - a_{11}g_3'(1) - a_{12}g_3(1)]c_3 + [g_4''(1) - a_{11}g_4'(1) - a_{12}g_4(1)]c_4 = 0, \\ \left[g_1'''(1) - a_{21}g_1'(1) - a_{22}g_1(1) - \tilde{k}\rho^2 g_1(1)\right] c_1 + [g_2'''(1) - a_{21}g_2'(1) - a_{22}g_2(1) - \tilde{k}\rho^2 g_2(1)]c_2 \\ +[g_3'''(1) - a_{21}g_3'(1) - a_{22}g_3(1) - \tilde{k}\rho^2 g_3(1)]c_3 + [g_4'''(1) - a_{21}g_4'(1) - a_{22}g_4(1) - \tilde{k}\rho^2 g_4(1)]c_4 = 0, \end{cases} \tag{39}$$

where $\tilde{k}$ is defined by (25).

From (36) and (37), for any $k, 1 \le k \le 4$,

$$\begin{cases} g_k(0) = 1 + \mathcal{O}(\frac{1}{\rho}), g_k'(0) = \rho\omega_k[1 + \mathcal{O}(\frac{1}{\rho})], \\ [g_k''(1) - a_{11}g_k'(1) - a_{12}g_k(1)] = (\rho\omega_k)^2 e^{\rho\omega_k}[1 + \mathcal{O}(\frac{1}{\rho})], \\ g_k'''(1) - a_{21}g_k'(1) - a_{22}g_k(1) - \tilde{k}\rho^2 g_k(1) \\ = (\rho\omega_k)^3 e^{\rho\omega_k}[1 + \mathcal{O}(\frac{1}{\rho})] - \tilde{k}\rho^2 e^{\rho\omega_k}[1 + \mathcal{O}(\frac{1}{\rho})], \end{cases} \tag{40}$$

and

$$|e^{\rho\omega_2}| \le 1, |e^{\rho\omega_1}| = \mathcal{O}(e^{-c|\rho|}) \text{ as } |\rho| \to \infty, \tag{41}$$

for some constant $c > 0$. Then we know that $g(z)$ is nonzero if and only if ρ satisfies the characteristic equation

$$\det \begin{pmatrix} [1] & [1] & [1] & [1] \\ \rho\omega_1[1] & \rho\omega_2[1] & \rho\omega_3[1] & \rho\omega_4[1] \\ (\rho\omega_1)^2 e^{\rho\omega_1}[1] & (\rho\omega_2)^2 e^{\rho\omega_2}[1] & (\rho\omega_3)^2 e^{\rho\omega_3}[1] & (\rho\omega_4)^2 e^{\rho\omega_4}[1] \\ (\rho\omega_1)^3 e^{\rho\omega_1}[1] & (\rho\omega_2)^3 e^{\rho\omega_2}[1] & (\rho\omega_3)^3 e^{\rho\omega_3}[1] & (\rho\omega_4)^3 e^{\rho\omega_4}[1] \end{pmatrix} = 0,$$

where $[1] = 1 + \mathcal{O}(\frac{1}{\rho})$. Since $\omega_4 = -\omega_1, \omega_3 = -\omega_2$, the above equation is equivalent to

$$\det \begin{pmatrix} [1] & [1] & e^{\rho\omega_2}[1] & e^{\rho\omega_1}[1] \\ \omega_1[1] & \omega_2[1] & -\omega_2 e^{\rho\omega_2}[1] & -\omega_1 e^{\rho\omega_1}[1] \\ \omega_1^2 e^{\rho\omega_1}[1] & \omega_2^2 e^{\rho\omega_2}[1] & \omega_2^2[1] & \omega_1^2[1] \\ \omega_1^3 e^{\rho\omega_1}[1] & \omega_2^3 e^{\rho\omega_2}[1] & -\omega_2^3[1] & -\omega_1^3[1] \end{pmatrix} = 0. \tag{42}$$

Noting that each element of the matrix in (42) is bounded, we may rewrite (42) as

$$\det \begin{pmatrix} 1 & 1 & e^{\rho\omega_2} & 0 \\ \omega_1 & \omega_2 & -\omega_2 e^{\rho\omega_2} & 0 \\ 0 & \omega_2^2 e^{\rho\omega_2} & \omega_2^2 & \omega_1^2 \\ 0 & \omega_2^3 e^{\rho\omega_2} & -\omega_2^3 & -\omega_1^3 \end{pmatrix} + \mathcal{O}\left(\frac{1}{\rho}\right) = 0, \tag{43}$$

which results in

$$e^{2\rho\omega_2} = \left(\frac{\omega_2 - \omega_1}{\omega_2 + \omega_1}\right)^2 + \mathcal{O}\left(\frac{1}{\rho}\right) = -1 + \mathcal{O}\left(\frac{1}{\rho}\right). \tag{44}$$

By solving (44), we obtain (8) by the same arguments as those of section 4, Chapter 2 of [4]. Since the matrix in (43) has rank 3 for each sufficiently large ρ_n, there is only one linearly independent solution g to (32) for $\rho = \rho_n$. Hence each λ_n is geometrically simple for n sufficiently large. □

In Remark 4, we will indicate that each eigenvalue of $\mathcal{A}$ must be geometrically simple. Noting (37), (38), and (42), we can write g, g'' as

$$g(z) = \det \begin{pmatrix} [1] & [1] & e^{\rho\omega_2}[1] & e^{\rho\omega_1}[1] \\ e^{\rho\omega_1 z}[1] & e^{\rho\omega_2 z}[1] & e^{\rho\omega_2(1-z)}[1] & e^{\rho\omega_1(1-z)}[1] \\ \omega_1^2 e^{\rho\omega_1}[1] & \omega_2^2 e^{\rho\omega_2}[1] & \omega_2^2[1] & \omega_1^2[1] \\ \omega_1^3 e^{\rho\omega_1}[1] & \omega_2^3 e^{\rho\omega_2}[1] & -\omega_2^3[1] & -\omega_1^3[1] \end{pmatrix}, \tag{45}$$

$$g''(z) = \rho^2 \det \begin{pmatrix} [1] & [1] & e^{\rho\omega_2}[1] & e^{\rho\omega_1}[1] \\ \omega_1^2 e^{\rho\omega_1 z}[1] & \omega_2^2 e^{\rho\omega_2 z}[1] & \omega_2^2 e^{\rho\omega_2(1-z)}[1] & \omega_1^2 e^{\rho\omega_1(1-z)}[1] \\ \omega_1^2 e^{\rho\omega_1}[1] & \omega_2^2 e^{\rho\omega_2}[1] & \omega_2^2[1] & \omega_1^2[1] \\ \omega_1^3 e^{\rho\omega_1}[1] & \omega_2^3 e^{\rho\omega_2}[1]] & -\omega_2^3[1] & -\omega_1^3[1] \end{pmatrix}. \tag{46}$$

LEMMA 3.2. *Let λ_n, ρ_n be defined as in Proposition* 2.2. *Then the unique (up to a scalar) associated solution g_n to* (32) *has the following asymptotic expansion:*

$$-\frac{\sqrt{2}}{4}(1+i)g_n(z) = \sin(n+\pi/2)z - \cos(n+\pi/2)z + e^{-(n+1/2)\pi z} + (-1)^n e^{-(n+1/2)\pi(1-z)} + \mathcal{O}(n^{-1}), \tag{47}$$

$$-\frac{\sqrt{2}}{4}(1+i)\rho_n^{-2}g_n''(z) = i[\cos(n+\pi/2)z - \sin(n+\pi/2)z + e^{-(n+1/2)\pi z} + (-1)^n e^{-(n+1/2)\pi(1-z)}] + \mathcal{O}(n^{-1}). \tag{48}$$

Moreover, it follows directly from (37) *and* (45) *that*

$$\rho_n^{-2}g_n'(z) = \mathcal{O}(n^{-1}). \tag{49}$$

Proof. It follows from (45) that

$$g_n(z) = \det\begin{pmatrix} 1 & 1 & e^{\rho_n\omega_2} & 0 \\ e^{\rho_n\omega_1 z} & e^{\rho_n\omega_2 z} & e^{\rho_n\omega_2(1-z)} & e^{\rho_n\omega_1(1-z)} \\ 0 & \omega_2^2 e^{\rho_n\omega_2} & \omega_2^2 & \omega_1^2 \\ 0 & \omega_2^3 e^{\rho_n\omega_2} & -\omega_2^3 & -\omega_1^3 \end{pmatrix} + \mathcal{O}\left(\frac{1}{\rho_n}\right).$$

After a simple calculation, we find that

$$\begin{aligned} g_n(z) &= \omega_1^2\omega_2^2[2\omega_1 e^{\rho_n\omega_1 z} + 2\omega_2 e^{\rho_n\omega_2} e^{\rho_n\omega_1(1-z)} \\ &\quad + (\omega_2+\omega_1)e^{\rho_n\omega_2} e^{\rho_n\omega_2(1-z)} + (\omega_2-\omega_1)e^{\rho_n\omega_2 z}] + \mathcal{O}(\tfrac{1}{\rho_n}) \\ &= \sqrt{2}(i-1)[\sin(n+\pi/2)z - \cos(n+\pi/2)z \\ &\quad + e^{-(n+1/2)\pi z} + (-1)^n e^{-(n+1/2)\pi(1-z)}] + \mathcal{O}(\tfrac{1}{n}). \end{aligned}$$

This is (47). Equation (48) can be proved similarly. □

Note that the asymptotic expansions (47) and (48) are exactly the same as those obtained in [1] for the eigenfunctions of system (2) with constant coefficients; i.e., $EI = \rho = \text{const}$. However, it should be pointed out that the estimates in [1] and [2] rely on the analytic expression of the eigenfunctions.

Proof of Proposition 2.3. The result follows directly from the following facts that are deduced from transformations (29), (31), and (49):

$$\begin{cases} -\frac{\sqrt{2}}{4}(1+i)e^{\frac{1}{4}\int_0^z a(\tau)d\tau} f_n(z) = -\frac{\sqrt{2}}{4}(1+i)g_n(z), \\ -\frac{\sqrt{2}}{4}(1+i)e^{\frac{1}{4}\int_0^z a(\tau)d\tau}\rho_n^{-2}f_n''(z) = -\frac{\sqrt{2}}{4}(1+i)\rho_n^{-2}g_n''(z) + \mathcal{O}\left(\frac{1}{n}\right), \\ \phi_n(x) = f_n(z), \rho_n^{-2}\phi_n''(x) = \frac{1}{h^2}\left(\frac{\rho(x)}{EI(x)}\right)^{1/2}\rho_n^{-2}f_n''(z) + \mathcal{O}\left(\frac{1}{n}\right). \end{cases} \tag{50}$$

□

Before proving Theorem 2.4, let us recall that for a closed linear operator A in a Hilbert space H, a nonzero $x \in H$ is called a generalized eigenvector of A, corresponding to an eigenvalue λ of A which has finite algebraic multiplicity, if there is a positive integer n such that $(\lambda - \mathrm{A})^n x = 0$. A sequence $\{x_n\}_1^\infty$ in H is called a

Riesz basis for H if there is an orthonormal basis $\{e_n\}_1^\infty$ in H and a linear bounded invertible operator T such that

$$\mathrm{T}e_n = x_n, n = 1, 2, \ldots.$$

It is seen that each Riesz basis sequence must be approximately normalized:

$$C_1 \le \|x_n\| \le C_2, C_1, C_2 > 0, n = 1, 2, \ldots.$$

Suppose that $\{\lambda_n\}_1^\infty \subset \sigma(\mathrm{A})$ and lie in some left half complex plane. If each λ_n has finite algebraic multiplicity m_n and $m_n = 1$ as $n > N$ for some integer $N > 1$, then there is a sequence of linearly independent generalized eigenvectors $\{x_{n_i}\}_{i=1}^{m_n}$ corresponding to λ_n. If $\{\{x_{n_i}\}_{i=1}^{m_n}\}_{n=1}^\infty$ forms a Riesz basis for H, then A generates a C_0-semigroup $e^{\mathrm{A}t}$ which can be represented as

$$e^{\mathrm{A}t}x = \sum_{n=1}^{\infty} e^{\lambda_n t} \sum_{i=1}^{m_n} a_{ni} \sum_{j=1}^{m_n} f_{nj}(t)x_{n_j} \text{ for any } x = \sum_{n=1}^{\infty}\sum_{i=1}^{m_n} a_{ni}x_{n_i} \in H,$$

where $f_{nj}(t)$ is a polynomial of t with order less than m_n. In particular, if $a < \mathrm{Re}\lambda < b$ for some real numbers a and b, then A generates a C_0-group on H. Moreover, the spectrum-determined growth condition holds for $e^{\mathrm{A}t}$.

In order to remove the requirement of the estimation of the low eigenpairs of the system, a corollary of Bari's theorem is recently reported in [1] (a simplified proof can be found in [2]), which provides a much less demanding approach in generating a Riesz basis for general discrete operators in the Hilbert spaces. The result is cited here.

THEOREM 3.3. *Let* A *be a densely defined discrete operator (that is,* $(\lambda - \mathrm{A})^{-1}$ *is compact for some* λ*) in a Hilbert space* H*. Let* $\{z_n\}_1^\infty$ *be a Riesz basis for* H*. If there are an* $N \ge 0$ *and a sequence of generalized eigenvectors* $\{x_n\}_{N+1}^\infty$ *of* A *such that*

$$\sum_{N+1}^{\infty} \|x_n - z_n\|^2 < \infty,$$

then

(i) *There are an* $M > N$ *and generalized eigenvectors* $\{x_{n0}\}_1^M$ *of* A *such that* $\{x_{n0}\}_1^M \cup \{x_n\}_{M+1}^\infty$ *forms a Riesz basis for* H*.*

(ii) *Consequently, let* $\{x_{n0}\}_1^M \cup \{x_n\}_{M+1}^\infty$ *correspond to eigenvalues* $\{\sigma_n\}_1^\infty$ *of* A*. Then* $\sigma(\mathrm{A}) = \{\sigma_n\}_1^\infty$*, where* σ_n *is counted according to its algebraic multiplicity.*

(iii) *If there is an* $M_0 > 0$ *such that* $\sigma_n \ne \sigma_m$ *for all* $m, n > M_0$*, then there is an* $N_0 > M_0$ *such that all* $\sigma_n, n > N_0$ *are algebraically simple.*

Remark 3. It follows from Theorem 3.3 that when A and B satisfy the conditions (i) and (ii) of Theorem 2.7, $A + B$ is of $[D]$-class.

In order to apply Theorem 3.3 to the operator $\mathcal{A}$ when we consider $\{x_n\}$ in Theorem 3.3 as the eigenfunctions of $\mathcal{A}$, we need a referring Riesz basis $\{z_n\}_1^\infty$ as well. For the system (2), this is accomplished by collecting (approximately) normalized eigenfunctions of the following free conservative system:

$$\begin{cases} \rho(x)y_{tt}(x,t) + (EI(x)y_{xx}(x,t))_{xx} = 0, & 0 < x < 1, t > 0, \\ y(0,t) = y_x(0,t) = y_{xx}(1,t) = (EIy_{xx})_x(1,t) = 0. \end{cases} \tag{51}$$

The system operator $\mathcal{A}_0 : D(\mathcal{A}_0)(\subset \mathrm{H}) \to \mathrm{H}$ associated with (51) is nothing but the operator $\mathcal{A}$ with $k = 0$:

$$\begin{cases} \mathcal{A}_0(f,g) = (g, -\frac{1}{\rho(x)}(EI(x)f''(x))''), \\ D(\mathcal{A}_0) = \{(f,g) \in (H_E^2 \cap H^4) \times H_E^2 | f''(0) = f'''(1) = 0\}. \end{cases} \tag{52}$$

$\mathcal{A}_0$ is skew-adjoint with compact resolvent in H. Since Propositions 2.2 and 2.3 still keep valid when $k = 0$, we have the following counterpart for the operator $\mathcal{A}_0$.

LEMMA 3.4. *Each $\mu \in \sigma(\mathcal{A}_0)$, with sufficiently large modulus, is geometrically simple and hence algebraically simple. The eigenvalues $\{\lambda_{n0}, \overline{\lambda_{n0}}\}$ and the corresponding eigenfunctions $\{(\lambda_{n0}^{-1}\phi_{n0}, \phi_{n0})\} \cup \{$their conjugates$\}$ of $\mathcal{A}_0$ have the following asymptotic expressions:*

$$\lambda_{n0} = \frac{\rho_n^2}{h^2}, h = \int_0^1 \left(\frac{\rho(\tau)}{EI(\tau)}\right)^{1/4} d\tau, \rho_n = \frac{1}{\sqrt{2}}\left(n + \frac{1}{2}\right)\pi(1+i) + \mathcal{O}(n^{-1}) \text{ as } n \to \infty, \tag{53}$$

where n is a large positive integer, and

$$\begin{cases} -\frac{\sqrt{2}}{4}(1+i)e^{\frac{1}{4}\int_0^z a(\tau)d\tau}\phi_{n0}(x) = \sin(n+\pi/2)z - \cos(n+\pi/2)z \\ +e^{-(n+1/2)\pi z} + (-1)^n e^{-(n+1/2)\pi(1-z)} + \mathcal{O}(n^{-1}), \\ -\frac{\sqrt{2}}{4}(1+i)e^{\frac{1}{4}\int_0^z a(\tau)d\tau}\lambda_{n0}^{-1}\phi_{n0}''(x) = i\left(\frac{\rho(x)}{EI(x)}\right)^{1/2}[\cos(n+\pi/2)z - \sin(n+\pi/2)z \\ +e^{-(n+1/2)\pi z} + (-1)^n e^{-(n+1/2)\pi(1-z)}] + \mathcal{O}(n^{-1}). \end{cases} \tag{54}$$

Proof of Theorem 2.4. Since $\mathcal{A}_0$ is a skew-adjoint discrete operator in H, from a well-known result in functional analysis, the set of all ω-linearly independent eigenfunctions of $\mathcal{A}_0$ forms an orthogonal basis for H. Since $(\phi_{n0}, \lambda_{n0}\phi_{n0})$ defined by (54) are approximately normalized, $\{(\phi_{n0}, \lambda_{n0}\phi_{n0})\} \cup \{$their conjugates$\}$ form a (orthogonal) Riesz basis for H. Combining (9), (10), (53), and (54), we see that there is an $N > 0$ such that

$$\sum_{n>N}^{\infty} \|(\lambda_n^{-1}\phi_n, \phi_n) - (\lambda_{n0}^{-1}\phi_{n0}, \phi_{n0})\|_{\mathrm{H}}^2 = \sum_{n>N}^{\infty} \mathcal{O}(n^{-2}) < \infty. \tag{55}$$

The same is true for their conjugates. Hence the conditions of Theorem 2.5 are satisfied with correspondence $\mathcal{A} = A, x_n = (\lambda_n^{-1}\phi_n, \phi_n), z_n = (\lambda_{n0}^{-1}\phi_{n0}, \phi_{n0})$. The proof is complete. □

Now we are in a position to show the exponential stability confirmed by Theorem 2.7. Since the spectrum-determined growth condition holds, which is claimed by Theorem 2.4, system (2) is exponentially stable if and only if there is an $\omega > 0$ such that

$$\mathrm{Re}\lambda < -\omega \quad \forall\ \lambda \in \sigma(\mathcal{A}).$$

LEMMA 3.5. *Let λ_n be defined by* (8). *Then there is an $\omega_0 > 0$ such that*

$$\lim_{n\to\infty} \mathrm{Re}\lambda_n = -\omega_0 < 0. \tag{56}$$

Proof. Let $(\lambda, \phi) = (\lambda_n, \phi_n)$ in (7), where ϕ_n is defined by (9). Multiplying $\overline{\phi_n}$ on both sides of the first equation in (7) and integrating from 0 to 1 with respect to x, we obtain

$$\lambda_n^2 \int_0^1 \rho(x)|\phi_n(x)|^2 dx + \int_0^1 EI(x)|\phi_n''(x)|^2 dx + k\lambda_n|\phi_n(1)|^2 = 0.$$

Since $\mathrm{Im}\lambda_n \neq 0$ for sufficiently large n, we have, from the above equation, that

$$2\mathrm{Re}\lambda_n \int_0^1 \rho(x)|\phi_n(x)|^2 dx = -k|\phi_n(1)|^2 \text{ as } n \to \infty.$$

Then by (9) and the Riemann–Lebesgue lemma, we have

$$\lim_{n\to\infty} |\phi_n(1)|^2 = 16e^{-\frac{1}{2}\int_0^1 a(\tau)d\tau}, \lim_{n\to\infty} \int_0^1 \rho(x)|\phi_n(x)|^2 dx = 4\int_0^1 \rho(x)e^{-\frac{1}{2}\int_0^z a(\tau)d\tau} dx,$$

where $z = z(x)$ is specified by (29). Hence

$$\lim_{n\to\infty} \mathrm{Re}\lambda_n = -2k \frac{e^{-\frac{1}{2}\int_0^1 a(\tau)d\tau}}{\int_0^1 \rho(x)e^{-\frac{1}{2}\int_0^z a(\tau)d\tau} dx} < 0.$$

The result follows. □

Proof of Theorem 2.5. By Lemma 3.5 and the spectrum-determined growth condition, we need only show that

$$\mathrm{Re}\lambda < 0 \text{ for any } \lambda \in \sigma(\mathcal{A}). \tag{57}$$

Since the system is dissipative, $\mathrm{Re}\lambda \leq 0$ for any $\lambda \in \sigma(\mathcal{A})$. Suppose that $\mathrm{Re}\lambda = 0$. Then from $\mathrm{Re}\langle \mathcal{A}Y, Y\rangle = -k|\phi(1)|^2$ for each $Y = (\phi, \lambda\phi)$, we have $\phi(1) = 0$. In this case, (7) becomes

$$\begin{cases} \lambda^2\rho(x)\phi(x) + (EI(x)\phi''(x))'' = 0, & 0 < x < 1, \\ \phi(0) = \phi'(0) = \phi''(1) = \phi'''(1) = \phi(1) = 0. \end{cases} \tag{58}$$

The proof is complete if we can show that there is only zero solution to (58). To this end, we follow the idea used in [17].

First, we claim that there is at least one zero of ϕ in (0,1). In fact, by $\phi(0) = \phi(1) = 0$, it follows from Rolle's theorem that there is a $\xi_1 \in (0,1)$ such that $\phi'(\xi_1) = 0$, which, together with $\phi'(0) = 0$, claims that $(EI\phi'')(\xi_2) = 0$ for some $\xi_2 \in (0, \xi_1)$, and so $(EI\phi'')'(\xi_3) = 0$ for some $\xi_3 \in (\xi_2, 1)$ by the condition $(EI\phi'')(1) = 0$. Hence there is a $\xi_4 \in (\xi_3, 1)$ such that $(EI\phi'')''(\xi_4) = 0$ by the condition $(EI\phi'')'(1) = 0$. However, $(EI\phi'')''(\xi_4) = -\lambda^2\rho(\xi)\phi(\xi_4)$; we conclude that $\phi(\xi_4) = 0$.

Next, we show that if there are n different zeros of ϕ in (0,1), then there are at least $n+1$ number of different zeros of ϕ in (0,1).

Indeed, suppose that

$$0 < \xi_1 < \xi_2 < \cdots < \xi_n < 1, \phi(\xi_i) = 0, i = 1, 2, \ldots, n.$$

Since $\phi(0) = \phi(1) = 0$, it follows from Rolle's theorem that there are $\eta_i, i = 1, 2, \ldots, n+1$,

$$0 < \eta_1 < \xi_1 < \eta_2 < \xi_2 < \cdots < \xi_n < \eta_{n+1} < 1$$

such that $\phi'(\eta_i) = 0$. Noting that $\phi'(0) = 0$, there are $\alpha_i, i = 1, 2, \ldots, n+1$,

$$0 < \alpha_1 < \eta_1 < \alpha_2 < \eta_2 < \cdots < \alpha_{n+1} < \eta_{n+1} < 1$$

such that $(EI\phi'')(\alpha_i) = 0$. Since $(EI\phi'')(1) = 0$, using Rolle's theorem again, we have $\beta_i, i = 1, 2, \ldots, n+1$,

$$0 < \alpha_1 < \beta_1 < \alpha_2 < \cdots < \alpha_{n+1} < \beta_{n+1} < 1$$

such that $(EI\phi'')'(\beta_i) = 0$. Finally, by the condition $(EI\phi'')'(1) = 0$, we have $\vartheta_i, i = 1, 2, \ldots, n+1$,

$$0 < \beta_1 < \vartheta_1 < \beta_2 < \cdots < \beta_{n+1} < \vartheta_{n+1} < 1$$

such that $(EI\phi'')''(\vartheta_i) = 0$. Therefore,

$$\phi(\vartheta_i) = 0, i = 1, 2, \ldots, n+1.$$

Third, by mathematical induction, there is an infinite number of different zeros $\{x_i\}_1^\infty$ of ϕ in (0,1). Let $x_0 \in [0, 1]$ be an accumulation point of $\{x_i\}_1^\infty$. It is obvious that

$$\phi^{(i)}(x_0) = 0, \qquad i = 0, 1, 2, 3.$$

Note that ϕ satisfies the linear differential equation $(EI(x)\phi''(x))'' + \lambda^2\rho(x)\phi(x) = 0$. Therefore, $\phi \equiv 0$ by the uniqueness of the solution of linear ordinary different equations. □

Remark 4. The proof of Theorem 2.5 shows that each eigenvalue of $\mathcal{A}$ must be geometrically simple. In fact, suppose that $(\phi_1, \lambda\phi_1), (\phi_2, \lambda\phi_2)$ are any two eigenfunctions of $\mathcal{A}$ corresponding to λ. Then one can choose constants c_1, c_2 not identical to zero simultaneously such that $\phi = c_1\phi_1 + c_2\phi_2$ satisfies $\phi(1) = 0$. Now ϕ satisfies (58), and so $\phi \equiv 0$. Hence ϕ_1 and ϕ_2 are linearly independent.

From previous discussions, we see that our method can be easily used to get the Riesz basis property for the following beam equation under boundary moment feedback control:

$$\begin{cases} \rho(x)y_{tt}(x,t) + (EI(x)y_{xx}(x,t))_{xx} = 0, & 0 < x < 1, t > 0, \\ y(0,t) = y_x(0,t) = y_{xxx}(1,t) = 0, \\ y_{xx}(1,t) = -ky_{xt}(1,t), & k > 0. \end{cases} \tag{59}$$

It should be noted that the referring Riesz basis applied with Theorem 3.3 is accomplished by collecting all eigenfunctions of the following conservative free system:

$$\begin{cases} \rho(x)y_{tt}(x,t) + (EI(x)y_{xx}(x,t))_{xx} = 0, & 0 < x < 1, t > 0, \\ y(0,t) = y_x(0,t) = y_{xxx}(1,t) = y_{xt}(1,t) = 0. \end{cases} \tag{60}$$

This is the same as that of the uniform case [1]. Moreover, the analysis in this paper shows that the low order perturbations do not affect the basis property. For example, if we assume $b(x) \in C^3[0, 1]$, then Theorem 2.7 is still valid for the following system:

$$\begin{cases} \rho(x)y_{tt}(x,t) + b(x)y_{xxx}(x,t) + (EI(x)y_{xx}(x,t))_{xx} = 0, & 0 < x < 1, t > 0, \\ y(0,t) = y_x(0,t) = y_{xx}(1,t) = 0, (EI(x)y_{xx})_x(1,t) = ky_t(1,t). \end{cases} \tag{61}$$

Let us turn to system (12). First, we give a short proof of Theorem 2.7 by virtue of Theorem 3.3.

Proof of Theorem 2.7. Obviously, $A+B$ is a discrete operator in H. Write $A+B=A_s+T$, where $A_s\phi_n=\lambda_n\phi_n$ for all $n\geq 1$ and T is a linear bounded operator on H. We may assume without loss of generality that $\|\phi_n\|=1$ for all $n\geq 1$. Since $\{\phi_n\}_1^\infty$ is a Riesz basis, there is a $K>0$ such that for any $\phi=\sum_{n=1}^\infty a_n\phi_n$ and any complex series $\{\beta_n\}, |\beta_n|\leq 1$,

$$\left\|\sum_{n=1}^{\infty}\beta_n a_n\phi_n\right\|\leq K\|\phi\|. \tag{62}$$

By (15), we have $d_n\to\infty$ as $n\to\infty$. Hence for any $C>K\|T\|$, there is an integer $M>N$ such that $2\|T\|K/d_n<1$ for all $n\geq M$ and

$$|\lambda-\lambda_m|\geq C \text{ for any } \lambda \text{ satisfying } |\lambda-\lambda_n|=C, n\geq M.$$

First, for any $\phi=\sum_{n=1}^\infty a_n\phi_n$ and λ satisfying $|\lambda-\lambda_n|=C, n\geq M$,

$$\|CR(\lambda,A_s)\phi\|=\left\|\sum_{n=1}^{\infty}\frac{C}{\lambda-\lambda_n}a_n\phi_n\right\|\leq K\|\phi\|,$$

and so $\|R(\lambda,A_s)\|\leq K/C$. Hence $\|R(\lambda,A_s)T\|\leq K\|T\|/C<1$. This shows that $\{\lambda||\lambda-\lambda_n|=C, n\geq M\}\subset\rho(A_s+T)$ since $\lambda\in\sigma(A_s+T)$ if and only if $1\in\rho(R(\lambda,A_s)T)$. Let $\Gamma_n=\{\lambda||\lambda-\lambda_n|=C\}, n\geq$M. Consider eigenprojectors

$$\begin{aligned}Q_n-P_n&=\frac{1}{2\pi i}\int_{\Gamma_n}R(\lambda,A_s+T)d\lambda-\frac{1}{2\pi i}\int_{\Gamma_n}R(\lambda,A_s)d\lambda\\&=\frac{1}{2\pi i}\sum_{m=1}^{\infty}\int_{\Gamma_n}[R(\lambda,A_s)T]^mR(\lambda,A_s)d\lambda.\end{aligned}$$

One can choose $C>0$ large enough such that

$$\|Q_n-P_n\|\leq C\sum_{m=1}^{\infty}(K\|T\|/C)^mK/C=K\frac{K\|T\|/C}{1-K\|T\|/C}<1. \tag{63}$$

Therefore, $\dim(Q_n)=\dim(P_n)$. Hence there exists a unique $\mu_n, |\mu_n-\lambda_n|<C$ such that $\mu_n\in\sigma(A_s+T)=\sigma(A+B)$. This is (i). Moreover, since $\|P_n\phi_n\|=\|\phi_n\|=1$, we see that $Q_n\phi_n\neq 0$ and

$$Q_n\phi_n=\phi_n+\frac{1}{2\pi i}\sum_{m=1}^{\infty}\int_{\Gamma_n}[R(\lambda,A_s)T]^mR(\lambda,A_s)d\lambda d\lambda\phi_n. \tag{64}$$

Next, take $\Lambda_n=\{\lambda||\lambda-\lambda_n|=d_n/2\}, n\geq M$. Then for any $\phi=\sum_{n=1}^\infty a_n\phi_n$ and $\lambda\in\Lambda_n, \|d_n/2R(\lambda,A_s)\phi\|=\|\sum_{m=1}^\infty\frac{d_n}{2}\frac{1}{\lambda-\lambda_m}a_m\phi_m\|\leq K\|\phi\|$, and thus $\|R(\lambda,A_s)\|\leq\frac{2}{d_n}K$. Since $\|R(\lambda,A_s)T\|\leq\frac{2}{d_n}\|T\|K<1$, we see that $\{\Lambda_n, n\geq M\}\subset\rho(A_s+T)=\rho(A+B)$. Now consider

$$\begin{aligned}\tilde{Q}_n-P_n&=\frac{1}{2\pi i}\int_{\Lambda_n}R(\lambda,A_s+T)d\lambda-\frac{1}{2\pi i}\int_{\Lambda_n}R(\lambda,A_s)d\lambda\\&=\frac{1}{2\pi i}\sum_{m=1}^{\infty}\int_{\Lambda_n}[R(\lambda,A_s)T]^mR(\lambda,A_s)d\lambda.\end{aligned}$$

We have

$$\|\tilde{Q}_n - P_n\| \le \frac{\frac{2}{d_n}\|T\|K}{1-\frac{2}{d_n}\|T\|K}\|T\|K \le \frac{L}{d_n}, \qquad n \ge M,$$

for some constant $L > 0$. We may consider

$$\|\tilde{Q}_n - P_n\| \le \frac{L}{d_n} < 1, \qquad n \ge M. \tag{65}$$

Hence $\dim(\tilde{Q}_n) = \dim(P_n) = 1$, and $Q_n = \tilde{Q}_n$ as $n \ge M$. Therefore, $\psi_n = Q_n\phi_n$ satisfies

$$\|\psi_n - \phi_n\|^2 \le L^2 d_n^{-2} \text{ as } n \ge M, \tag{66}$$

proving the theorem. □

Note that the eigenproblem of (12) is to find the nonzero solution ψ such that

$$\begin{cases} \mu^2\rho(x)\psi(x) + \mu b(x)\psi(x) + (EI(x)\psi''(x))'' = 0, & 0 < x < 1, \\ \psi(0) = \psi'(0) = \psi''(1) = 0, \\ (EI(x)\psi'')'(1) = \mu k\psi(1), \end{cases} \tag{67}$$

and the eigenfunction of $\mathcal{A} + \mathcal{B}$ is of the form $(\psi, \mu\psi)$.

Proof of Corollary 2.9. Let $(\mu, \psi) = (\mu_n, \psi_n)$ in (67), where ψ_n is determined by (17). Multiplying $\overline{\psi_n}$ on both sides of the first equation in (67) and integrating from 0 to 1 with respect to x, we obtain

$$\mu_n^2 \int_0^1 \rho(x)|\psi_n(x)|^2dx + \mu_n \int_0^1 b(x)|\psi_n(x)|^2dx + \int_0^1 EI(x)|\psi_n''(x)|^2dx + k\lambda_n|\psi_n(1)|^2 = 0.$$

Since Im $\mu_n \ne 0$ for sufficiently large n, we have, from the above equation, the following:

$$\mathrm{Re}\mu_n = -\frac{1}{2}\frac{\int_0^1 b(x)|\psi_n(x)|^2dx + k|\psi_n(1)|^2}{\int_0^1 \rho(x)|\psi_n(x)|^2dx} \text{ as } n \to \infty. \tag{68}$$

It follows from (17) and (18) that $\|\psi_n - \phi_n\|_{L^2(0,1)} \to 0, \|\psi_n' - \phi_n'\|_{L^2(0,1)} \to 0$ as $n \to \infty$. By the trace theorem $|\psi_n(1) - \phi_n(1)| \to 0$. Therefore,

$$\mathrm{Re}\mu_n \to -\frac{1}{2}\frac{\int_0^1 b(x)|\phi_n(x)|^2dx + k|\phi_n(1)|^2}{\int_0^1 \rho(x)|\phi_n(x)|^2dx} \text{ as } n \to \infty. \tag{69}$$

Similar to the proof of Lemma 3.5, we obtain (19). □

Proof of Example 3. It follows from the proof of Corollary 2.9 that for any eigenfunction $(\psi, \mu\psi)$ of $\mathcal{A} + \mathcal{B}$

$$\mu^2 \int_0^1 |\psi(x)|^2dx + \mu \int_0^1 b(x)|\psi(x)|^2dx + \int_0^1 |\psi''(x)|^2dx = 0.$$

If Im $\mu = 0$, then from the above equation

$$(\mathrm{Re}\mu)^2 \int_0^1 |\psi(x)|^2dx + \mathrm{Re}\mu \int_0^1 b(x)|\psi(x)|^2dx + \int_0^1 |\psi''(x)|^2dx = 0.$$

Hence $\mathrm{Re}\mu < 0$. If $\mathrm{Im}\ \mu \neq 0$,

$$\mathrm{Re}\mu = -\frac{1}{2}\frac{\int_0^1 b(x)|\psi(x)|^2dx}{\int_0^1 |\psi(x)|^2dx} \leq -\frac{1}{2}\frac{b_0\int_a^b |\psi(x)|^2dx}{\int_0^1 |\psi(x)|^2dx} < 0.$$

Therefore, for any $\mu \in \sigma(\mathcal{A}+\mathcal{B})$, $\mathrm{Re}\mu < 0$. This, together with (20), gives the exponential stability of system (22), which is indicated in [18]. By similar reasoning, when $k = 0, b(x) \geq 0$ for $x \in [0,1]$ and $b(x) > b_0 > 0$ for all x in some subset $(a,b) \subset [0,1]$, system (12) is also exponential stable. □

Finally, we give the proof of Proposition 2.10. The validity of Proposition 2.10 deduces Lemma 3.5 automatically.

Proof of Proposition 2.10. Like the transformation from (7) to (32), (67) can be transformed into

$$\begin{cases} g^{(4)}(z) + a_1(z)g''(z) + a_2(z)g'(z) + a_3(z)g(z) + \lambda h^4\tilde{b}(z)g(z) + \lambda^2 h^4 g(z) = 0, \\ g(0) = g'(0) = 0, \\ g''(1) = a_{11}g'(1) + a_{12}g(1), \\ g'''(1) = a_{21}g'(1) + \left[\lambda\dfrac{kh^3}{EI(1)}\left(\dfrac{\rho(1)}{EI(1)}\right)^{-3/4} + a_{22}\right]g(1), \end{cases} \tag{70}$$

where the functions are the same as those in (32). By Theorem 2.8 and Corollary 2.9, all eigenvalues of $\mathcal{A}+\mathcal{B}$ with sufficiently large modulus must be located on the left complex plane under the assumption (23). Following [4], by noticing the smooth assumption (3) and (23), we know that for $\lambda = \rho^2/h^2$, $|\rho|$ sufficiently large,

$$g^{(4)}(z) + a_1(z)g''(z) + a_2(z)g'(z) + a_3(z)g(z) + \lambda h^4\tilde{b}(z) + \lambda^2 h^4 g(z) = 0$$

admits four linearly independent solutions $g_k, k = 1,2,3,4$ for any $\rho \in S$, which satisfy

$$\begin{cases} g_k(z) = e^{\rho\omega_k z}[1 + \frac{L_k(z)}{\rho} + \mathcal{O}(\frac{1}{\rho^2})], \\ g_k'(z) = \rho\omega_k e^{\rho\omega_k z}[1 + \frac{L_k(z)}{\rho} + \mathcal{O}(\frac{1}{\rho^2})], \\ g_k''(z) = (\rho\omega_k)^2 e^{\rho\omega_k z}[1 + \frac{L_k(z)}{\rho} + \mathcal{O}(\frac{1}{\rho^2})], \\ g_k'''(z) = (\rho\omega_k)^3 e^{\rho\omega_k z}[1 + \frac{L_k(z)}{\rho} + \mathcal{O}(\frac{1}{\rho^2})], \qquad k = 1,2,3,4, \end{cases} \tag{71}$$

where

$$L_k(z) = -\frac{1}{4\omega_k}\int_0^z a_1(\tau)d\tau + \frac{h^2}{4}\omega_k\int_0^z \tilde{b}(\tau)d\tau. \tag{72}$$

Similar to (37)–(40), by noting (71), we can write the characteristic equation (42) as

$$\det\begin{pmatrix} 1 & 1 & e^{\rho\omega_2} & 0 \\ \omega_1 & \omega_2 & -\omega_2 e^{\rho\omega_2} & 0 \\ 0 & \omega_2^2 e^{\rho\omega_2}\left[1+\dfrac{\ell_2}{\rho}\right] & \omega_2^2\left[1+\dfrac{\ell_3}{\rho}\right] & \omega_1^2\left[1+\dfrac{\ell_4}{\rho}\right] \\ 0 & \omega_2^3 e^{\rho\omega_2}\left[1+\dfrac{\ell_2}{\rho}\right] - \dfrac{\tilde{k}}{\rho}e^{\rho\omega_2} & -\omega_2^3\left[1+\dfrac{\ell_3}{\rho}\right] - \dfrac{\tilde{k}}{\rho} & -\omega_1^3\left[1+\dfrac{\ell_4}{\rho}\right] - \dfrac{\tilde{k}}{\rho} \end{pmatrix} = \mathcal{O}\left(\frac{1}{\rho^2}\right), \tag{73}$$

where $\ell_k = L_k(1)$. A direct computation yields

$$e^{2\rho\omega_2} = -1 + 2\frac{\tilde{k}}{\rho}\omega_2 + \frac{2\ell_2}{\rho} + \mathcal{O}\left(\frac{1}{\rho^2}\right). \tag{74}$$

Substituting $\rho = -(n+1/2)\pi\omega_2 + \mathcal{O}(n^{-1})$ into (74), the term $\mathcal{O}(n^{-1})$ satisfies

$$-2\omega_2\mathcal{O}(n^{-1}) = \frac{2\tilde{k}}{(n+1/2)\pi} - \frac{2\ell_2}{(n+1/2)\pi\omega_2} + \mathcal{O}(n^{-2});$$

hence

$$\mathcal{O}(n^{-1}) = \frac{\tilde{k}}{(n+1/2)\pi\omega_2} - \frac{2\ell_2}{(n+1/2)\pi\omega_2}\frac{1}{2\omega_2} + \mathcal{O}(n^{-2}).$$

Therefore,

$$\rho = -(n+1/2)\pi\omega_2 + \frac{2\tilde{k}}{2(n+1/2)\pi\omega_2} + \frac{2\ell_2}{(n+1/2)\pi\omega_2}\frac{1}{2\omega_2} + \mathcal{O}(n^{-2}),$$

which produces

$$\lambda h^2 = \rho^2 = -2\tilde{k} + i[(n+1/2)\pi]^2 - \frac{2\ell_2}{\omega_2} + \mathcal{O}(n^{-1}).$$

The required result then follows. □

It is seen that Proposition 2.10 coincides with the estimates in [1] for the uniform system (1) with $k_1 = 0, b = 0$.

Thus, from (24), condition (23) can be replaced by

$$\int_0^1 \tilde{b}(z)dz = \frac{1}{h}\int_0^1 \frac{b(x)}{\rho(x)}\left(\frac{\rho(x)}{EI(x)}\right)^{1/4} dx > 0. \tag{75}$$

For the case of $EI = \rho = 1$, the result can be found in [6].

REFERENCES

[1] B.-Z. Guo and R. Yu, *On Riesz basis property of discrete operators with application to an Euler-Bernoulli beam equation with boundary linear feedback control*, IMA J. Math. Control Inform., 18 (2001), pp. 241–251.

[2] B.-Z. Guo, *Riesz basis approach to the stabilization of a flexible beam with a tip mass*, SIAM J. Control Optim., 39 (2001), pp. 1736–1747.

[3] B.-Z. Guo and K. Y. Chan, *Riesz basis generation, eigenvalues distribution, and exponential stability for an Euler–Bernoulli beam with joint feedback control*, Rev. Mat. Complut., 14 (2001), pp. 205–229.

[4] M. A. Naimark, *Linear Differential Operators*, Vol. I, Ungar, New York, 1967.

[5] G. D. Birkhoff, *On the asymptotic character of the solutions of certain linear differential equations containing a parameter*, Trans. Amer. Math. Soc., 9 (1908), pp. 219–231.

[6] H. Wang and G. Chen, *Asymptotic locations of eigenfrequencies of Euler–Bernoulli beam with nonhomogeneous structural and viscous damping coefficients*, SIAM J. Control Optim., 29 (1991), pp. 347–367.

[7] B. P. Rao, *Optimal energy decay rate in a damped Rayleigh beam*, in Optimization Methods in Partial Differential Equations, Contemp. Math. 209, S. Cox and I. Lasiecka, eds., AMS, Providence, RI, 1997, pp. 211–229.

[8] S. Li, J. Yu, Z. Liang, and G. Zhu, *Stabilization of high eigenfrequencies of a beam equation with generalized viscous damping*, SIAM J. Control Optim., 37 (1999), pp. 1767–1779.

[9] S. Cox and E. Zuazua, *The rate at which energy decays in a damped string*, Comm. Partial Differential Equations, 19 (1994), pp. 213–243.

[10] A. Shubov, *Basis property of eigenfunctions of nonselfadjoint operator pencils generated by the equation of nonhomogenerous damped string*, Integral Equations Operator Theory, 25 (1996), pp. 289–328.

[11] M. A. Shubov, *Spectral operators generated by damped hyperbolic equations*, Integral Equations Operator Theory, 28 (1997), pp. 358–372.
[12] F. Conrad and Ö. Morgül, *On the stabilization of a flexible beam with a tip mass*, SIAM J. Control Optim., 36 (1998), pp. 1962–1986.
[13] G. Chen, S. A. Fulling, F. J. Narcowich, and C. Qi, *An asymptotic average decay rate for the wave equation with variable coefficient viscous damping*, SIAM J. Appl. Math., 50 (1990), pp. 1341–1347.
[14] B. R. Li, *The perturbation theory of a class of linear operators with applications*, Acta Math. Sinica, 21 (1978), pp. 206–222 (in Chinese).
[15] A. G. Ramm, *On the basis property for the root vectors of some nonselfadjoint operators*, J. Math. Anal. Appl., 80 (1981), pp. 57–66.
[16] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1966.
[17] D. X. Feng and G. T. Zhu, *The spectral property of the system operator arising from a class of elastic vibration problems*, Sciences Bulletin Ser. A, 24 (1981), pp. 1473–1475 (in Chinese).
[18] G. Chen, S. A. Fulling, F. J. Narcowich, and S. Sun, *Exponential decay of energy of evolution equations with locally distributed damping*, SIAM J. Appl. Math., 51 (1991), pp. 266–301.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 6, pp. 1924–1937

ANALYSIS OF THE HAMILTON–JACOBI EQUATION IN NONLINEAR CONTROL THEORY BY SYMPLECTIC GEOMETRY*

NOBORU SAKAMOTO†

Abstract. In this paper, the geometric property and structure of the Hamilton–Jacobi equation arising from nonlinear control theory are investigated using symplectic geometry. The generating function of symplectic transforms plays an important role in revealing the structure of the Hamilton–Jacobi equation. It is seen that many fundamental properties of the Riccati equation can be generalized in the Hamilton–Jacobi equation, and, therefore, the theory of the Hamilton–Jacobi equation naturally contains that of the Riccati equation.

Key words. Hamilton–Jacobi equation, nonlinear control theory, symplectic geometry

AMS subject classifications. 93C10, 93B36, 58F05

PII. S0363012999362803

1. Introduction.

The Hamilton–Jacobi equation plays a fundamental role in nonlinear control theory. The solvability of many important problems, such as optimal regulation [15], the H^∞ control problem [3, 11, 12, 21, 22], dissipative system theory [10, 26], and factorization theory [4] is represented by that of the Hamilton–Jacobi equation, and feedback functions depend on the solution of the Hamilton–Jacobi equation. When the system under consideration is linear and time-invariant, the Hamilton–Jacobi equation reduces to the Riccati equation. The most successful theory in linear control theory is H^∞ robust control. It naturally extends to nonlinear control systems. However, there are very few applications of nonlinear H^∞ control that have been reported, while linear H^∞ theory is now commonly used by industrial engineers. One of the biggest reasons for this is that little is known about the solution method, structure, and property of the Hamilton–Jacobi equation.

Only a few research papers have reported on the study of exact solutions of the Hamilton–Jacobi equation. In [21, 22], it is proved that if the Riccati equation constructed from the linear approximation of the Hamilton–Jacobi equation can be solved, then there exists, locally, a solution of the Hamilton–Jacobi equation, and nonlinear L^2 gain is also discussed by the linearization argument. After [21, 22], attention in nonlinear H^∞ theory has been paid not to the Hamilton–Jacobi equation itself, which represents the essential difficulty of nonlinearity, but to the derivation of the Hamilton–Jacobi equation, which is rather similar to that of the Riccati equation in linear H^∞ theory. For approximation methods of the solution of the Hamilton–Jacobi equation, we refer to [15] for the method by Taylor series expansion, [6, 7] for the Galerkin approximation, and [18] for the method by the state-dependent Riccati equation. One may also find extensive surveys on the research of the approximation of the Hamilton–Jacobi equation in [6, 7]. For the use of the theory of viscosity solutions in nonlinear H^∞ control when the system is not necessarily in the control-affine form, we refer to [5, 24, 25].

In this paper, we investigate the geometric property and structure of the Hamilton–Jacobi equation by using symplectic geometry. Symplectic geometry also plays a cen-

*Received by the editors October 12, 1999; accepted for publication (in revised form) November 1, 2001; published electronically March 20, 2002.
http://www.siam.org/journals/sicon/40-6/36280.html
†Department of Aerospace Engineering, Graduate School of Engineering, Nagoya University, Furocho, Chikusa-ku, Nagoya 464-8603, Japan (sakamoto@nuae.nagoya-u.ac.jp).

tral role in [21, 22], but the linearization argument is not employed in the present paper. Without linear approximation, it will be shown that major properties of the Riccati equation can be generalized and similar structures can be found in the Hamilton–Jacobi equation. In other words, the theory of the Riccati equation can be naturally embedded in that of the Hamilton–Jacobi equation. There exists a precise geometric theory in the Riccati equation developed in the course of the research from linear quadratic regulator theory to modern H^∞ robust control theory. Symplectic geometry provides a geometric framework in which we can investigate the Hamilton–Jacobi equation almost as well as we can for the Riccati equation.

The organization of the paper is as follows. In section 2, the theory of partial differential equations of the first order [8, 19] is outlined in the context of the Hamilton–Jacobi equation. According to the theory, to solve the Hamilton–Jacobi equation, all one has to do is solve a system of ordinary differential equations and/or integrate a completely integrable Pfaffian system. In section 3, attention is paid to a special kind of solution, a stabilizing solution, which plays an important role in control theory. The well-known theory [2, 20] by Arimoto and Potter, that is, a necessary and sufficient condition of the existence of a stabilizing solution, is reviewed as a part of the theory of the Hamilton–Jacobi equation. In control theory, we are interested in obtaining as many solutions as possible. To do this, more information on the structure of the Hamilton–Jacobi equation will be necessary, which is indicated by Example 1 of section 4. In section 4.1, the theory of the generating function for symplectic transforms is introduced; this will play a central role in a later subsection. Given a solution to the Hamilton–Jacobi equation, an auxiliary equation is derived that determines the other solutions using the generating function of symplectic transforms. It will also be seen that the auxiliary equation even characterizes the whole structure of the Hamilton–Jacobi equation. The linear control theoretic explanation of this structure will also be provided.

2. Overview of the theory of partial differential equations of the first order. In this section, we outline, by using a symplectic geometry machinery, the essential parts of the theory of partial differential equations of the first order that will be necessary for the analysis of the Hamilton–Jacobi equation.

Let us consider a partial differential equation of the form

$$\text{(PD)} \qquad F(x_1, \dots, x_n, p_1, \dots, p_n) = 0,$$

where F is a C^∞ function of $2n$ variables, $x_1, \dots, x_n$ are independent variables, z is an unknown function, and $p_1 = \partial z/\partial x_1, \dots, p_n = \partial z/\partial x_n$. Let M be an n dimensional space for $(x_1, \dots, x_n)$. We regard the $2n$ dimensional space for $(x, p) = (x_1, \dots, x_n, p_1, \dots, p_n)$ as the cotangent bundle T^*M of M. T^*M is a symplectic manifold with $\theta = \sum_{i=1}^n dx_i \wedge dp_i$.

Let $\pi : T^*M \to M$ be a natural projection, and let $V \subset T^*M$ be a hypersurface defined by $F = 0$. Define a submanifold

$$\Lambda_Z = \{(x, p) \in T^*M \mid p_i = \partial z/\partial x_i(x), i = 1, \dots, n\}$$

for a smooth function $z(x)$. Then, $z(x)$ is a solution of (PD) if and only if $\Lambda_Z \subset V$. Furthermore, $\pi|_{\Lambda_Z} : \Lambda_Z \to M$ is a diffeomorphism, and Λ_Z is a Lagrangian submanifold because $\dim \Lambda_Z = n$ and

$$\theta|_{\Lambda_Z} = \sum_{1 \leqslant i < j \leqslant n} \left(\frac{\partial^2 z}{\partial x_j \partial x_i} - \frac{\partial^2 z}{\partial x_i \partial x_j} \right) dx_i \wedge dx_j = 0.$$

Conversely, it is well known (see, e.g., [1, 19]) that for a Lagrangian submanifold Λ passing through $q \in T^*M$ on which $\pi|_\Lambda : \Lambda \to M$ is a diffeomorphism, there exists a neighborhood U of q and a function $z(x)$ defined on $\pi(U)$ such that

$$\Lambda \cap U = \{(x,p) \in U \mid p_i = \partial z/\partial x_i(x), i = 1, \dots, n\}.$$

Therefore, finding a solution of (PD) is equivalent to finding a Lagrangian submanifold $\Lambda \subset V$ on which $\pi|_\Lambda : \Lambda \to M$ is a diffeomorphism.

Let $f_1 = F$. To construct such a Langrangian submanifold passing through $q \in T^*M$, and hence to obtain a solution defined on a neighborhood of $\pi(q)$, it suffices to find functions $f_2, \dots, f_n \in \mathsf{F}(T^*M)$ with $df_1(q) \wedge \cdots \wedge df_n(q) \neq 0$ such that $\{f_i, f_j\} = 0$ $(i, j = 1, \dots, n)$, where $\{\cdot, \cdot\}$ is the Poisson bracket, and

$$\left| \frac{\partial(f_1, \cdots, f_n)}{\partial(p_1, \cdots, p_n)} \right| (q) \neq 0. \tag{2.1}$$

For if functions $\varphi, \psi \in \mathsf{F}(T^*M)$ satisfy $\varphi|_\Lambda = \psi|_\Lambda = 0$ for $\Lambda = \{f_1 = \cdots = f_n = 0\}$, then there exist functions $a_i, b_j \in \mathsf{F}(T^*M)$ $(i, j = 1, \dots, n)$ such that $\varphi = \sum_{i=1}^n a_i f_i$ and $\psi = \sum_{j=1}^n b_j f_j$, and, therefore, we have $\{\varphi, \psi\}|_\Lambda = 0$ from

$$\{\varphi, \psi\} = \sum_{i,j=1}^n a_i f_j \{f_i, b_j\} + b_j f_i \{a_i, f_j\} + f_i f_j \{a_i, b_j\},$$

where we used the fact that an n dimensional submanifold Λ is Lagrangian if and only if $\{f, g\}|_\Lambda = 0$ for all $f, g \in \mathsf{F}(T^*M)$ satisfying $f|_\Lambda = g|_\Lambda = 0$. Note that the condition (2.1) implies, by the implicit function theorem, that $\pi|_\Lambda$ is a diffeomorphism of some neighborhood of q.

Since $\{F, \cdot\}$ is the Hamiltonian vector field X_F with Hamiltonian F, the functions $f_2, \dots, f_n$ above are integrals of X_F. The ordinary differential equation that gives the integral curve of X_F is

$$\begin{cases} \dfrac{dx_i}{dt} = \dfrac{\partial F}{\partial p_i}, \\ \dfrac{dp_i}{dt} = -\dfrac{\partial F}{\partial x_i} \end{cases} \qquad (i = 1, \dots, n). \tag{2.2}$$

An integral of X_F is a function that is constant along solutions of (2.2), and, therefore, we seek integrals of the following Pfaffian equations together with the condition (2.1):

$$\frac{dp_1}{-\partial F/\partial x_1} = \cdots = \frac{dp_n}{-\partial F/\partial x_n} = \frac{dx_1}{\partial F/\partial p_1} = \cdots = \frac{dx_n}{\partial F/\partial p_n},$$

which is called the Lagrange–Charpit system.

The theory for general partial differential equations is developed as contact geometry, and the complete treatment of them, that is, how the solvability of them can be reduced to the existence theory of solutions for ordinary differential equations, can be found in [19]. [14] also contains the geometric theory of first order partial differential equations.

3. Stabilizing solutions of the Hamilton–Jacobi equation. In what follows, we consider the Hamilton–Jacobi equation in nonlinear control theory

$$\text{(HJ)} \qquad H(x,p) = p^T f(x) - \frac{1}{2} p^T R(x) p + q(x) = 0,$$

where $f : M \to \mathbb{R}^n$, $R : M \to \mathbb{R}^{n\times n}$, $q : M \to \mathbb{R}$ are all C^∞ and $R(x)$ are symmetric for all $x \in M$. We also assume that f and q satisfy $f(0) = 0$, $q(0) = 0$, and $\frac{\partial q}{\partial x}(0) = 0$.

A stabilizing solution of (HJ) is defined as follows.

DEFINITION 3.1. *A solution $z(x)$ of* (HJ) *is said to be a stabilizing solution if $p(0) = 0$, and 0 is an asymptotically stable equilibrium of the vector field $f(x) - R(x)p(x)$, where $p(x) = (\partial z/\partial x)^T(x)$. A solution $z(x)$ is called an antistabilizing solution if $p(0) = 0$ and $-f(x) + R(x)p(x)$ has an asymptotically stable equilibrium at 0.*

For example, let us consider a classical nonlinear optimal regulator problem (see, e.g., [15]),

$$\dot{x} = f(x) + g(x)u, \quad x(0) = x_0,$$
$$J(x_0) = \int_0^\infty q(x(t)) + u(t)^T u(t)\, dt,$$

where $f(0) = 0$ and $q(x)$ is positive definite. The optimal feedback is

$$u = -\frac{1}{2} g(x)^T \left(\frac{\partial V}{\partial x}\right)^T (x),$$

where $V(x)$ is a stabilizing solution of the following Hamilton–Jacobi equation:

$$\frac{\partial V}{\partial x} f(x) - \frac{1}{4} \frac{\partial V}{\partial x} g(x) g(x)^T \left(\frac{\partial V}{\partial x}\right)^T + q(x) = 0, \quad V(0) = 0.$$

When the system is linear, the above equation reduces to a Riccati equation, which will be discussed in the next subsection. See also [4, 11, 12, 21, 22, 23] for details of the role of the stabilizing solution in nonlinear control theory.

Let V be the hypersurface in T^*M defined by (HJ). According to the theory of partial differential equations of the first order described in the previous section, a solution of (HJ) can be considered as a Lagrangian submanifold contained in V such that $\pi|_\Lambda$ is a diffeomorphism. However, (HJ) reduces to the Riccati equation when systems are linear time-invariant, and a precise theory is established for the Riccati equation. So the natural question would be how the theory of partial differential equations for (HJ) reduces to the theory of the Riccati equation. We clarify the relation between them in sections 3.1 and 3.2.

3.1. The Riccati equation. The algebraic Riccati equation

$$\text{(RIC)} \qquad PA + A^T P - PRP + Q = 0$$

plays a fundamental role in linear control theory such as linear quadratic regulator, H^∞ control, factorization theory, etc. In (RIC), A is a real square matrix of dimension n and R and Q are symmetric matrices of dimension n. A symmetric matrix P is said to be a stabilizing solution of (RIC) if it is a solution of (RIC) and $A - RP$ is stable.

The $2n \times 2n$ matrix

$$\mathrm{Ham} = \begin{pmatrix} A & -R \\ -Q & -A^T \end{pmatrix}$$

is called the Hamiltonian matrix of (RIC). A necessary and sufficient condition for the existence of a stabilizing solution [2, 20, 9, 13] is that (i) Ham has no eigenvalues on the imaginary axis, and (ii) the generalized eigenspace $\mathbb{E}_-$ for n stable eigenvalues satisfies the following complementary condition:

$$\mathbb{E}_- \oplus \mathrm{Im} \begin{pmatrix} 0 \\ I \end{pmatrix} = \mathbb{R}^{2n}.$$

The Hamilton–Jacobi equation corresponding to (RIC) is

$$(\mathrm{HJ})' \qquad H'(x,p) = p^T A x - \frac{1}{2} p^T R p + \frac{1}{2} x^T Q x = 0.$$

Let $V' := \{(x,p) \in T^*M \mid H'(x,p) = 0\}$. In this case, we look for a Lagrangian subspace, regarding T^*M as a linear space of $2n$ dimension, with the properties in section 2. A Lagrangian subspace is an n dimensional subspace of T^*M on which θ vanishes. Note that V' is a quadric surface in T^*M such as an ellipsoid or a hyperboloid.

By using the properties of Ham, if Ham has no eigenvalues on the imaginary axis, it is shown that there exist an $n \times n$ stable matrix E_1, an unstable matrix E_2, and $n \times n$ matrices T_1, T_2, T_3, and T_4 satisfying

$$\mathrm{Ham} \begin{pmatrix} T_1 & T_3 \\ T_2 & T_4 \end{pmatrix} = \begin{pmatrix} T_1 & T_3 \\ T_2 & T_4 \end{pmatrix} \begin{pmatrix} E_1 & 0 \\ 0 & E_2 \end{pmatrix}, \tag{3.1}$$

$$T^T J T = J, \tag{3.2}$$

where $J = \left(\begin{smallmatrix} 0 & I \\ -I & 0 \end{smallmatrix} \right)$. Thus it is seen that V' contains Lagrangian subspaces. If, in addition, $\mathbb{E}_-$ satisfies the complementary condition, there exists a Lagrangian subspace $\Lambda \subset V'$ such that $\pi|_\Lambda$ is an isomorphism. See [16, 23] for details.

3.2. Lagrangian submanifold for stabilizing solutions. The observation on Lagrangian subspace for (RIC) is naturally generalized for (HJ) using symplectic geometry (see also [12, 23]).

Equation (3.1) takes the form

$$X_H \circ g = Dg \cdot X_A \tag{3.3}$$

with an immersion g from a neighborhood of 0 in $\mathbb{R}^n$ into T^*M with $g(0) = (0,0)$ and a vector field X_A on $\mathbb{R}^n$. Write $g(u) = (g_1(u), g_2(u)) = (g_{11}(u), \ldots, g_{1n}(u), g_{21}(u), \ldots, g_{2n}(u))$. If $g_1 : \mathbb{R}^n \to M$ is a local diffeomorphism and X_A has an asymptotically stable equilibrium at the origin, the Lagrange submanifold described by g as its image yields a stabilizing solution.

Indeed, from (3.3), we have

$$F_t \circ g = g \circ \tilde{F}_t, \quad t \geqslant 0, \tag{3.4}$$

in a neighborhood of $0 \in \mathbb{R}^n$, where $F_t(x,p)$ and $\tilde{F}_t(u)$ are integral curves of X_H and X_A, respectively. Taking the derivative of (3.4) and using the fact that F_t is a symplectic transform yield

$$Dg(u)^T J\, Dg(u) = D\tilde{F}_t(u)^T Dg(\tilde{F}_t(u))^T J\, Dg(\tilde{F}_t(u))\, D\tilde{F}_t(u).$$

Thus, for a neighborhood U' of $0 \in \mathbb{R}^n$, it follows that

$$Dg(u)^T J\, Dg(u) = Dg_1(u)^T Dg_2(u) - Dg_2(u)^T Dg_1(u) = 0, \quad u \in U', \tag{3.5}$$

since X_A is asymptotically stable. Equation (3.5) corresponds to (3.2). Taking the derivative of (3.4) with respect to t and setting $t = 0$ and $x = g_1(u)$ give

$$X_H(x, g_2 \circ g_1^{-1}(x)) = \begin{pmatrix} I \\ Dg_2(u)Dg_1^{-1}(x) \end{pmatrix} \cdot Dg_1(u) X_A(u). \tag{3.6}$$

Define $p(x) = g_2 \circ g_1^{-1}(x)$ for $x \in U = g_1(U') \subset M$. Then, premultiplying (3.6) by $[Dp(x) \quad -I]$, we have

$$\begin{aligned} &Dp(x)f(x) - Dp(x)R(x)p(x) + \left(\frac{\partial f}{\partial x}\right)^T (x)p(x) \\ &\qquad - \frac{1}{2}\left(\frac{\partial R(x)p}{\partial x}\right)^T (x)p(x) + \left(\frac{\partial q}{\partial x}\right)^T (x) \\ &= \frac{\partial^T}{\partial x}\left\{p(x)^T f(x) - \frac{1}{2}p(x)^T R(x)p(x) + q(x)\right\} = 0, \end{aligned} \tag{3.7}$$

where $\frac{\partial R(x)p}{\partial x}$ denotes the Jacobian matrix of $R(x)p(x)$ for fixed $p(x)$. Therefore, we obtain (HJ) from $f(0) = 0$, $q(0) = 0$, $p(0) = g_2 \circ g_1^{-1}(0) = g_2(0) = 0$. The solution is obtained from

$$dz = g_{21} \circ g_1^{-1}(x)dx_1 + \cdots + g_{2n} \circ g_1^{-1}(x)dx_n,$$

the integrability of which follows from $Dp(x)^T = Dp(x)$. The solution is a stabilizing one since $f(x) - R(x)p(x) = ((g_1)_* X_A)(x)$, the push-forward (see, e.g., [1]) of X_A.

Comparing this to [9], one can see a parallel structure to the Riccati equation.

4. The structure of the Hamilton–Jacobi equation. Let us begin this section with the following example.

Example 1. Consider the equation

$$H' = -p_1 x_1 + p_2(x_1 + x_2) - \frac{1}{2}p_2^2 + \frac{1}{2}x_1^2 = 0. \tag{4.1}$$

This can be seen as a Riccati equation for

$$A = \begin{pmatrix} -1 & 0 \\ 1 & 1 \end{pmatrix}, \quad R = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad Q = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

The corresponding Hamiltonian vector field is $X_{H'} = (-x_1, x_1 + x_2, p_1 - p_2 - x_1, -p_2)$, and the Lagrange–Charpit system is

$$\frac{dx_1}{-x_1} = \frac{dx_2}{x_1 + x_2 - p_2} = \frac{dp_1}{p_1 - p_2 - x_1} = \frac{dp_2}{-p_2}.$$

The integrals H' and p_2/x_1 are readily obtained. The integral p_2/x_1 satisfies (2.1), and with a nonzero constant c we get a solution

$$P = \begin{pmatrix} 1/2 + c - c^2/2 & c \\ c & 0 \end{pmatrix}.$$

However, this solution is not a stabilizing one because $A-RP$ is not stable. To find a stabilizing solution, we need to find another integral. It is a simple task to state but not an easy one to handle to find out that $(-2x_1+2p_1-p_2)dx_1+2x_1dp_1-x_1dp_2=0$, $(-2x_1+2p_1-p_2)dx_1+2x_1dp_1-x_1dp_2=d(2x_1p_1-x_1^2-x_1p_2)$, and, therefore, $2x_1p_1-x_1^2-x_1p_2$ is an integral of the Lagrange–Charpit system. From this integral (with zero constant) and (4.1), we get

$$P=\begin{pmatrix}1&1\\1&2\end{pmatrix},\quad \begin{pmatrix}1/2&0\\0&0\end{pmatrix}.$$

The former is a stabilizing solution.

This example illustrates the difficulty of finding all solutions even for the Riccati equation by means of the method in section 2. Some solutions are obtained rather easily, and some are not. It would be, then, a natural desire to want to find another solution by using information from the solution on hand. We try that by using the generating function of symplectic transforms.

4.1. The generating function of symplectic transforms. In this subsection, we give a brief introduction to the theory of the generating function of symplectic transforms.

Let $x'_1(x,\,p),\dots,x'_n(x,\,p)$ be n independent functions defined on a neighborhood of $(0,\,0)\in T^*M$ satisfying $\{x'_k,\,x'_{k'}\}=0$ for $1\leqslant k,\,k'\leqslant n$ and

$$\left|\frac{\partial(x'_1,\dots,x'_n)}{\partial(x_{j_1},\dots,x_{j_{n-l}},p_{i_1},\dots,p_{i_l})}\right|(0,\,0)\neq 0 \tag{4.2}$$

for subsets $I=\{i_1,\dots,i_l\}$ and $J=\{j_1,\dots,j_{n-l}\}$ of $\{1,\dots,n\}$ with $I\cap J=\phi$. By the implicit function theorem,

$$\begin{cases} p_i=h_i(x',\,x_I,\,p_J) & (i\in I),\\ x_j=h_j(x',\,x_I,\,p_J) & (j\in J)\end{cases}$$

where $x_I=(x_{i_1},\dots,x_{i_l})$ and $p_J=(p_{j_1},\dots,p_{j_{n-l}})$. Since $\Lambda=\{(x,\,p)\,|\,x'_1=\cdots=x'_n=0\}$ is a Lagrangian submanifold, it follows that

$$0=\sum_{k=1}^{n}dx_k\wedge dp_k\bigg|_\Lambda=\left(\sum_{i\in I}dx_i\wedge dh_i+\sum_{j\in J}dh_j\wedge dp_j\right)\bigg|_\Lambda,$$

which implies

$$\frac{\partial h_i}{\partial x_\mu}=\frac{\partial h_\mu}{\partial x_i},\quad \frac{\partial h_i}{\partial p_j}=-\frac{\partial h_j}{\partial x_i},\quad \frac{\partial h_j}{\partial p_\nu}=\frac{\partial h_\nu}{\partial p_j}\quad (i,\,\mu\in I;\,j,\,\nu\in J). \tag{4.3}$$

These guarantee the existence of a function $\Phi(x',\,x_I,\,p_J)$ that satisfies

$$\begin{cases}\dfrac{\partial\Phi}{\partial x_i}=h_i & (i\in I),\\[2ex] \dfrac{\partial\Phi}{\partial p_j}=-h_j & (j\in J)\end{cases}$$

(see [19] for a proof).

Now define a function

$$\Omega(x',\, x_I,\, p_J) = \Phi(x',\, x_I,\, p_J) + \psi(x') + \sum_{j\in J} x_j p_j,$$

with an arbitrary function $\psi(x')$. Then

$$\begin{aligned}\sum_{k=1}^{n} p_k dx_k - d\Omega &= \sum_{k=1}^{n} p_k dx_k - \sum_{i\in I} h_i dx_i + \sum_{j\in J} h_j dp_j \\ &\quad - \sum_{k=1}^{n} \frac{\partial(\Phi+\psi)}{\partial x'_k} dx'_k - d\left(\sum_{j\in J} x_j p_j\right) \\ &= -\sum_{k=1}^{n} \frac{\partial(\Phi+\psi)}{\partial x'_k} dx'_k\end{aligned}$$

shows that the transformation defined by

$$\begin{cases} x'_k = x'_k(x,\, p), \\ p'_k = -\dfrac{\partial(\Phi+\psi)}{\partial x'_k}(x'(x,\, p)) \end{cases} \qquad (1 \leqslant k \leqslant n)$$

is a symplectic transform because $\sum_{k=1}^{n} dp_k \wedge dx_k = \sum_{k=1}^{n} p'_k \wedge dx'_k$.

Conversely, if a transformation $x' = x'(x,\, p)$, $p' = p'(x,\, p)$ is symplectic, then $\{x'_k,\, x'_{k'}\} = 0$ for $1 \leqslant k,\, k' \leqslant n$. Also, it can be shown that there exist subsets $I = \{i_1, \ldots, i_l\}$ and $J = \{j_1, \ldots, j_{n-l}\}$ with $I \cap J = \phi$ such that (4.2) holds. Therefore, there exists a function $\Phi(x',\, x_I,\, p_J)$ such that

$$\left|\frac{\partial^2 \Phi}{\partial(x_I,\, p_J)\partial x'}\right| \neq 0, \quad \begin{cases} p_i = \dfrac{\partial \Phi}{\partial x_i}, \quad x_j = -\dfrac{\partial \Phi}{\partial p_j} & (i \in I;\, j \in J), \\ p'_k = -\dfrac{\partial \Phi}{\partial x'_k} & (k = 1, \ldots, n). \end{cases}$$

See [17, 19] for details of the generating function of symplectic transforms.

4.2. Realization of the entire structure of (HJ) using the generating function of symplectic transforms. Motivated by Example 1, we propose the method of obtaining other solutions from one solution. This is done by clarifying the whole structure of the Hamilton–Jacobi equation that determines the rest of the solutions. The role of the generating function of symplectic transforms is significant.

THEOREM 4.1. *Let $z(x)$ be a solution of* (HJ) *defined on a neighborhood of $0 \in M$ and $x'_k = p_k - p_k(x)$ for $1 \leqslant k \leqslant n$. Suppose there exists a solution $z''(x')$ around $x' = 0$ to the auxiliary equation*

$$x'^T f^*(p'') - \frac{1}{2} x'^T R(p'') x' = 0, \tag{4.4}$$

where $p''_k = \partial z''/\partial x'_k$ $(1 \leqslant k \leqslant n)$ and $f^*(x) = f(x) - R(x)p(x)$. Then, together with $x'_k(x,\, p)$ $(1 \leqslant k \leqslant n)$, the functions $p'_1(x,\, p), \ldots, p'_n(x,\, p)$ defined by

$$p'_k(x,\, p) = -x_k + p''_k(x'(x,\, p)) = -x_k + \frac{\partial z''}{\partial x'_k}(x'(x,\, p)), \quad 1 \leqslant k \leqslant n,$$

form a symplectic transform

$$\alpha : \begin{cases} x'_k = p_k - p_k(x), \\ p'_k = -x_k + \dfrac{\partial z''}{\partial x'_k}(x'(x,\, p)) \end{cases} \qquad (1 \leqslant k \leqslant n), \tag{4.5}$$

and the Hamiltonian vector field X_H leaves the submanifold $\Lambda_+ = \{(x,\, p) \,|\, p'_1 = \cdots = p'_n = 0\}$ invariant.

Proof. We note that (4.2) holds for $I = \{1, \ldots, n\}$. Define a function $\Phi(x',\, x) = \sum_{k=1}^n x'_k x_k + z(x)$. Then it satisfies

$$\frac{\partial \Phi}{\partial x_k} = x'_k + p_k(x), \quad 1 \leqslant k \leqslant n.$$

Therefore, from the theory of the generating function of symplectic transforms, (4.5) is a symplectic transform for an arbitrary function $z''(x')$. Next we determine z'' so that Λ_+ is invariant for X_H; in other words, $X_H \cdot p'_k|_{p'_1 = \cdots = p'_n = 0} = 0$ for $1 \leqslant k \leqslant n$. This is equivalent to

$$\{H,\, p'_k\}|_{p'_1 = \cdots = p'_n = 0} = 0 \quad (1 \leqslant k \leqslant n).$$

Rewrite $H(x,\, p)$ in $(x',\, p')$ coordinates as

$$\begin{aligned} H(x,\, p) &= (x' + p(x))^T f(x) - \frac{1}{2}(x' + p(x))^T R(x)(x' + p(x)) + q(x) \\ &= p(x)^T f(x) - \frac{1}{2} p(x)^T R(x) p(x) p(x) + q(x) \\ &\quad + x'^T f(x) - x'^T R(x) p(x) - \frac{1}{2} x'^T R(x) x' \\ &= x'^T f^*(-p' + p''(x')) - \frac{1}{2} x'^T R(-p' + p''(x')) x', \end{aligned}$$

where we used the relation (HJ). Denote $\{\, ,\, \}'$ the Poisson bracket with respect to $(x',\, p')$ coordinates. Then $\{\cdot, \cdot\} = \{\cdot, \cdot\}'$ because $\alpha : (x,\, p) \mapsto (x',\, p')$ is a symplectic transform. Therefore,

$$\begin{aligned} &\{H,\, p'_k\}|_{p'_1 = \cdots = p'_n = 0} = 0 \qquad (1 \leqslant k \leqslant n) \\ &\Leftrightarrow \{H,\, p'_k\}'|_{p'_1 = \cdots = p'_n = 0} = 0 \quad (1 \leqslant k \leqslant n) \\ &\Leftrightarrow \left.\frac{\partial H}{\partial x'_k}\right|_{p'_1 = \cdots = p'_n = 0} = 0 \qquad (1 \leqslant k \leqslant n) \\ &\Leftrightarrow \frac{\partial}{\partial x'} \left\{ x'^T f^*(p''(x')) - \frac{1}{2} x'^T R(p''(x')) x' \right\} = 0 \\ &\Leftrightarrow x'^T f^*(p''(x')) - \frac{1}{2} x'^T R(p''(x')) x' = \text{const}\,. \end{aligned}$$

Substituting $x' = 0$ into the last equation yields (4.4). □

Remark. Let us interpret Theorem 4.1 in the linear case. Let P be a symmetric solution of (RIC). Then

$$\text{Ham} \begin{pmatrix} I \\ P \end{pmatrix} = \begin{pmatrix} I \\ P \end{pmatrix} E_1$$

for $E_1 = A - RP$. It can be verified that

$$\begin{pmatrix} x' \\ p' \end{pmatrix} = \begin{pmatrix} p - Px \\ -x + Sx' \end{pmatrix} = \begin{pmatrix} p - Px \\ Sp - (SP + I)x \end{pmatrix}$$

is a symplectic transform for any symmetric S. Also, a direct computation shows that $T = \left(\begin{smallmatrix} I & S \\ P & PS+I \end{smallmatrix}\right)$ satisfies $T^T J T = J$. (That is, T is symplectic.) If S is a solution of the Lyapunov equation

$$E_1 S + S E_1^T = R, \tag{4.6}$$

which corresponds to (4.4), then it follows that

$$\mathrm{Ham}\begin{pmatrix} S \\ PS + I \end{pmatrix} = -\begin{pmatrix} S \\ PS + I \end{pmatrix} (A - RP)^T$$

from (RIC) and (4.6). This means that (3.1) holds for

$$T = \begin{pmatrix} I & S \\ P & PS + I \end{pmatrix}, \quad E_1 = A - RP, \quad E_2 = -(A - RP)^T = -E_1^T.$$

What we did here was to obtain T_3 and T_4 (that is, to establish the second half of (3.1)) from a solution (the first half of (3.1)) by means of solving (4.6).

Example 1 (continued). The stabilizing solution of (4.1) can be obtained more easily by following the method in Theorem 4.1. The solution $P = \left(\begin{smallmatrix} 1/2+c-c^2/2 & c \\ c & 0 \end{smallmatrix}\right)$ was not a stabilizing solution. Let $c = 1$ for the sake of simplicity. Then $f^*(x) = (-x_1, x_2)$, $x_1' = p_1 - x_1 - x_2$, and $x_2' = p_2 - x_1$. The auxiliary equation (4.4) is

$$-x_1' p_1'' + x_2' p_2'' - \frac{1}{2} {x_2'}^2 = 0,$$

and the Lagrange–Charpit system is

$$\frac{dx_1'}{-x_1'} = \frac{dx_2'}{x_2'} = \frac{dp_1''}{p_1''} = \frac{dp_2''}{x_2' - p_2''}.$$

From the equation for the first and third parts, the integral $x_1' p_1''$ is found, and we have $p_1'' = 0$ and $p_2'' = \frac{1}{2} x_2'$. Thus, from (4.5),

$$\begin{aligned} p_1' &= -x_1 + 0 = -x_1, \\ p_2' &= -x_2 + \frac{1}{2} x_2' = \frac{1}{2} p_2 - \frac{1}{2} x_1 - x_2. \end{aligned}$$

Note that

$$\begin{aligned} &dx_1' \wedge dp_1' + dx_2' \wedge dp_2' \\ &= (dp_1 - dx_1 - dx_2) \wedge (-dx_1) + (dp_2 - dx_1) \wedge \left(\frac{1}{2} dp_2 - \frac{1}{2} dx_1 - dx_2\right) \\ &= dx_1 \wedge dp_1 + dx_2 \wedge dp_2, \end{aligned}$$

and the submanifolds $\{x_1' = 0\}$, $\{x_2' = 0\}$, $\{p_1' = 0\}$, and $\{p_2' = 0\}$ are invariant for $X_{H'}$. The stabilizing solution is obtained from $\{x_1' = p_2' = 0\}$.

Equation (3.1) suggests that Ham can be decomposed into two parts E_1 and E_2 by a similarity transformation. Also, in the remark after Theorem 4.1, if P is a stabilizing solution, then E_1 is a stable matrix, and, therefore, so is $-E_2$. This is a consequence from the well-known property of Ham that if λ is an eigenvalue of Ham, then so is $-\lambda$, which can be verified from $\mathrm{Ham}J + J\mathrm{Ham} = 0$. Are these properties of the Riccati equation carried over to the Hamilton–Jacobi equation? More specifically, is it possible to find a coordinate system in which X_H is decomposed into two noninteracting subsystems? And if $z(x)$ is a stabilizing solution and the conditions in Theorem 4.1 are fulfilled, is one of the subsystems of X_H in the new coordinates asymptotically stable and the other asymptotically unstable? The following example says that the first property may be partially true, but the second is not.

Example 2. The equation

$$-2p_1x_1^3 + 2p_2x_2 - p_2^2 + x_1^4 = 0$$

is of Hamilton–Jacobi type with a nonhyperbolic equilibrium. One still can obtain a stabilizing solution $z(x) = \frac{1}{4}x_1^2 + x_2^2$ using the method in section 2. The auxiliary equation (4.4) associated with this solution is

$$x_1' {p_1''}^3 + x_2' p_2'' + \frac{1}{2} {x_2'}^2 = 0.$$

The Lagrange–Charpit system is

$$\frac{dx_1'}{3x_1' {p_1''}^2} = \frac{dx_2'}{x_2'} = \frac{dp_1''}{-{p_1''}^3} = \frac{dp_2''}{-p_2'' - x_2'};$$

hence $p_1'' = 0$ and $p_2'' = -\frac{1}{2}x_2'$ are readily obtained from the equation for the second and fourth parts. The symplectic transform (4.5) is

$$\alpha : \begin{cases} x_1' = p_1 - \dfrac{1}{2}x_1, \\ x_2' = p_2 - 2x_2, \\ p_1' = -x_1, \\ p_2' = -\dfrac{1}{2}p_2. \end{cases}$$

The push-forward $\alpha_* X_H$ of X_H by α is

$$(\alpha_* X_H)(x',\, p') = D\alpha(\alpha^{-1}(x',\, p'))X_H(\alpha^{-1}(x',\, p')) = \begin{pmatrix} 3x_1' {p_2'}^2 \\ x_2' \\ -{p_1'}^3 \\ -p_2' \end{pmatrix}.$$

It can be seen that $(\alpha_* X_H)(0,\, p') = (0,\, 0,\, -x_1^3,\, -x_2)$, so the second half is an asymptotically stable vector field. However, the first half of $-(\alpha_* X_H)(x',\, 0) = (0,\, -x_2',\, 0,\, 0)$ is not asymptotically stable.

THEOREM 4.2. *Assume that the conditions of Theorem* 4.1 *are satisfied and* $p''(0) = 0$. *Then the following hold.*

(i) $(\alpha_* X_H)(0,\, p') = \begin{pmatrix} 0 \\ -f^*(-p') \end{pmatrix}$, *which equals* $\begin{pmatrix} 0 \\ f^*(x) \end{pmatrix}$ *in* x *coordinates.*

(ii) $(\alpha_* X_H)(x',\, 0) = \begin{pmatrix} f^{**}(x') \\ 0 \end{pmatrix}$, *where*

$$f^{**}(x') = -\left(\frac{\partial f^*}{\partial x}\right)^T (p''(x'))x' + \frac{1}{2}\left(\frac{\partial R(x)x'}{\partial x}\right)^T (p''(x'))x'.$$

Proof. We first claim that the inverse transformation α^{-1} of

$$\alpha : \begin{cases} x' = p - p(x), \\ p' = -x + p''(x'(x,\, p)) = -x + p''(p - p(x)) \end{cases}$$

is given by

$$\alpha^{-1} : \begin{cases} x = p''(x') - p', \\ p = p(p''(x') - p') + x'. \end{cases}$$

To see this, one can check that $\alpha \circ \alpha^{-1} = id_{(x',\, p')}$ and $\alpha^{-1} \circ \alpha = id_{(x,\, p)}$ by direct computations. The derivative of α is

$$D\alpha(x,\, p) = \begin{pmatrix} -\dfrac{\partial p}{\partial x}(x) & I \\ -I - \dfrac{\partial p''}{\partial x'}(p - p(x))\dfrac{\partial p}{\partial x}(x) & \dfrac{\partial p''}{\partial x'}(p - p(x)) \end{pmatrix}.$$

Since $z(x)$ is a solution of (HJ),

$$\begin{aligned} &\frac{\partial p}{\partial x}(x)f(x) - \frac{\partial p}{\partial x}(x)R(x)p(x) + \left(\frac{\partial f}{\partial x}\right)^T(x)p(x) \\ &\qquad - \frac{1}{2}\left(\frac{\partial R(x)p}{\partial x}\right)^T(x)p(x) + \left(\frac{\partial q}{\partial x}\right)^T(x) = 0 \end{aligned} \tag{4.7}$$

for all x in a neighborhood of $0 \in \mathbb{R}^n$. Premultiplying

$$X_H(\alpha^{-1}(0,\, p')) = \begin{pmatrix} f(-p') - R(-p')p(-p') \\ -\left(\dfrac{\partial f}{\partial x}\right)^T(-p')p + \dfrac{1}{2}\left(\dfrac{\partial R(x)p}{\partial x}\right)^T(-p')p(-p') - \left(\dfrac{\partial q}{\partial x}\right)^T(-p') \end{pmatrix}$$

by $D\alpha(\alpha^{-1}(0,\, p'))$ and using (4.7), the first part of the theorem follows.

The second part of the theorem is proved by premultiplying

$$\begin{aligned} &X_H(\alpha^{-1}(x',\, 0)) \\ &= X_H(p''(x'),\, p(p''(x')) + x') \\ &= \begin{pmatrix} f(p''(x')) - R(p''(x'))p(p''(x')) - R(p''(x'))x' \\ -\left(\dfrac{\partial f}{\partial x}\right)^T(p''(x'))\cdot(p(p''(x')) + x') \\ +\dfrac{1}{2}\left(\dfrac{\partial R(x)\cdot(p(p''(x')) + x')}{\partial x}\right)^T(p''(x'))\cdot(p(p''(x')) + x') - \left(\dfrac{\partial q}{\partial x}\right)^T(p''(x')) \end{pmatrix}, \end{aligned}$$

by

$$D\alpha(\alpha^{-1}(x',\, 0)) = \begin{pmatrix} -\dfrac{\partial p}{\partial x}(p''(x')) & I \\ -I - \dfrac{\partial p''}{\partial x'}(x')\dfrac{\partial p}{\partial x}(p''(x')) & \dfrac{\partial p''}{\partial x'}(x') \end{pmatrix},$$

where we use

$$\left(\frac{\partial R(x)p(p''(x'))}{\partial x}\right)^T(p''(x'))x' = \left(\frac{\partial R(x)x'}{\partial x}\right)^T(p''(x'))p(p''(x'))$$

and

$$f^*(p'(x')) - R(p''(x'))x' + \frac{\partial p''}{\partial x'}\left\{\left(\frac{\partial f^*}{\partial x}\right)^T (p''(x'))x' - \frac{1}{2}\left(\frac{\partial R(x)x'}{\partial x}\right)^T (p''(x'))x'\right\} = 0,$$

which is derived by taking the derivative of (4.4) with respect to x'. □

COROLLARY 4.3. *Under the assumptions in Theorem* 4.2,

$$Df^{**}(0) = -Df^*(0)^T.$$

Proof. This follows immediately from Theorem 4.2. □

Remark. Theorem 4.2 generalizes the eigenequation (3.1) and the result $E_2 = -E_1^T$ in the theory of the Riccati equation (see the remark after Theorem 4.1). It should also be noted that one can answer the question raised before Example 2 by the expression of f^{**} in Theorem 4.2.

5. Concluding remarks. In this paper, the geometric property and structure of the Hamilton–Jacobi equation have been investigated using symplectic geometry. Many fundamental properties of the Riccati equation can be generalized in the Hamilton–Jacobi equation, and, therefore, the theory of the Hamilton–Jacobi equation naturally contains that of the Riccati equation.

It should be noted that before the modern linear robust control such as H^∞ theory became a key technology in industry, there was a tremendous amount of research undertaken on the Riccati equation from an analytical viewpoint as well as a numerical viewpoint. The lack of applicability of nonlinear control theory up to this time, such as nonlinear H^∞ theory, is mainly due to the lack of understanding of the Hamilton–Jacobi equation. More research of the Hamilton–Jacobi equation needs to be done for the development of applicable nonlinear control theory.

REFERENCES

[1] R. A. ABRAHAM AND J. E. MARSDEN, *Foundations of Mechanics*, 2nd ed., Benjamin/Cummings, Reading, MA, 1978.

[2] S. ARIMOTO, *Optimal feedback control minimizing the effects of noise disturbances C*, Trans. SICE C, 2 (1966), pp. 1–7 (in Japanese).

[3] J. A. BALL, J. W. HELTON, AND M. L. WALKER, *H_∞ control for nonlinear systems with output feedback*, IEEE Trans. Automat. Control, 38 (1993), pp. 546–559.

[4] J. A. BALL AND A. J. VAN DER SCHAFT, *J-Inner-outer factorization, J-spectral factorization, and robust control for nonlinear systems*, IEEE Trans. Automat. Control, 41 (1996), pp. 379–392.

[5] M. BARDI AND I. CAPUZZO-DOLCETTA, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman equations*, Birkhäuser Boston, Boston, 1997.

[6] R. W. BEARD, G. N. SARDIS, AND J. T. WEN, *Galerkin approximations of the generalized Hamilton-Jacobi-Bellman equation*, Automatica J. IFAC, 33 (1997), pp. 2159–2177.

[7] R. W. BEARD AND T. W. MCLAIN, *Successive Galerkin approximation algorithms for nonlinear optimal and robust control*, Internat. J. Control, 71 (1998), pp. 717–743.

[8] C. CARATHÉODORY, *Calculus of Variations and Partial Differential Equations of the First Order. Part* I*: Partial Differential Equations of the First Order*, Holden-Day, San Francisco, London, Amsterdam, 1965.

[9] B. A. FRANCIS, *A Course in H_∞ Control Theory*, Springer-Verlag, New York, 1986.

[10] D. J. HILL AND P. J. MOYLAN, *The stability of nonlinear dissipative systems*, IEEE Trans. Automat. Control, 21 (1976), pp. 708–711.

[11] J. IMURA, H. MAEDA, T. SUGIE, AND T. YOSHIKAWA, *Robust stabilization of nonlinear systems by H_∞ state feedback*, Systems Control Lett., 24 (1995), pp. 103–114.

[12] A. Isidori and A. Astolfi, *Disturbance attenuation and H_∞ control via measurement feedback in nonlinear systems*, IEEE Trans. Automat. Control, 37 (1992), pp. 1283–1293.
[13] P. Lancaster and L. Rodman, *Algebraic Riccati Equations*, Oxford University Press, Oxford, UK, 1995.
[14] P. Libermann and C.-M. Marle, *Symplectic Geometry and Analytical Mechanics*, D. Reidel, Boston, 1987.
[15] D. L. Lukes, *Optimal regulation of nonlinear dynamical systems*, SIAM J. Control, 7 (1969), pp. 75–100.
[16] K. R. Meyer and G. R. Hall, *Introduction to Hamiltonian Dynamical Systems and the N-Body Problem*, Springer-Verlag, New York, 1991.
[17] D. McDuff and D. Salamon, *Introduction to Symplectic Topology*, 2nd ed., Oxford University Press, Oxford, UK, 1998.
[18] C. P. Mracek and J. R. Cloutier, *Control designs for the nonlinear benchmark problem via the state-dependent Riccati equation method*, Internat. J. Robust Nonlinear Control, 8 (1998), pp. 401–433.
[19] T. Oshima and H. Komatsuf, *Partial Differential Equations of the First Order*, Iwanami, Tokyo, 1977 (in Japanese).
[20] J. E. Potter, *Matrix quadratic solutions*, SIAM J. Appl. Math., 14 (1966), pp. 496–501.
[21] A. J. van der Schaft, *On a state space approach to nonlinear H_∞ control*, Systems Control Lett., 16 (1991), pp. 1–8.
[22] A. J. van der Schaft, *L_2-gain analysis of nonlinear systems and nonlinear state feedback H_∞ control*, IEEE Trans. Automat. Control, 37 (1992), pp. 770–784.
[23] A. J. van der Schaft, *L_2-Gain and Passivity Techniques in Nonlinear Control*, 2nd ed., Springer-Verlag, New York, 2000.
[24] P. Soravia, *H^∞ control of nonlinear systems: Differential games and viscosity solutions*, SIAM J. Control Optim., 34 (1996), pp. 1071–1097.
[25] P. Soravia, *Equivalence between nonlinear H^∞ control problems and existence of viscosity solutions of Hamilton-Jacobi-Issacs equations*, Appl. Math. Optim., 39 (1999), pp. 17–32.
[26] J. C. Willems, *Dissipative dynamical systems* I, II, Arch. Rational Mech. Anal., 45 (1972), pp. 321–393.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 6, pp. 1938–1964

STABILIZATION IN PROBABILITY OF NONLINEAR STOCHASTIC SYSTEMS WITH GUARANTEED COST*

S. BATTILOTTI† AND A. DE SANTIS†

Abstract. We deal with nonlinear dynamical systems, consisting of a linear nominal part plus model uncertainties, nonlinearities, and both additive and multiplicative random noise, modeled as a Wiener process. In particular, we study the problem of finding suitable measurement feedback control laws such that the resulting closed-loop system is stable in some probabilistic sense and a given cost functional is minimized. We give a Lyapunov-based separation result which splits the control design into a *state feedback* problem and a *filtering problem*. Finally, we point out constructive algorithms for solving the state feedback and filtering problems with arbitrarily large region of attraction for a wide class of nonlinear systems, which at least include *feedback linearizable systems*.

Key words. stochastic nonlinear systems, optimal control, robust stabilization, filtering

AMS subject classifications. 93E11, 93E20, 93D15, 93D21, 49K30, 60H10

PII. S0363012901375026

1. Introduction.

In modeling dynamical systems, the stochastic framework is suitable for taking into account either randomly varying system parameters or stochastic exogenous inputs. It is important in many practical situations to require, besides stability in some sense, some optimal and robustness performances, which can be usually described through a suitable cost functional. These performances may include tracking errors and physical constraints, due, for example, to control actuators or sensors with limited range.

According to the existing literature (see [20], [21], [16], [17], [18]; see also the textbooks [13] and [19]), by stability it is usually meant that

- the probability that the trajectory, stemming from x_0, leaves an ϵ-ball around the origin goes to zero as x_0 tends to the origin;
- the trajectory, stemming from x_0, goes asymptotically to zero almost surely.

This stability, known as *stability in probability*, is either *local* or *global* according to whether x_0 is in some (small) neighborhood of the origin or, respectively, it is *any point of the state space*. In [13] Lyapunov-based conditions are given for guaranteeing stability in probability, and they require the solution of partial differential inequalities (PDIs). In [15] and [17] it has been proved that a step-by-step algorithm (*backstepping*) can be successfully implemented for globally solving these PDIs whenever the state is available for feedback and the uncertainties have an upper triangular structure; by using the same backstepping design, in [16] the problem of global output feedback stabilization in probability is solved for the following class of nonlinear systems with triangular structure:

$$\begin{aligned} dx_i(t) &= x_{i+1}(t)dt + \varphi_i^T(y(t))dw(t), \quad i = 1, \dots, n-1, \\ dx_n(t) &= u(t)dt + \varphi_n^T(y(t))dw(t), \\ y(t) &= x_1(t), \end{aligned} \tag{1}$$

*Received by the editors May 10, 2001; accepted for publication (in revised form) November 1, 2001; published electronically March 20, 2002. An abridged version of this paper was presented at the IEEE Conference on Decision and Control, Sydney, Australia, 2000.
http://www.siam.org/journals/sicon/40-6/37502.html

†Dipartimento di Informatica e Sistemistica, Via Eudossiana 18, 00184 Roma, Italy (battilotti@dis.uniroma1.it, desantis@dis.uniroma1.it).

where $w(t)$ is a Wiener process. Even in a deterministic framework, as shown through some counterexamples in [11], the class of systems for which global stabilization can be achieved using output feedback can be only slightly enlarged with respect to (1). Indeed, as shown in [11], the system

$$\begin{aligned} \dot{x}_1(t) &= x_2(t), \\ \dot{x}_2(t) &= x_2^j(t) + u(t), \quad j \geq 3, \\ y(t) &= x_1(t) \end{aligned} \tag{2}$$

cannot be globally stabilized by any C^0 finite-dimensional output feedback dynamic controller. On the other hand, the earlier works of Esfandiari and Khalil [8], [9] have shown that feedback linearizable systems, such as, for example, (2), are instead *semiglobally* stabilizable via output feedback. *Semiglobal stabilization* was introduced in [4] and requires a local asymptotic stability of the closed-loop system plus a region of attraction containing any a priori given compact set of the state space. The basic ingredients for achieving semiglobal stability via output feedback are *control saturations* and *high-gain observers* [8], [9], [12]: large values of the observer gain guarantee that the error between the state and its estimate, generated by the observer itself, goes to zero "sufficiently fast," while input saturations rule out destabilizing effects such as *peaking* [6], which is a phenomenon occurring when one is trying to force some state variables to zero as fast as possible, causing an impulsive-like behavior of some others.

A first objective of our paper is to extend the notion of semiglobal stabilization to the following class of nonlinear stochastic systems:

$$\begin{aligned} dx(t) &= (Ax(t) + B_2u(t) + B_1\Phi(t, u(t), x(t)))dt + H(t, x(t))dw(t), \\ dy(t) &= (C_2x(t) + C_1\Phi(t, u(t), x(t)))dt + K(t, x(t))dw(t), \end{aligned} \tag{3}$$

where $w(t) \in \mathbb{R}^s$ is a Wiener process, $u(t) \in \mathbb{R}^m$ is the control, $x(t) \in \mathbb{R}^n$ is the state, $y(t) \in \mathbb{R}^p$ are the measurements, and $\Phi(t, u(t), x(t)) \in \mathbb{R}^r$ are model uncertainties and nonlinearities. The system

$$\begin{aligned} \dot{x}(t) &= Ax(t) + B_2u(t), \\ y(t) &= C_2x(t) \end{aligned} \tag{4}$$

can be understood as a *nominal system*, i.e., a system under nominal conditions, and Φ, Kdw, and Hdw are model nonlinearities and parameter uncertainties. We will not assume any *global growth condition* on Φ, H, and K as in [1]. Moreover, we will consider families of admissible controllers

$$\begin{aligned} u(t) &= \eta(F(k)\sigma(t)), \\ d\sigma(t) &= (L(k)\sigma(t) + B_2u(t))dt + G(k)dy(t), \quad \sigma \in \mathbb{R}^n, \end{aligned} \tag{5}$$

for $k \in \mathbb{R}^+$ and for some matrices $F(k)$, $L(k)$, and $G(k)$ and C^0 function $\eta : \mathbb{R}^m \to \mathbb{R}^m$, *linear* near the origin. This is a reasonable structure for the controller since near the origin (3) behaves as its own linearization.

Given numbers $\alpha, \beta \in [0, 1)$ and a pair of compact sets $\mathcal{B}^e \subset \Omega^e$ containing the origin, we use the same notion of stabilization in probability ($(\Omega^e, \mathcal{B}^e, \alpha, \beta)$-*SP* given in [3]. This notion requires that for sufficiently large k the trajectories of the closed-loop system, resulting from (3), with initial condition in Ω^e, remain inside some compact

set $\Omega^e(k) \supseteq \Omega^e$ of the state space, eventually enter any given neighborhood of the target set $\mathcal{B}^e$ in finite time, and remain therein with probability at least $(1-\alpha)(1-\beta)$. The numbers α and β are given *risk margins*: the first one quantifies the risk of leaving $\Omega^e(k)$ with initial condition in Ω^e rather than getting close to the target, while the second one gives a risk margin for remaining close to the target. If $\mathcal{B}^e = \{0\}$ and Ω^e can be taken to be any a priori given compact set of the state space and α and β any numbers in $[0, 1)$, our definition extends to a stochastic setting the notion of *semiglobal stabilization* as introduced in [4], and in what follows we will refer to this property as *semiglobal stabilization in probability*. If $\Omega^e = \mathbb{R}^{2n}$ and $\mathcal{B}^e$ can be taken any a priori given compact subset of Ω^e and α and β any a priori given numbers in $[0, 1)$, our definition gives a stochastic analogue of the concept of *practical stabilization*, which will be referred to as *practical stabilization in probability*.

As a second step, we define our optimality and robustness criteria. First, we give a set of admissibility constraints which impose precise characteristics to Φ, H, K, and η, generally satisfied under mild assumptions. As will be clear, these constraints lead to an optimal controller (5) in which

$$\begin{aligned} u(t) &= F(k)\sigma(t), \\ d\sigma(t) &= (L(k)\sigma(k) + B_2 u(t))dt + G(k)dy(t), \quad \sigma \in \mathbb{R}^n, \end{aligned} \tag{6}$$

is a *worst-case linear controller for the nominal system with disturbance* Φ.

The optimality criteria are formulated in terms of achieving either a guaranteed value or the minimum value of some cost functionals, according to whether multiplicative or additive noise is taken into account. These functionals penalize the "distance" from a reference situation for which the worst-case linear controller (6) is designed, and in the linear case they reduce to a standard quadratic cost (see [18] and [19] for comparisons with other inverse optimal schemes for deterministic and stochastic nonlinear systems).

We show that the problem of finding a stabilizing optimal controller can be split into two lower dimensional problems: one is related to the case in which the *state x is available for feedback* and the other to the possibility of *constructing an observer for estimating the state x*. Furthermore, we show that the conditions of our theorems can be actually met with an arbitrarily large region of attraction for a wide class of nonlinear stochastic systems with noiseless outputs, which include at least *feedback linearizable systems*, and we show that *control saturations* and *high gain observers* are instrumental in accomplishing this task exactly as in the case of deterministic systems. We do this into two steps. First, we give a semiglobal in probability backstepping design procedure for solving the state feedback problem, which stands as a *practical semiglobal* version of the corresponding *global* result proved in [15] and [17]. On the other hand, our step-by-step procedure is computationally simpler for the choice at each step of both the Lyapunov functions and the change of coordinates. Finally, we give some constructive tools for the observer design.

2. Notation and basic notions. First, we give some notation extensively used throughout the paper.

- If $\|v\|$ denotes the 2-norm of any given vector v, by $\|A\|$ we denote the induced 2-norm of any given matrix A; by $\|v\|_A$ we denote the A-norm of v, i.e., $\|v\|_A = \sqrt{v^T A v}$; let col $(v_1, \ldots, v_n)$ be the column vector with the ith entry equal to v_i.
- By $\mathcal{SP}^n$ (resp., $\mathcal{SN}^n$) we denote the set of $n \times n$ positive (resp., negative) definite symmetric matrices; by $\mathcal{SSP}^n$ we denote the set of $n \times n$ positive

semidefinite symmetric matrices; $\mathbb{R}^+$ denotes the set of positive real numbers and $\mathbb{R}^{\geq}$ the set of nonnegative real numbers.

- For any vector-valued function $\eta : \mathbb{R}^s \to \mathbb{R}^r$, we denote by η_i (or $[\eta]_i$) its ith component.
- For any given set $\mathcal{S}$, we denote by $\overline{\mathcal{S}}$ its closure and by $\partial\mathcal{S}$ its boundary; moreover, given $\delta > 0$ and a set $\mathcal{S}$, by *δ-neighborhood* of $\mathcal{S}$ we denote the set $\mathcal{S}_\delta = \{z : \inf_{y\in\mathcal{S}} \|z - y\| < \delta\}$.
- For any sequence of sets $\{\mathcal{S}_k\}$, $\liminf_{k\to\infty} \mathcal{S}_k = \cup_{k=1}^{\infty} \cap_{i\geq k} \mathcal{S}_i$ and $\limsup_{k\to\infty} \mathcal{S}_k = \cap_{k=1}^{\infty} \cup_{i\geq k} \mathcal{S}_i$. It is easy to see that if $\liminf_{k\to\infty} \mathcal{S}_k \supset \mathcal{V}$, then there exists k° such that $\mathcal{S}_k \supseteq \mathcal{V}$ for all $k \geq k^\circ$. Similarly, if $\limsup_{k\to\infty} \mathcal{S}_k \subset \mathcal{V}$, then there exists k° such that $\mathcal{S}_k \subseteq \mathcal{V}$ for all $k \geq k^\circ$.

In the remaining part of this section, we briefly recall some notions of stochastic processes, referring the reader to standard textbooks for the basic concepts [23], [24]. We assume that the reader is familiar with the basic notions of probability theory and stochastic processes $\{x(t), t \in \mathbb{R}\}$ on a given probability space $(\Omega, \mathcal{F}, \mathbf{P})$. (We assume that the probability space and all the σ-algebras we consider are completed with all the subsets of sets having null measure.) We denote by $\mathbf{E}\{\cdot\}$ the expectation and by $\mathbf{P}\{\cdot|\cdot\}$ ($\mathbf{E}\{\cdot|\cdot\}$) the conditional probability (expectation).

An important definition regards the notion of *Markov time*. Let $\{\mathcal{F}_t, t \in \mathbb{R}\}$ be an increasing family of right continuous σ-algebras contained in $\mathcal{F}$ (*filtration*).

DEFINITION 2.1. *A nonnegative random variable τ, $\tau \leq +\infty$, is called an $\mathcal{F}_t$ Markov time if for all $t \geq 0$ $\{\omega : \tau(\omega) \leq t\} \in \mathcal{F}_t$ (i.e., it is $\mathcal{F}_t$ adapted). If $\mathbf{P}\{\tau < \infty\} = 1$, then τ is called a stopping time.*

A stochastic process $\{x(t), t \in \mathbb{R}\}$ is a *Wiener process* (with respect to $\{\mathcal{F}_t, t \in \mathbb{R}\}$) if $\mathbf{E}\{x(t)|\mathcal{F}_s\} = x(s)$ and $\mathbf{E}\{(x(t) - x(s))^2|\mathcal{F}_s\} = t - s$ for $t \geq s$. A stochastic process $\{x(t), t \in \mathbb{R}\}$ is a *Markov process* if for any collections $t_1 < \cdots < t_N$ and $r_1, \ldots, r_N$,

$$\mathbf{P}\{x_{t_N} < r_N | x_{t_1} = r_1, \ldots, x_{t_{N-1}} = r_{N-1}\} = \mathbf{P}\{x_{t_N} < r_N | x_{t_{N-1}} = r_{N-1}\}. \tag{7}$$

For the corresponding definitions in the multidimensional case, we refer to [25].

By a *stochastic differential equation*, we mean the following equation:

$$dx(t) = f(x(t), t)dt + g(x(t), t)dw(t) \tag{8}$$

with initial condition $x(t_0) = \overline{x}$, where $\{w(t), t \in \mathbb{R}\}$ is a Wiener process (with respect to $\{\mathcal{F}_t, t \in \mathbb{R}\}$). The solution $x(t, t_0, \overline{x})$ of (8), whenever it exists, is a Markov process satisfying

$$x(t, t_0, \overline{x}) = \overline{x} + \int_{t_0}^{t} f(x(s, t_0, \overline{x}), s)ds + \int_{t_0}^{t} g(x(s, t_0, \overline{x}), s)dw(s) \tag{9}$$

almost surely (a.s.). The last integral is called the *Itô integral*. It is well known [13] that if

$$\begin{aligned} \|f(t, x_1) - f(x_2, t)\| + \|g(t, x_1) - g(t, x_2)\| &\leq K\|x_1 - x_2\|, \\ \|f(t, x)\| + \|g(t, x)\| &\leq H(1 + \|x\|) \end{aligned} \tag{10}$$

for all (x_1, t), (x_2, t), and (x, t) in $\mathcal{Z} \times [t_0, T]$, with $\mathcal{Z}$ a compact set containing $\overline{x}$, then there exists an a.s. unique stochastic process $x(t)$, sample continuous and satisfying (9) on $[t_0, \tau_{\mathcal{Z},T}(t)]$, where $\tau_{\mathcal{Z},T}(t) = \min(t, \tau_{\mathcal{Z}}, T)$ and $\tau_{\mathcal{Z}}$ is the Markov time (relative to the σ-algebra generated by $\{x(s), s \leq t\}$) defined as the first time at which $x(t)$ reaches the boundary of $\mathcal{Z}$ [13].

An important property of solutions of stochastic differential equations is *regularity*. Consider a sequence of increasing bounded domains $\{\mathcal{Z}(n)\}$, containing the origin, such that the distance of the boundary from the origin goes to infinity as n tends to infinity, and let $\{\tau_{\mathcal{Z}(n)}\}$ be the corresponding sequence of Markov times. Since $\{\tau_{\mathcal{Z}(n)}\}$ is nondecreasing, its limit exists. We will say that the solution is *regular* if $\lim_{n\to\infty} \tau_{\mathcal{Z}(n)} = \infty$ a.s. Any regular solution can be uniquely (a.s.) extended for all $t \geq t_0$.

Any solution $x(t)$ of (8) satisfies the following *strong Markov property* [24]:

$$\begin{aligned}&\mathbf{P}\{x(t+\tau, t_0, \overline{x}) \in A\}\\ &\quad = \int \mathbf{P}\{\tau \in ds;\ x(\tau, t_0, \overline{x}) \in dz\}\mathbf{P}\{x(t+\tau, s, z) \in A\},\end{aligned} \tag{11}$$

where τ is any given Markov time (relative to the σ-algebra generated by $\{x(s), s \leq t\}$). In (11) we can substitute $\mathbf{P}\{\cdot\}$ with its conditioned version $\mathbf{P}\{\cdot|\cdot\}$ as long as $\mathbf{P}\{\cdot|\cdot\}$ is regular, i.e., it is a function $p(\omega, A)$, measurable for each fixed A and a probability for each fixed ω [24].

From now on, we will denote $x(t, t_0, \overline{x})$, if not otherwise stated, simply by $x(t)$. Given a C^2 (measurable) function $V : \mathbb{R}^n \to \mathbb{R}$, define

$$\mathcal{L}V(x) = \frac{\partial V}{\partial x}(x) f(x,t) + \frac{1}{2}\mathbf{Tr}\left\{g^T(x,t)\frac{\partial^2 V}{\partial x^2}(x) g(x,t)\right\}. \tag{12}$$

Proposition 2.1 (Dynkin's formula). *Let $\overline{x} \in \mathcal{Z}$ a.s. The solution $x(t)$ of (8) satisfies on $[t_0, \tau_{\mathcal{Z},T}(t)]$ the following equation:*

$$\mathbf{E}\{V(x(\tau_{\mathcal{Z},T}(t)))\} - V(\overline{x}) = \mathbf{E}\left\{\int_{t_0}^{\tau_{\mathcal{B},T}(t)} \mathcal{L}V(x(s))ds\right\}. \tag{13}$$

The integral appearing in the right-hand side of (13) is meant in the sense that

$$\int_{t_0}^{\tau_{\mathcal{Z},T}(t)} \mathcal{L}V(x(s))ds = \int_{t_0}^{t} \xi_{\tau_{\mathcal{Z},T} > t}\mathcal{L}V(x(s))ds,$$

where $\xi_{\tau_{\mathcal{Z},T}>t}$ is the indicator function corresponding to the event $\{\tau_{\mathcal{Z},T} > t\}$.

Also, we will use extensively the following (generalized) *Čebyšev inequality*:

$$\mathbf{P}\{\eta \notin \mathcal{S}\} \leq \frac{\mathbf{E}\{V(\eta)\}}{\inf_{s\in\mathbb{R}^n\setminus\mathcal{S}}\{V(s)\}}, \tag{14}$$

where $\mathcal{S} \subset \mathbb{R}^n$, $V(\cdot)$ is real nonnegative, and η is a given random variable such that $\mathbf{E}\{V(\eta)\}$ exists. Finally, we recall the following fundamental formula of the differential calculus.

Proposition 2.2 (Itô's rule). *Given a C^2 function $\varphi : \mathbb{R}^n \to \mathbb{R}$ and if $x(t)$ is a solution of (8), then*

$$d\varphi(x(t)) = \frac{\partial \varphi}{\partial x}(x(t))dx(t) + \frac{1}{2}\mathbf{Tr}\left\{g^T(x(t),t)\frac{\partial^2\varphi}{\partial x^2}(x(t))g(x(t),t)\right\}dt. \tag{15}$$

3. Main motivations. To understand some key points of the forthcoming results, we illustrate some known facts under the assumptions that Φ, H, and K are bounded by a linear function. In particular, we consider the problem of stabilizing (3) under the assumption that $H = \sum_{j=1}^{s} H_j x$ and

$$\|\Phi(t,u,x)\|^2 \le \frac{\|x\|_E^2 + \|u\|_{R_1}^2}{\gamma^2} \tag{16}$$

for all x, u, and t and for some $E \in \mathcal{SSP}^n$, $R_1 \in \mathcal{SP}^m$, and $\gamma > 0$.

First, assume that the state vector x is available for feedback. Assume also the existence of $P_{SF}, Q_{SF} \in \mathcal{SP}^n$ such that

$$A^T P_{SF} + P_{SF} A + P_{SF}\left(\frac{1}{\gamma^2} B_1 B_1^T - B_2 R_1^{-1} B_2^T\right) P_{SF} + E + \sum_{j=1}^{s} H_j^T P_{SF} H_j = -Q_{SF}. \tag{17}$$

Let $V_{SF}(x) = \|x\|_{P_{SF}}^2$ and $F = -R^{-1} B_2^T P_{SF}$. Let us pretend that Φ is an "external" disturbance such that $\int_0^\infty \mathbf{E}\{\|\Phi\|^2\} dt < \infty$. Along the trajectories of the system

$$dx = (Ax + B_2 u + B_1 \Phi)dt + H dw \tag{18}$$

by (17)

$$\begin{aligned} &\mathcal{L}V_{SF} + \|x\|_E^2 + \|u\|_{R_1}^2 - \gamma^2 \|\Phi\|^2 \\ &\qquad = \|u - Fx\|_{R_1}^2 - \gamma^2 \|\Phi - \Phi_*\|^2 - \|x\|_{Q_{SF}}^2, \end{aligned} \tag{19}$$

where $\Phi_* = \frac{1}{\gamma^2} B_1^T P_{SF} x$ is the *worst-case disturbance* [14], since it maximizes the left-hand part of (19).

By taking the expectations in (19) and integrating between $[0, \infty)$, it follows that the $\mathcal{L}_2$ gain of the closed-loop system (18), with $u = Fx$, from any Φ to

$$z = \sqrt{\|x\|_E^2 + \|u\|_{R_1}^2}, \tag{20}$$

is *less than or equal to* γ [14] or, in other words, the admissible controller $u = Fx$ attains for (18) a *guaranteed level of attenuation* γ (in terms of the *expectation of energy*) of the effect of Φ over the "cost" z. On the other hand, since $\|x\|_E^2 + \|u\|_{R_1}^2 \ge \gamma^2 \|\Phi(t,u,x)\|^2$ *for all* t, u, x by (16), it follows from (19) that the trajectories of (3), with $u = Fx$, tend to zero as $t \to \infty$ in the quadratic mean and, therefore, in probability.

When the state vector x is not available for feedback, we should replace x by some estimate. To this aim, assume $B_1 C_1^T = 0$, $K = \sum_{j=1}^{s} K_j x$ the existence of $R_2 \in \mathcal{SP}^p$, with $R_2 \ge C_1 C_1^T$, and $P_m, Q_m \in \mathcal{SP}^n$ such that

$$\begin{aligned} &\left(A + \frac{1}{\gamma^2} B_1 B_1^T P_{SF}\right)^T P_m + P_m \left(A + \frac{1}{\gamma^2} B_1 B_1^T P_{SF}\right) + \frac{1}{\gamma^2} P_m B_1 B_1^T P_m \\ &\qquad + F^T R_1 F - \gamma^2 C_2^T R_2^{-1} C_2 = -Q_m \end{aligned} \tag{21}$$

and

$$Q_{SF} - \sum_{j=1}^{s} (H_j - GK_j)^T P_m (H_j - GK_j) > 0, \tag{22}$$

where $G = \gamma^2 P_m^{-1} C_2^T R_2^{-1}$.

Along the trajectories of

$$\begin{aligned} dx &= (Ax + B_2 u + B_1 \Phi)dt + H dw, \\ d\sigma &= \left(\left(A + \frac{1}{\gamma^2} B_1 B_1^T P_{SF} - G C_2 \right) \sigma + B_2 u \right) dt + G dy, \\ dy &= (C_2 x + C_1 \Phi)dt + K dw, \end{aligned} \tag{23}$$

with $G = \gamma^2 P_m^{-1} C_2^T R_2^{-1}$ and $e = x - \sigma$, we obtain

$$\begin{aligned} \mathcal{L}V_m + \|Fe\|_{R_1}^2 - \gamma^2 \|\Phi - \Phi_*\|^2 = &- \|e\|_{Q_m}^2 + \sum_{j=1}^{s} x^T (H_j - GK_j)^T P_m (H_j - GK_j) x \\ &- \gamma^2 \|\Phi - \Phi_*^e\|^2, \end{aligned} \tag{24}$$

where $\Phi_*^e = \frac{1}{\gamma^2}(B_1 - GC_1)^T P_m e + \frac{1}{\gamma^2} B_1^T P_{SF} x$. Summing up (19) and (24),

$$\begin{aligned} &\mathcal{L}V_{SF} + \mathcal{L}V_m + \|x\|_E^2 + \|u\|_{R_1}^2 - \gamma^2 \|\Phi\|^2 \\ &\quad = \|u - Fx\|_{R_1}^2 - \|Fe\|_{R_1}^2 - \|x\|^2_{Q_{SF} - \sum_{j=1}^{s} (H_j - GK_j)^T P_m (H_j - GK_j)} - \|e\|_{Q_m}^2 \\ &\qquad - \gamma^2 \|\Phi - \Phi_*^e\|^2. \end{aligned} \tag{25}$$

Note that Φ_*^e is the *worst-case disturbance* for (23) since it maximizes the left-hand part of (25).

By taking the expectations in (25) and integrating between $[0, \infty)$, it follows that the $\mathcal{L}_2$ gain of (23), with $u = F\sigma$, from any Φ such that $\int_0^\infty \mathbf{E}\{\|\Phi\|^2\} dt < \infty$ to (20) is *less than or equal to* γ [14]. Since $\|x\|_E^2 + \|u\|_{R_1}^2 \geq \gamma^2 \|\Phi(t,u,x)\|^2$ *for all* t, u, x by (16), it follows from (25) that the trajectories of (3) together with the admissible controller

$$\begin{aligned} u &= F\sigma, \\ d\sigma &= \left(\left(A + \frac{1}{\gamma^2} B_1 B_1^T P_{SF} - G C_2 \right) \sigma + B_2 u \right) dt + G dy \end{aligned} \tag{26}$$

tend to zero as $t \to \infty$ in the quadratic mean and, therefore, in probability [1], [14].

Besides asymptotic stability in the quadratic mean, it is possible to require some optimal performances. In particular, assume that $K = 0$ and that $C_1^T C_1$ is nonsingular, and define the following cost functional:

$$J = \lim_{T \to \infty} \int_{t_0}^{T} \sum_{j=1}^{2} \mathbf{E}\{\mathcal{W}_j(t, u(t), x(t), e(t))\} dt, \tag{27}$$

where

$$\begin{aligned} \mathcal{W}_1(t,u,x,e) &= \widetilde{\mathcal{W}}_1(t,u,x,e) + \gamma^2 \|\Phi(t,u,x) - \Phi_*^e\|^2, \\ \widetilde{\mathcal{W}}_1(t,u,x,e) &= -\gamma^2 \|\Phi(t,u,x)\|^2 + \|x\|_E^2 + \|u\|_{R_1}^2, \\ \mathcal{W}_2(t,u,x,e) &= \|e\|_{Q_m(k)}^2 + \|x\|^2_{Q_{SF} - \sum_{j=1}^{s} (H_j - GK_j)^T P_m (H_j - GK_j)}. \end{aligned} \tag{28}$$

Note that the term $\mathcal{W}_1(t,x,e)$ penalizes the distance of Φ from a *worst-case situation*, with respect to which the $\mathcal{H}_\infty$ controller is designed, and the term $\mathcal{W}_2(t,x,e)$ represents a *parametric quadratic cost*. By using standard arguments, it is possible to prove

that $F = -R_1^{-1}B_2^T P_{SF}$ and $G = \gamma^2 P_m^{-1} C_2^T R_2^{-1}$ characterize the optimal controller within the class (26). Indeed, from (25) and with $V^e = V_{SF} + V_m$ and $x^e = (x, \sigma)$

$$\mathcal{L}V^e(x^e) + \sum_{j=1}^{2} \mathcal{W}_j(t, u, x, e) = 0. \tag{29}$$

From Dynkin's formula and (29)

$$\mathbf{E}\{V^e(x^e(t))\} - V^e(x_0^e) = -\int_{t_0}^{t} \mathbf{E}\left\{\sum_{j=1}^{2} \mathcal{W}_j(t, u(t), x(t), e(t))\right\} ds. \tag{30}$$

Differentiating both sides of (30), since V^e is a quadratic function of x and e and by (22) and (28) $\sum_{j=1}^{2} \mathcal{W}_j(t, u, x, e)$ is lower bounded by a quadratic function of x and e, we obtain for some $\widetilde{\lambda}, \widetilde{\nu} > 0$

$$\mathbf{E}\{V^e(x^e(t))\} \leq \widetilde{\lambda} V^e(x_0^e) e^{-\widetilde{\nu}(t-t_0)}. \tag{31}$$

We conclude that $\mathbf{E}\{V^e(x^e(t))\} \to 0$ as $t \to \infty$ and from (30), with $t \to \infty$, $J = V^e(x_0)$.

Since

$$\mathcal{L}V^e + \sum_{j=1}^{2} \mathbf{E}\{\mathcal{W}_j(t, u, x, e)\} \geq 0 \tag{32}$$

for any other F and G, we conclude that J achieves its minimum with $F = -R_1^{-1}B_2^T P_{SF}$ and $G = \gamma^2 P_m^{-1} C_2^T R_2^{-1}$ [1], [14].

In the case that Φ, H, and K contain some *nonlinear term* as, for example, in

$$\begin{aligned} dx_1(t) &= x_2(t)dt, \\ dx_2(t) &= (x_2^3(t) + u(t))dt + x_1^2(t)dw, \\ y(t) &= x_1(t), \end{aligned} \tag{33}$$

(16) cannot be satisfied for all t, u, and x by some *constant* E, and known results in the literature do not apply. However, (16) can be still satisfied on some compact set of the state space for some E, provided the trajectories are ensured to stay in this compact set (and eventually approach a given target set) at least with some guaranteed probability. These requirements are well represented by the notion of semiglobal stability in probability, introduced in [3]. Moreover, since in (33) both *additive and multiplicative noise* affects the system, the cost functional J is no longer suitable, and a *time average* version of that functional should be considered (for the state feedback case, see the work of [15]). All of these facts will be discussed in detail in the next section.

4. Problem formulation. Let us consider nonlinear stochastic systems Σ of the form (3), where $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, $y(t) \in \mathbb{R}^p$, $w(t)$ is an s-dimensional Wiener process, and $\Phi(t, u(t), x(t)) \in \mathbb{R}^r$ represents model uncertainties and nonlinearities. The discussion in section 3 suggests considering as the class of *candidate controllers* $\{\mathcal{C}(k)\}$

$$\begin{aligned} u &= \eta(F(k)\sigma), \\ d\sigma &= (L(k)\sigma + B_2 u)dt + G(k)dy, \quad \sigma \in \mathbb{R}^n, \end{aligned} \tag{34}$$

with

$$L(k) = A + \frac{1}{\gamma^2(k)} B_1 B_1^T P_{SF}(k) - G(k) C_2 \tag{35}$$

for some sequence of real positive extended numbers $\{\Delta(k)\}$, $k \in \mathbb{R}^+$, matrices $\{F(k)\}$ and $\{G(k)\}$, positive numbers $\{\gamma(k)\}$, and symmetric positive definite matrices$\{P_{SF}(k)\}$ and $\eta : \mathbb{R}^m \to \mathbb{R}^m$ any C^0 function such that

$$\|\eta(s)\| \le \Delta(k) \quad \forall s, \tag{36}$$

$$\eta(s) = s, \qquad \|s\| \le s_0, \tag{37}$$

for some $s_0 > 0$. In other words, any candidate controller is the composition of a *linear controller* with a *static nonlinearity* η, which is *bounded by* $\Delta(k)$ (unbounded if $\Delta(k) = \infty$) and *it is the identity function near the origin*. While $\eta(s)$ is designed in such a way to counteract the destabilizing effects due to large values of $G(k)$ (peaking), $\Delta(k)$ accounts for possible limitations on the control u (as an example, saturations of the control actuators).

For the stability analysis of the closed-loop system (3)–(34), we also define a class of *candidate Lyapunov functions* $\{V_k^e\}$

$$V_k^e(x, \sigma) = \|x\|^2_{P_{SF}(k)} + \varphi(\|x - \sigma\|^2_{P_m(k)}), \tag{38}$$

where $\{P_m(k)\}$ is a sequence in $\mathcal{SP}^n$ and $\varphi : \mathbb{R}^{\ge} \to \mathbb{R}$ is any (at least) C^2, positive definite and proper function such that

$$\frac{\partial^2 \varphi}{\partial s^2}(s) \le 0 < \frac{\partial \varphi}{\partial s}(s) \le 1 \tag{39}$$

for all $s \ge 0$. Conditions (39) imply that over any compact set containing the origin any *candidate Lyapunov function is bounded from below and above by a quadratic function*, and, as will be clear in the next sections, the function φ is instrumental in enlarging the region of attraction of the closed-loop system.

Next we define some *admissibility constraints* for the noise coefficients H and K and for the uncertainty term Φ. To this aim, define the following compact sets:

$$\begin{aligned} \Omega(k) &= \{x \in \mathbb{R}^n : \|x\|^2_{P_{SF}(k)} \le k\}, \\ \mathcal{U}_{\Delta(k)} &= \{u \in \mathbb{R}^m : \|u\| \le \Delta(k)\}. \end{aligned} \tag{40}$$

Let $\{E(k)\}$, $\{R_1(k)\}$, and $\{c_1(k)\}$ be sequences in $\mathcal{SSP}^n$, $\mathcal{SP}^m$, and $\mathbb{R}^{\ge}$, respectively. Define

$$\begin{aligned} \Phi_*^e &= \frac{1}{\gamma^2(k)} \left[B_1^T P_{SF}(k) x + \frac{\partial \varphi}{\partial s}\Big|_{s=\|e\|^2_{P_m(k)}} (B_1 - G(k) C_1)^T P_m(k) e \right], \\ \widetilde{\mathcal{P}}_1(t, u, x, e, k) &= -\gamma^2(k) \|\Phi(t, u, x)\|^2 + \|x\|^2_{E(k)} + \|u\|^2_{R_1(k)} + c_1(k), \\ \mathcal{P}_1(t, u, x, e, k) &= \widetilde{\mathcal{P}}_1(t, u, x, e, k) + \gamma^2 \|\Phi(t, u, x) - \Phi_*^e\|^2, \end{aligned} \tag{41}$$

and let $\mathcal{F}(k)$ be the class of C^0 functions $\Phi : \mathbb{R} \times \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^r$ such that $\widetilde{\mathcal{P}}_1(t, u, x, e, k) \ge 0$ for all t, $u \in \mathcal{U}_{\Delta(k)}$, $x \in \Omega(k)$, and $e \in \mathbb{R}^n$. Note that Φ_*^e is the uncertainty Φ which maximizes $\mathcal{L}V_k^e - \gamma^2 \|\Phi\|^2$ along the trajectories of

$$\begin{aligned} dx &= (Ax + B_2 \eta(F(k)\sigma) + B_1 \Phi) dt + H dw, \\ d\sigma &= (L(k)\sigma(k) + B_2 \eta(F(k)\sigma)) dt + G(k) dy, \end{aligned} \tag{42}$$

or, in other words, it is the *worst-case uncertainty* for the closed-loop system (see section 3). Note also that $\mathcal{P}_1$ penalizes the distance of Φ from being the worst-case uncertainty Φ_*^e and, at the same time, from being a linear (parametrized by k) function of x and u over the sets $\Omega(k)$ and $\mathcal{U}_{\Delta(k)}$, which represent a worst-case situation with respect to which the matrices $F(k)$ and $G(k)$ are designed according to an $\mathcal{H}_\infty$ strategy.

Let $\{\widehat{H}_j(k)\}$, $j = 1, \dots, s$, be a sequence in $\mathbb{R}^{n\times n}$, and let $\{c_2(k)\}$ be a sequence in $\mathbb{R}^{\geq}$. Define

$$\begin{aligned}\mathcal{P}_2(t,u,x,e,k) = &-\mathbf{Tr}\{H^T(t,x)P_{SF}(k)H(t,x))\}\\ &+\sum_{j=1}^{s} x^T\widehat{H}_j^T(k)P_{SF}(k)\widehat{H}_j(k)x + c_2(k).\end{aligned} \tag{43}$$

Let $\mathcal{H}(k)$ be the class of C^0 functions $H : \mathbb{R}\times\mathbb{R}^n \to \mathbb{R}^{n\times s}$ such that $\mathcal{P}_2(t,u,x,e,k) \geq 0$ for all $t \geq 0$, $u \in \mathcal{U}_{\Delta(k)}$, $x \in \Omega(k)$, and $e \in \mathbb{R}^n$. Note that $\mathcal{P}_2$ penalizes the distance of $\mathbf{Tr}\{H^T(t,x)P_{SF}(k)H(t,x))\}$ from a sum of quadratic functions over the sets $\Omega(k)$ and $\mathcal{U}_{\Delta(k)}$. Since $\Omega(k)$ and $\mathcal{U}_{\Delta(k)}$ are compact sets, the admissibility constraints on $\widetilde{\mathcal{P}}_1$ and $\mathcal{P}_2$ can be always met whenever Φ and H are *locally Lipschitz*, uniformly with respect to t (compare with [1]).

Let $\{Q_m(k)\}$ and $\{c_3(k)\}$ be sequences in $\mathcal{SP}^n$ and $\mathbb{R}^{\geq}$, respectively, and let

$$M(t,x,k) = H(t,x) - G(k)K(t,x). \tag{44}$$

Define $\mathcal{K}(k)$ and $\mathcal{D}(k)$ as the class of C^0 functions $K : \mathbb{R}\times\mathbb{R}^n \to \mathbb{R}^{p\times s}$ and, respectively, the class of pairs of C^0 functions $\eta : \mathbb{R}^m \to \mathbb{R}^m$ and $\varphi : \mathbb{R}^{\geq} \to \mathbb{R}$ satisfying (36), (37), and (39) and such that

$$\begin{aligned}\mathcal{P}_3(t,u,x,e,k) = &-\|\eta(F(k)(x-e)) - F(k)x\|^2_{R_1(k)} + \|x\|^2_{Q_{SF}(k)} + c_3(k)\\ &+\frac{\partial\varphi}{\partial s}\Big|_{s=\|e\|^2_{P_m(k)}}[\|F(k)e\|^2_{R_1(k)} + \|e\|^2_{Q_m(k)}\\ &-\mathbf{Tr}\{M^T(t,x,k)P_m(k)M(t,x,k))\}] > 0\end{aligned} \tag{45}$$

for all $(x,e) \in (\Omega(k)\times\mathbb{R}^n)\backslash(0,0)$. Note that $c_1(k)$, $c_2(k)$, and $c_3(k)$ take into account additive noise and nonzero equilibrium points.

Finally, let

$$\begin{aligned}&\mathcal{P}_4(t,u,x,e,k)\\ &\quad= \frac{1}{\gamma^2(k)}\frac{\partial\varphi}{\partial s}\Big|_{s=\|e\|^2_{P_m(k)}}\Bigg[\left(1-\frac{\partial\varphi}{\partial s}\Big|_{s=\|e\|^2_{P_m(k)}}\right)\\ &\qquad\qquad\times e^TP_m(k)(B_1-G(k)C_1)(B_1-G(k)C_1)^TP_m(k)e\\ &\qquad\qquad+e^TP_m(k)G(k)(R_2(k)-C_1C_1^T)G^T(k)P_m(k)e\Bigg]\\ &\qquad-\frac{\partial^2\varphi}{\partial s^2}\Big|_{s=\|e\|^2_{P_m(k)}} e^TP_m(k)M(t,x,k)M^T(t,x,k)P_m(k)e\end{aligned} \tag{46}$$

for some sequence $\{R_2(k)\}$ in $\mathcal{SP}^p$ such that $R_2(k) \geq C_1C_1^T$. Note that, by (39), (46) is nonnegative for all $e \in \mathbb{R}^n$, and if $C_1C_1^T$ is nonsingular, then we can take directly

$R_2(k) = C_1 C_1^T$. Note also that $\mathcal{P}_4$ penalizes the distance from the situation for which φ is *linear* (i.e., quadratic Lyapunov functions) and $R_2(k) = C_1 C_1^T$.

In what follows, we will refer to $\Phi \in \mathcal{F}(k)$, $H \in \mathcal{H}(k)$, $K \in \mathcal{K}(k)$, and $(\eta, \varphi) \in \mathcal{D}(k)$ as *admissible functions*. Moreover, any choice of $\{P_{SF}(k)\}$, $\{P_m(k)\}$, $\{Q_{SF}(k)\}$, $\{Q_m(k)\}$, $\{R_1(k)\}$, $\{R_2(k)\}$, $\{\gamma(k)\}$, $\{\Delta(k)\}$, $\{E(k)\}$, $\{c_j(k)\}$, $j = 1, 2, 3$, $\{\widehat{H}_j(k)\}$, $j = 1, \ldots, s$, for which $\Phi \in \mathcal{F}(k)$, $H \in \mathcal{H}(k)$, and $K \in \mathcal{K}(k)$ will be referred to as *admissible parametrization*.

Denote by $x_k^e(t, t_0, x_0^e) = col(x_k(t, t_0, x_0^e), \sigma_k(t, t_0, x_0^e))$ the trajectory of the closed-loop system $\Sigma \circ \mathcal{C}(k)$ at time $t \geq t_0$ stemming from $x_0^e = col(x_0, \sigma_0)$. With some abuse of notation, wherever there is no ambiguity, we will use $x_k^e(t)$ instead of $x_k^e(t, t_0, x_0^e)$. Moreover, let $e_k(t) = x_k(t) - \sigma_k(t)$.

Our goal is to find an admissible parametrization which minimizes the expectation of the sum of the $\mathcal{P}_j$'s: this corresponds to achieving *semiglobal inverse optimality* with respect to a "reference" system for which a worst-case linear controller can be designed. At the same time, local optimality is guaranteed since the closed-loop system behaves locally as its reference system. To this aim, we introduce two sequences of cost functionals $\{J_h(k)\}$, $h = 1, 2$, defined as follows:

$$J_1(k) = \lim_{T \to \infty} \frac{1}{T - t_0} \int_{t_0}^{T} \mathbf{E} \left\{ \sum_{j=1}^{4} \mathcal{P}_j(t, u(t), x_k(t), e_k(t), k) \right\} dt \tag{47}$$

and

$$J_2(k) = \lim_{T \to \infty} \int_{t_0}^{T} \mathbf{E} \left\{ \sum_{j=1}^{4} \mathcal{P}_j(t, u(t), x_k(t), e_k(t), k) \right\} dt. \tag{48}$$

Note that $J_h(k) \geq 0$, $h = 1, 2$, for any $\Phi \in \mathcal{F}(k)$, $H \in \mathcal{H}(k)$, $K \in \mathcal{K}(k)$, and $(\eta, \varphi) \in \mathcal{D}(k)$. While $J_1(k)$ is more suitable in the case of both *additive and multiplicative noise*, $J_2(k)$ is not suitable for the case of *additive noise*, since the constant $c_j(k) \neq 0$ for at least one j would cause $J_2(k)$ to diverge.

The aim of this paper is to study under which conditions it is possible to modify the behavior of (3) in such a way that $J_1(k)$ achieves a guaranteed value (resp., $J_2(k)$ achieves its minimum) and to obtain stability in some "stochastic" sense. To make the last point precise, let us give the following definition.

DEFINITION 4.1. *Let $\alpha, \beta \in [0, 1)$ and $\Omega^e, \mathcal{B}^e \subset \mathbb{R}^{2n}$ be compact sets. The system (3) is said to be $(\Omega^e, \mathcal{B}^e, \alpha, \beta)$-stabilizable in probability (or $(\Omega^e, \mathcal{B}^e, \alpha, \beta)$-SP) if there exist a sequence of admissible control laws $\{\mathcal{C}(k)\}$, a sequence of compact sets $\{\Omega^e(k)\}$, and open sets $\{\mathcal{B}^e(k)\}$ of $\mathbb{R}^{2n}$ such that*

(i) $\liminf_{k \to \infty} \Omega^e(k) \supset \Omega^e \supset \mathcal{B}^e \supseteq \limsup_{k \to \infty} \mathcal{B}^e(k)$;

(ii) *for each $\delta > 0$ and $\Phi \in \mathcal{F}(k)$,*

$$\liminf_{k \to \infty} \inf_{x_0^e \in \overline{\mathcal{B}}^e(k)} \mathbf{P}\{x_k^e(t) \in \overline{\mathcal{B}}_\delta^e \ \forall t \geq t_0\} \geq 1 - \beta; \tag{49}$$

(iii) *for each $\delta > 0$ and $\Phi \in \mathcal{F}(k)$,*

$$\begin{aligned} \liminf_{k \to \infty} \inf_{x_0^e \in \Omega^e \setminus \mathcal{B}^e(k)} \mathbf{P}\{ & x_k^e(t) \in \Omega^e(k) \ \forall t \geq t_0 \\ & \text{and } x_k^e(t + \tau_{\mathbb{R}^{2n} \setminus \mathcal{B}^e(k)}) \in \mathcal{B}_\delta^e \ \forall t \geq 0 \\ & \text{and } \tau_{\mathbb{R}^{2n} \setminus \mathcal{B}^e(k)} < \infty\} \geq (1 - \alpha)(1 - \beta). \end{aligned} \tag{50}$$

Note that the events in (49) and (50) are measurable by separability and measurability (on the product σ-algebra $\mathcal{F} \times \mathcal{I}$, where $\mathcal{I}$ is the Borel σ-algebra on the line) of the process $x_k^e(t)$ and by being $\{\tau_{\mathbb{R}^{2n} \setminus \mathcal{B}^e(k)} \leq t\}$ adapted to the σ-algebra generated by $\{x_k^e(s),\ s \leq t\}$.

The set Ω^e gives the *guaranteed region of attraction* of the closed-loop system $\Sigma \circ \mathcal{C}(k)$, while $\mathcal{B}^e$ represents the *target set*. From (i) it follows that there exists k° such that $\Omega^e(k) \supset \Omega^e \supset \mathcal{B}^e \supset \mathcal{B}^e(k) \subset \mathcal{B}_\delta^e$ for all $k \geq k^\circ$. Property (ii) is a *local* property with respect to $\mathcal{B}^e$: for each δ-neighborhood of $\mathcal{B}^e$, there exists sufficiently large k° for which the probability that the trajectories $x_k^e(t)$ of the closed-loop system $\Sigma \circ \mathcal{C}(k)$, starting from $\overline{\mathcal{B}}^e(k)$, stay forever in $\overline{\mathcal{B}}_\delta^e$ is at least $1-\beta$ for all $k \geq k^\circ$. Property (iii) is a property *in the large* with respect to Ω^e: there exists sufficiently large k° for which the trajectories of $\Sigma \circ \mathcal{C}(k)$ starting inside Ω^e remain inside $\Omega^e(k)$, eventually enter any given δ-neighborhood of the target set $\mathcal{B}^e$ in finite time, and remain therein with probability at least $(1-\alpha)(1-\beta)$ for all $k \geq k^\circ$. The numbers α and β are given *risk margins: the first one quantifies the risk of leaving the compact set $\Omega^e(k)$ with initial condition in Ω^e rather than getting close to the target, while the second one gives a risk margin for remaining close to the target*. Note also that (iii) requires that $\tau_{\mathbb{R}^{2n} \setminus \mathcal{B}^e(k)} < \infty$. As will be clear in the next section, under the standard assumptions of local existence and uniqueness a.s. of trajectories, each Markov time $\tau_{\mathbb{R}^{2n} \setminus \mathcal{B}^e(k)}$, conditioned to $x_k(t) \in \Omega^e(k)$ for all $t \geq t_0$, is always finite and $\tau_{\mathbb{R}^{2n} \setminus \mathcal{B}^e(k)} \to \infty$ as $k \to \infty$ as long as $\limsup_{k \to \infty} \mathcal{B}^e(k) = \{0\}$. In particular, this implies that, if $\mathcal{B}^e = \{0\}$, the trajectory approaches the origin as $t \to \infty$.

The roles of the risk margins, region of attraction, and target set are peculiar of our setup and become unessential in the classical definitions given in [13]. If $\mathcal{B}^e = \{0\}$, $\alpha = \beta = 0$, and $\{\mathcal{C}(k)\} = \mathcal{C}$ for all k, Definition 4.1 recovers the classical definition of *asymptotic stability in probability* [13]. If in addition $\Omega^e = \mathbb{R}^{2n}$, Definition 4.1 gives the notion of *asymptotic stability in probability in the large* [13]. On the other hand, if $\mathcal{B}^e = \{0\}$ and Ω^e can be taken to be any a priori given compact set of $\mathbb{R}^{2n}$ and α and β any a priori given numbers in $[0, 1)$, our definition gives a stochastic analogue of the concept of *semiglobal stabilization*, as introduced in [4]. If $\Omega^e = \mathbb{R}^{2n}$ and $\mathcal{B}^e$ can be taken to be any a priori given compact set of $\mathbb{R}^{2n}$ and α and β any a priori given number in $[0, 1)$, Definition 4.1 extends to a stochastic setting the concept of *practical stabilization*.

All of the above remarks can be straightforwardly extended to the definition of stability in the quadratic mean.

We are ready to formulate our problems.

Problem I: Nonlinear stabilization in probability with guaranteed cost. Let $\Phi \in \mathcal{F}(k)$, $H \in \mathcal{H}(k)$, $K \in \mathcal{K}(k)$, $\mathcal{B}^e \subset \Omega^e$ be compact sets of $\mathbb{R}^{2n}$, $\alpha, \beta \in [0, 1)$, $x_0^e \in \Omega^e$ and $\{\overline{\omega}(k)\}$ a given sequence in $\mathbb{R}^{\geq}$. Find an admissible parametrization and $(\eta, \varphi) \in \mathcal{D}(k)$ such that

- (*guaranteed cost*) along the trajectories of the closed-loop systems $\Sigma \circ \mathcal{C}(k)$,

$$\liminf_{k \to \infty} \mathbf{Pr}\{J_1(k) \leq \overline{\omega}(k)\} \geq (1 - \alpha); \tag{51}$$

- (*stability*) $\Sigma \circ \mathcal{C}(k)$ is $(\Omega^e, \mathcal{B}^e, \alpha, \beta)$-stable in probability.

Problem II: Nonlinear stabilization in quadratic mean with optimality. Let $\Phi \in \mathcal{F}(k)$, $H \in \mathcal{H}(k)$, $K \in \mathcal{K}(k)$, $\mathcal{B}^e \subset \Omega^e$ be compact sets of $\mathbb{R}^{2n}$, $\alpha, \beta \in [0, 1)$, $x_0^e \in \Omega^e$. Find an admissible parametrization and $(\eta, \varphi) \in \mathcal{D}(k)$ such that

- (*optimality*) along the trajectories of the closed-loop systems $\Sigma \circ \mathcal{C}(k)$,

$$\liminf_{k\to\infty} \mathbf{Pr}\{J_2(k) \le \widetilde{J}_2(k)\} \ge (1-\alpha), \tag{52}$$

  where $\widetilde{J}_2(k)$ is the value of $J_2(k)$ corresponding to any other admissible parametrization;
- (*stability*) $\Sigma \circ \mathcal{C}(k)$ is $(\Omega, \{0\}, \alpha, 0)$-stable in the quadratic mean.

5. Main results. Let

$$\begin{aligned} H(t,x) &= (\, H_1(t,x) \cdots H_s(t,x)\,), \\ K(t,x) &= (\, K_1(t,x) \cdots K_s(t,x)\,), \end{aligned} \tag{53}$$

and, without loss of generality, assume that $B_1 C_1^T = 0$ and $H(t,x)K^T(t,x) = 0$ for all x and t.

THEOREM 5.1. *Assume that there exist an admissible parametrization and* $(\eta, \varphi) \in \mathcal{D}(k)$ *such that*

- (state feedback (SF))

$$\begin{aligned} A^T P_{SF}(k) + P_{SF}(k)A + \frac{1}{\gamma^2(k)} P_{SF}(k) B_1 B_1^T P_{SF}(k) + E(k) \\ - F^T(k) R_1(k) F(k) \\ + \sum_{j=1}^{s} \widehat{H}_j^T(k) P_{SF}(k) \widehat{H}_j(k) = -Q_{SF}(k), \end{aligned} \tag{54}$$

  where

$$F(k) = -R_1^{-1}(k) B_2^T P_{SF}(k); \tag{55}$$

- (output injection (OI))

$$\begin{aligned} P_m(k) \left(A + \frac{1}{\gamma^2(k)} B_1 B_1^T P_{SF}(k) \right) \\ + \left(A + \frac{1}{\gamma^2(k)} B_1 B_1^T P_{SF}(k) \right)^T P_m(k) + F^T(k) R_1 F(k) \\ + \frac{1}{\gamma^2(k)} P_m(k) B_1 B_1^T P_m(k) - \gamma^2(k) C_2^T R_2^{-1}(k) C_2 = -Q_m(k); \end{aligned} \tag{56}$$

- (risk margins (RM)) *if*

$$\Omega^e(k) = \{(x,\sigma) \in \mathbb{R}^n \times \mathbb{R}^n : V_k^e(x,\sigma) \le k\} \tag{57}$$

  and $\{\mathcal{B}^e(k)\}$, *a sequence of open sets of* $\mathbb{R}^{2n}$, *are such that*

$$\limsup_{k\to\infty} \mathcal{B}^e(k) \subseteq \mathcal{B}^e \subset \Omega^e \subset \liminf_{k\to\infty} \Omega^e(k), \tag{58}$$

  then for each $\delta > 0$,

$$\limsup_{k\to\infty} \sup_{(x,\sigma)\in\Omega^e \setminus \mathcal{B}^e(k)} \frac{V_k^e(x,\sigma)}{k} \le \alpha, \tag{59}$$

$$\limsup_{k\to\infty} \sup_{(x,\sigma)\in\partial\mathcal{B}^e(k)} \frac{V_k^e(x,\sigma)}{\inf_{(s_1,s_2)\in\mathbb{R}^{2n}\setminus\overline{\mathcal{B}}_\delta^e} V_k^e(s_1,s_2)} \le \beta, \tag{60}$$

and

$$\sum_{j=1}^{4} \mathcal{P}_j(t, \eta(F(k)\sigma), x, x-\sigma, k) - \sum_{j=1}^{3} c_j(k) \geq Q_k^e(x,\sigma) \tag{61}$$

for all t *and* $(x,\sigma) \in \Omega^e(k)$ *and for some sequence of quadratic* C^0 *positive definite functions* $\{Q_k^e\}$.

Under the above assumptions, the controller (34) *with* $F(k)$ *as in* (55) *and*

$$G(k) = \gamma^2(k) P_m^{-1}(k) C_2^T R_2^{-1}(k) \tag{62}$$

solves problem I *with* $\overline{\omega}(k) = \sum_{j=1}^{3} c_j(k)$. *If, in addition,* $c_j(k) = 0$ *for all* $j = 1,2,3$, *and* $K(t,x) = 0$ *for all* t, x, *and* $j = 1, \ldots, r$, *the same controller* (34) *solves problem* II.

Proof. Throughout the proof, unless otherwise stated, we will omit k and the arguments of Φ, K, and H. Moreover, we can assume $k \geq k^*$, where k^* is such that $\Omega^e(k) \supseteq \Omega^e \supset \mathcal{B}^e(k)$ for all $k \geq k^*$. (This is always possible by (58).)

Let $V_{SF}(x) = \|x\|^2_{P_{SF}}$ and V_k^e, M and e as in section 4. The closed-loop system is

$$\begin{aligned} dx &= \left(\left(A + \frac{1}{\gamma^2} B_1 B_1^T P_{SF} \right) x + B_2 u + B_1 \widetilde{\Phi} \right) dt + H dw, \\ de &= (Le + (B_1 - GC_1)\widetilde{\Phi}) dt + M dw, \end{aligned} \tag{63}$$

where $\widetilde{\Phi} = \Phi - \frac{1}{\gamma^2} B_1^T P_{SF} x$ and $u = \eta(F\sigma)$.

By (12)

$$\begin{aligned} \mathcal{L}\varphi &= \frac{\partial \varphi}{\partial e} \left[Le + (B_1 - GC_1)\widetilde{\Phi} \right] + \frac{1}{2} \mathbf{Tr} \left\{ M^T \frac{\partial^2 \varphi}{\partial e^2} M \right\}, \\ \mathcal{L}V_{SF} &= 2x^T P_{SF} \left[\left(A + \frac{1}{\gamma^2} B_1 B_1^T P_{SF} \right) x + B_2 u + B_1 \widetilde{\Phi} \right] + \mathbf{Tr}\{H^T P_{SF} H\}. \end{aligned} \tag{64}$$

Moreover,

$$\frac{\partial^2 \varphi}{\partial e^2} = 2 \frac{\partial^2 \varphi}{\partial s^2}\Big|_{s=\|e\|^2_{P_m}} P_m e e^T P_m + 2 \frac{\partial \varphi}{\partial s}\Big|_{s=\|e\|^2_{P_m}} P_m. \tag{65}$$

Since $HK^T = 0$ and $\mathbf{Tr}(AB) = \mathbf{Tr}(BA)$, by (65) one has

$$\begin{aligned} &\frac{1}{2} \mathbf{Tr} \left\{ M^T \frac{\partial^2 \varphi}{\partial e^2} M \right\} \\ &\quad = \frac{\partial \varphi}{\partial s}\Big|_{s=\|e\|^2_{P_m}} \mathbf{Tr}\{M^T P_m M\} + \frac{\partial^2 \varphi}{\partial s^2}\Big|_{s=\|e\|^2_{P_m}} e^T P_m M M^T P_m e. \end{aligned} \tag{66}$$

Since $B_1 C_1^T = 0$ and $HK^T = 0$, using (56) and (66), we have

$$
\begin{aligned}
&\mathcal{L}\varphi + \frac{\partial\varphi}{\partial s}\Big|_{s=\|e\|^2_{P_m}} \|Fe\|^2_{R_1} - \gamma^2\|\widetilde{\Phi}\|^2 \\
&\quad = \frac{\partial\varphi}{\partial s}\Big|_{s=\|e\|^2_{P_m}} [2e^T P_m(Le + (B_1 - GC_1)\widetilde{\Phi}) + \|Fe\|^2_{R_1}] - \gamma^2\|\widetilde{\Phi}\|^2 \\
&\qquad + \frac{\partial\varphi}{\partial s}\Big|_{s=\|e\|^2_{P_m}} \mathbf{Tr}\{M^T P_m M\} + \frac{\partial^2\varphi}{\partial s^2}\Big|_{s=\|e\|^2_{P_m}} e^T P_m M M^T P_m e \\
&\quad = \frac{\partial\varphi}{\partial s}\Big|_{s=\|e\|^2_{P_m}} e^T[P_m L + L^T P_m + F^T R_1 F]e - \gamma^2\|\Phi - \Phi^e_*\|^2 \\
&\qquad + \frac{1}{\gamma^2} e^T P_m \left(\frac{\partial\varphi}{\partial s}\Big|_{s=\|e\|^2_{P_m}}\right)^2 (B_1 B_1^T + GC_1 C_1^T G^T) P_m e \\
&\qquad + \frac{\partial^2\varphi}{\partial s^2}\Big|_{s=\|e\|^2_{P_m}} e^T P_m M M^T P_m e + \frac{\partial\varphi}{\partial s}\Big|_{s=\|e\|^2_{P_m}} \mathbf{Tr}\{M^T P_m M\} \\
&\quad = \frac{\partial\varphi}{\partial s}\Big|_{s=\|e\|^2_{P_m}} \Bigg[e^T \left(P_m L + L^T P_m + F^T R_1 F + \frac{1}{\gamma^2} P_m (B_1 B_1^T + G R_2 G^T) P_m \right) e \\
&\qquad\qquad + \mathbf{Tr}\{M^T P_m M\} \Bigg] - \gamma^2\|\Phi - \Phi^e_*\|^2 - \mathcal{P}_4 \\
&\quad = -\|u - Fx\|^2_{R_1} + \|x\|^2_{P_{SF}} + \frac{\partial\varphi}{\partial s}\Big|_{s=\|e\|^2_{P_m}} \|Fe\|^2_{R_1} - \mathcal{P}_1 + \widetilde{\mathcal{P}}_1 - \mathcal{P}_3 - \mathcal{P}_4 + c_3. \qquad (67)
\end{aligned}
$$

Moreover, for all u, by completing the square and using (54)

$$
\mathcal{L}V_{SF} = \|u - Fx\|^2_{R_1} - \gamma^2\|\widetilde{\Phi}\|^2 - \|x\|^2_{Q_{SF}} - \widetilde{\mathcal{P}}_1 - \mathcal{P}_2 + c_1 + c_2. \qquad (68)
$$

Summing up together (67) and (68), we conclude that

$$
\mathcal{L}V^e_k + \sum_{j=1}^{4} \mathcal{P}_j = \overline{\omega} = \sum_{j=1}^{3} c_j. \qquad (69)
$$

To prove our theorem, we are left with proving the following facts:

- $J_1(k) \le \overline{\omega}(k)$ (resp., $J_2(k)$ achieves its minimum), conditionally to the event $\{x^e_k(t) \in \Omega^e(k),\ t \ge t_0\}$;
- $\liminf_{k\to\infty} \mathbf{Pr}\{(x^e_k(t) \in \Omega^e(k),\ t \ge t_0\} \ge 1 - \alpha$, i.e., the event $\{x^e_k(t) \in \Omega^e(k),\ t \ge t_0\}$ has a guaranteed probability $1 - \alpha$ for sufficiently large k;
- $\Sigma \circ \mathcal{C}(k)$ is $(\Omega^e, \mathcal{B}^e, \alpha, \beta)$-stable in probability ($\Sigma \circ \mathcal{C}(k)$ is $(\Omega^e, \{0\}, \alpha, 0)$-stable in quadratic mean, respectively).

First, we prove that $J_1(k) = \overline{\omega}(k)$ (resp., $J_2(k)$ achieves its minimum), conditionally to the event $\{x^e_k(t) \in \Omega^e(k),\ t \ge t_0\}$. By (69) and Dynkin's formula, for each $T > t_0$

$$
\begin{aligned}
&\frac{1}{T - t_0} \int_{t_0}^{T} \mathbf{E}\left\{ \sum_{j=1}^{4} \mathcal{P}_j(s, u, x^e_k(s), k) \right\} ds \\
&\qquad = \frac{1}{T - t_0}(V^e_k(x^e_0) - \mathbf{E}\{V^e_k(x^e_k(T))\}) + \overline{\omega}(k) \qquad (70)
\end{aligned}
$$

for all $x^e_0 \in \Omega^e(k)$. From (70), letting $T \to \infty$ and since $\sum_{j=1}^{4} \mathcal{P}_j \ge 0$ by admissibility and $V^e_k \ge 0$, we obtain $0 \le J_1(k) = \overline{\omega}(k)$. If, in addition, $c_j(k) = 0$ for all $j = 1, 2, 3$ and $K(t, x) = 0$ for all t, x, it is easy to see that (70) holds with $\overline{\omega}(k) = 0$.

As a consequence of the above facts, $x_k^e(t)$ being defined for all $t \geq t_0$ a.s., by Dynkin's formula

$$\mathbf{E}\{V_k^e(x_k^e(t))\} - V_k^e(x_0^e) = \int_{t_0}^{t} \mathbf{E}\{\mathcal{L}V_k^e(x_k^e(s))\}ds. \tag{71}$$

Differentiating both sides of (71), taking into account that V_k^e is lower bounded by a quadratic function of x and e, from (61) and (69) we obtain for some sequences $\{\widetilde{\lambda}(k)\}, \{\widetilde{\nu}(k)\}$ of positive numbers

$$\mathbf{E}\{V_k^e(x_k^e(t))\} \leq \widetilde{\lambda}(k)V_k^e(x_0^e)e^{-\widetilde{\nu}(k)(t-t_0)} \tag{72}$$

conditionally to the event $\{x_k^e(t) \in \Omega^e(k),\ t \geq t_0\}$. We conclude that $\mathbf{E}\{V_k^e(x_k^e(t))\} \to 0$ as $t \to \infty$ and $J_2(k) = V_k^e(x_0)$.

Since

$$\mathcal{L}V_k^e + \sum_{j=1}^{4} \mathcal{P}_j \geq 0 \tag{73}$$

for any other $F(k)$ and $G(k)$, we conclude that $J_2(k)$ achieves its minimum with $F(k)$ and $G(k)$ as in (55) and (62), respectively.

Using (69) and (RM), the $(\Omega^e, \mathcal{B}^e, \alpha, \beta)$ stability of $\Sigma \circ \mathcal{C}(k)$ is a consequence of the following lemma, which is proved in the appendix. If, in addition, $c_j = 0$ for all j and $K(t,x) = 0$ for all t and x, from (72) we conclude also the $(\Omega^e, \{0\}, \alpha, 0)$ stability in quadratic mean of $\Sigma \circ \mathcal{C}(k)$. From (124) (see the appendix), we also infer that $\liminf_{k\to\infty} \mathbf{Pr}\{x_k^e(t) \in \Omega^e(k),\ t \geq t_0\} \geq 1 - \alpha$. □

LEMMA 5.1. *The system* (3) *is* $(\Omega^e, \mathcal{B}^e, \alpha, \beta)$*-SP if there exist a sequence of admissible control laws* $\{\mathcal{C}(k)\}$*, a sequence of (at least)* C^2*, positive definite and proper functions* $\{V_k^e(x^e)\}$*, a sequence of* C^0*, positive definite functions* $\{Q_k^e(x^e)\}$*, and open sets* $\{\mathcal{B}^e(k)\}$*,* $\mathcal{B}^e(k) \subset \mathbb{R}^{2n}$*, containing the origin, such that*

(iv) $\liminf_{k\to\infty} \Omega^e(k) \supset \Omega^e \supset \mathcal{B}^e \supseteq \limsup_{k\to\infty} \mathcal{B}^e(k)$ *and* $\Omega^e(k) \supset \mathcal{B}^e(k)$ *for all* k*, where*

$$\Omega^e(k) = \{z \in \mathbb{R}^{2n} : V_k^e(z) \leq k\};$$

(v) $\mathcal{L}V_k^e(x^e) \leq -Q_k^e(x^e)$ *for all* k*,* t*,* $\Phi \in \mathcal{F}(k)$*, and* $x^e \in \Omega^e(k) \backslash \mathcal{B}^e(k)$;

(vi) $\limsup_{k\to\infty} \sup_{x^e \in \Omega^e \backslash \mathcal{B}^e(k)} \frac{V_k^e(x^e)}{k} \leq \alpha$ *and*

$$\limsup_{k\to\infty} \sup_{x^e \in \partial\mathcal{B}^e(k)} \frac{V_k^e(x^e)}{\inf_{z \in \mathbb{R}^{2n} \backslash \overline{\mathcal{B}}_\delta^e} V_k^e(z)} \leq \beta$$

for each $\delta > 0$.

Remark 5.1. We note that, as a consequence of (59), if

$$\limsup_{k\to\infty} \frac{V_k^e(x^e)}{k} = 0 \tag{74}$$

for each x^e, then the risk margin α can be taken to be *any* number in $[0, 1)$ and any a priori given compact set can be included in Ω^e. Thus, condition (74), together with (iv)–(vi) of Lemma 5.1, guarantee *semiglobal stabilization in probability.*

On the other hand, if $\limsup_{k\to\infty} \mathcal{B}^e(k) = \{0\}$, then for each $\delta > 0$

$$\limsup_{k\to\infty} \sup_{x^e \in \partial\mathcal{B}^e(k)} \frac{V_k^e(x^e)}{\inf_{z\in\mathbb{R}^{2n}\setminus\overline{\mathcal{B}}_\delta^e} V_k^e(z)} = 0 \tag{75}$$

and the risk margin β can be taken to be *any* number in $[0, 1)$. Moreover, any a priori given compact set can be chosen as target set, and condition (75) together with (iv)–(vi) of Lemma 5.1 guarantee *practical stabilization in probability*.

Remark 5.2. The proof of Lemma 5.1 is based on a *probabilistic invariance property* which extends to a stochastic setup the following well-known property: if there exists a C^1 proper and positive definite function $V_k^e : \mathbb{R}^{2n} \to \mathbb{R}$ such that, along the trajectories $x_k^e(t, t_0, x_0^e)$ of $\Sigma \circ \mathcal{C}(k)$, $\dot{V}_k^e$ is definite negative on $\Omega^e(k)\backslash\mathcal{B}^e(k)$ and (iv) and (vi) hold, then any trajectory $x_k^e(t, t_0, x_0^e)$ starting from $\Omega^e \subseteq \Omega^e(k)$ stays forever in $\Omega^e(k)$, eventually enters any given δ-neighborhood of $\mathcal{B}^e$ in finite time, and remains therein. In our setting, this invariance property corresponds to an event which occurs with probability at least $(1-\alpha)(1-\beta)$. For the above reasons, α and β can be thought of as *risk margins*. In the deterministic case, (vi) corresponds to a precise geometric property of the level sets of V_k^e for sufficiently large k: one is that Ω^e is contained in $\Omega^e(k)$, and the other is that $\overline{\mathcal{B}}^e(k)$ is contained in some level set of V_k^e which is, in turn, contained in $\overline{\mathcal{B}}_\delta^e$.

Remark 5.3. On the other hand, conditions of Theorem 5.1 ensure that stability in probability in the sense of Definition 4.1 is achieved together with *robust performances* with respect to parameter variations and model uncertainties. Moreover, it is easily seen from the proof that the conditions of Theorem 5.1 guarantee properties (ii) and (iii) of Definition 4.1 to hold *uniformly with respect to* $\Phi \in \mathcal{F}(k)$.

Remark 5.4 (linear case). Consider the class of systems (3) with $H_j(t, x) = H_j x$ and $K_j(t, x) = K_j x$ for all $j = 1, \ldots, s$, x, and t and, in addition, with $\Phi(t, u, x)$ satisfying $\widetilde{\mathcal{P}}_1 \geq 0$ *for all* x, u, and t. Pick admissible $\eta(s) = s$ (i.e., *linear controllers*) and $\varphi(s) = s$ (i.e., *quadratic* Lyapunov functions). With our positions the admissibility constraint $\mathcal{P}_3 > 0$ boils down to the following *matrix inequality*:

$$Q_{SF}(k) > \sum_{j=1}^{s} (H_j - GK_j)^T P_m(k)(H_j - GK_j) \tag{76}$$

(see (22)). This recovers the stabilization results with the optimality of [14] and [1]. Moreover, if $H_j = 0$ and $K_j = 0$ for all j (i.e., deterministic case), then the constraint on $\mathcal{P}_3$ is trivially satisfied (see [2]).

6. Stochastic stabilization with guaranteed cost for feedback linearizable systems. The conditions of Theorem 5.1 do not provide any constructive procedure to find an admissible parametrization with the functions η and φ. In the next two sections, we want to outline algorithms for accomplishing this task for the following class of nonlinear stochastic systems:

$$\begin{aligned} dx &= (Ax + B(u + \Phi(t, u, x)))dt + Bh(t, x)dw, \\ y &= Cx \end{aligned} \tag{77}$$

with (A, B, C) invertible with no invariant zeros and $\Phi(t, u, x)$ and $h(t, x)$ norm-bounded from above by a locally Lipschitz function of x and u, uniformly with respect to t. Moreover, we will assume that $\Phi(t, 0, 0) = 0$ and $h(t, 0) = 0$ for all t. (The cases $\Phi(t, 0, 0) \neq 0$ or $h(t, 0) \neq 0$ can be treated in a similar way with heavier calculations.)

First, we give a *semiglobal in probability backstepping* design procedure for solving the state feedback problem (SF); then we give a recursive procedure to satisfy (OI) and (RM). We remark that Theorem 5.1 still holds if one replaces (54) with

$$\begin{aligned}\mathcal{L}V_{SF} &= \|u - F(k)x\|^2_{R_1(k)} - \gamma^2(k)\|\widetilde{\Phi}\|^2 - \|x\|^2_{Q_{SF}(k)} \\ &\quad - \widetilde{\mathcal{P}}_1(t,u,x,e,k) - \mathcal{P}_2(t,u,x,e,k) + c_1(k) + c_2(k)\end{aligned} \tag{78}$$

and

$$\widetilde{\mathcal{P}}_1(t,u,x,e,k) = -\gamma^2(k)\|\Phi(t,u,x)\|^2 + \|x\|^2_{E(k)} + \|u - \Gamma(k)x\|^2_{R_1(k)} + c_1(k) \tag{79}$$

for some sequence of matrices $\Gamma(k)$ and with $F(k) = -R_1^{-1}(k)B_2^T P_{SF}(k) + \Gamma(k)$. In order to keep the backstepping algorithm as simple as possible, it is convenient to satisfy (78) rather than (54). Moreover, the choice of $\Gamma(k)$ gives an additional flexibility in the optimal control design.

Preliminarly, by [10] there exists a change of coordinates $z = Zx$ such that (77) reads out in the new coordinates

$$\begin{aligned} dz &= [\widehat{A}z + \widehat{B}(u + \Phi(t,u,Z^{-1}z))]dt + Bh(t,Z^{-1}z)dw, \\ y &= \widehat{C}z, \end{aligned} \tag{80}$$

where

$$\widehat{A} = \begin{pmatrix} a_{11} & 1 & 0 & \cdots & 0 & 0 \\ a_{12} & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ a_{n-1,1} & 0 & 0 & \cdots & 0 & 1 \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{n,n-1} & a_{nn} \end{pmatrix},$$

$$\widehat{B} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix},$$

$$\widehat{C} = (\,1 \quad 0 \quad 0 \quad \cdots \quad 0 \quad 0\,).$$

To simplify notation, we will omit the hats and denote z by x.

6.1. Backstepping design. The main result of the section is the following.

THEOREM 6.1. *The system* (80) *is semiglobally stabilizable in quadratic mean with optimality through a linear state feedback controller.*

As a first step toward the proof of Theorem 6.1, rewrite (80) as

$$d\pi_0 = (A_0\pi_0 + B_0 x_n)dt, \tag{81}$$

$$dx_n = (\widetilde{f}_n(t,u,x) + u)dt + \widetilde{h}_n(t,x)dw \tag{82}$$

with $\pi_0 = \mathrm{col}(x_1,\ldots,x_{n-1})$. We will prove (78) directly on (81)–(82), and this amounts to introducing the following definition.

DEFINITION 6.1. *We will say that*

$$d\pi = (A(k)\pi + B_1(k)\Phi(t,u,\pi) + B_2(k)u)dt + H(t,\pi)dw \tag{83}$$

satisfies the property DI if there exist C^0 *functions* $P_{SF}, Q_{SF} : \mathbb{R}^+ \to \mathcal{SP}^n$, $R_1, \gamma : \mathbb{R}^+ \to \mathbb{R}^+$, $E : \mathbb{R}^+ \to \mathcal{SSP}^n$, *and* $\Gamma : \mathbb{R}^m \times \mathbb{R}^n$ *such that* $\Phi \in \mathcal{F}(k)$, $H \in \mathcal{H}(k)$, *and for all* $\pi \in \Omega(k) = \{\pi \in \mathbb{R}^n : V_{SF}(\pi) \le k\}$, u, *and* t *one has*

$$\|\Phi(t,u,\pi)\|^2 \le \frac{\|\pi\|^2_{E(k)} + \|u - \Gamma(k)\pi\|^2_{R_1(k)}}{\gamma^2(k)} \tag{84}$$

and

$$\begin{aligned} &\mathcal{L}V_{SF} + \|\pi\|^2_{E(k)} + \|u - \Gamma(k)\pi\|^2_{R_1(k)} - \gamma^2(k)\|\Phi(t,u,\pi)\|^2 \\ &- \mathbf{Tr}\{H^T(t,\pi)P_{SF}H(t,\pi)\} + \sum_{j=1}^{r} \pi^T \widehat{H}_j(k) P_{SF} \widehat{H}_j(k)\pi \\ &= -\|\pi\|^2_{Q_{SF}(k)} + \|u - F(k)\pi\|^2_{R_1(k)} - \gamma^2(k)\|\Phi(t,u,\pi) - \frac{1}{\gamma^2(k)} B_1^T(k) P_{SF}(k)\pi\|^2, \end{aligned} \tag{85}$$

where

$$V_{SF}(\pi) = \|\pi\|^2_{P_{SF}(k)}, \tag{86}$$

$$F(k) = -R_1^{-1}(k)B_2^T(k)P_{SF}(k) + \Gamma(k). \tag{87}$$

We have the following result, which roughly states that (81)–(82) satisfies the DI property in some new coordinates π.

LEMMA 6.1. *There exists a* C^0 *function* $\lambda : \mathbb{R}^+ \to (0,1)$ *and a change of coordinates*

$$\begin{aligned} \pi &= \begin{pmatrix} \pi_0 \\ \zeta \end{pmatrix}, \\ \zeta &= \lambda(k)(x_n - F_0\pi_0), \end{aligned} \tag{88}$$

such that (81)–(82) *reads out as*

$$d\pi = (A(k)\pi + B_1(k)\Phi(t,u,\pi) + B_2(k)u)dt + H(t,\pi)dw \tag{89}$$

and satisfies DI, with

$$\begin{aligned} \pi &= \begin{pmatrix} \pi_0 \\ \zeta \end{pmatrix}, \\ \zeta &= \lambda(k)(x_n - F_0\pi_0), \\ F_0 &= -R_0^{-1}B_0^T P_0, \\ A(k) &= \begin{pmatrix} A_0 + B_0F_0 & \dfrac{B_0}{\lambda(k)} \\ -\lambda(k)F_0(A_0 + B_0F_0) & -F_0B_0 \end{pmatrix}, \\ B_1(k) &= \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \\ B_2(k) &= \lambda(k)\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \\ H(t,\pi) &= \lambda(k)\begin{pmatrix} 0 \\ \widetilde{h}_n(t, \frac{\zeta}{\lambda(k)} + F_0\pi_0) \end{pmatrix}, \\ \Phi(t,u,\pi) &= \lambda(k)\begin{pmatrix} 0 \\ \widetilde{f}_n(t,u, \frac{\zeta}{\lambda(k)} + F_0\pi_0) \end{pmatrix} \end{aligned} \tag{90}$$

and $R_0 > 0$ *and* $P_0 \in \mathcal{SP}^{n-1}$ *such that*

$$A_0P_0 + P_0A_0^T - P_0B_0R_0^{-1}B_0^TP_0 + I = 0. \tag{91}$$

Proof. For simplicity, throughout the proof and whenever there is no ambiguity, we omit the arguments of the functions involved.

Pick a C^0 function $\widetilde{P} : \mathbb{R}^+ \to (0,1)$ such that $\lim_{k\to\infty} \widetilde{P}(k) = 0$, and define

$$\Omega(k) = \{(v_1, v_2) \in \mathbb{R}^{n-1} \times \mathbb{R} : \|v_1\|^2_{P_0} + \widetilde{P}(k)v_2^2 \le k\}. \tag{92}$$

We have by our assumptions on $\widetilde{h}_n$

$$\|\widetilde{h}_n\|^2 \le 2\left[\|\widetilde{h}_n - \widetilde{h}_n\Big|_{\zeta=0}\|^2 + \|\widetilde{h}_n\Big|_{\zeta=0}\|^2\right] \le \|\pi_0\|^2_{\widetilde{M}(k)} + \widetilde{N}(k)\zeta^2 \tag{93}$$

for all t, π, and u such that $\pi \in \Omega(k)$, where $\widetilde{h}_n|_{\zeta=0}$ denotes $\widetilde{h}_n$ evaluated for $\zeta = 0$ and for some C^0 functions $\widetilde{M} : \mathbb{R}^+ \to \mathcal{SSP}^{n-1}$ and $\widetilde{N} : \mathbb{R}^+ \to \mathbb{R}^{\ge}$. We remark that $\widetilde{M}$ can be chosen as function of P_0 only. Pick $\lambda : \mathbb{R}^+ \to (0,1)$ such that

$$\lambda^2(k)\widetilde{M}(k) < \frac{I}{2}. \tag{94}$$

By the Îto rule

$$d\zeta = (\lambda(k)(\widetilde{f}_n + u - \mathit{\Gamma}(k)\pi)dt + \widetilde{h}d\widetilde{w}), \tag{95}$$

where

$$\mathit{\Gamma}(k) = F_0\left(A_0 + B_0F_0 \qquad \frac{B_0}{\lambda(k)}\right).$$

Find C^0 functions $\widetilde{Q} : \mathbb{R}^+ \to \mathbb{R}^+$, $\gamma : \mathbb{R}^+ \to \mathbb{R}^+$, $\widetilde{R} : \mathbb{R}^+ \to \mathbb{R}^+$, $\widetilde{E}_1 : \mathbb{R}^+ \to \mathbb{R}^{\ge}$, and $\widetilde{E}_2 : \mathbb{R}^+ \to \mathcal{SSP}^{n-1}$ such that

- for all t, π, u such that $\pi \in \Omega(k)$, one has

$$\lambda^2(k)\widetilde{f}_n^2 \le \frac{\zeta^2\widetilde{E}_1(k) + \|\pi_0\|^2_{\widetilde{E}_2(k)} + |u - \mathit{\Gamma}(k)\pi|^2\widetilde{R}(k)}{\gamma^2(k)} \tag{96}$$

and

$$\widetilde{E}_2(k) < \frac{I}{2}; \tag{97}$$

- the following equality holds:

$$\frac{\widetilde{P}^2(k)}{\gamma^2(k)} - \frac{\lambda^2(k)\widetilde{P}^2(k)}{\widetilde{R}(k)} + \widetilde{E}_1(k) + \lambda^2(k)\widetilde{P}(k)\widetilde{N}(k) = -\widetilde{Q}(k) \tag{98}$$

with

$$\widetilde{Q}(k) > \frac{R_0}{\lambda^2(k)}. \tag{99}$$

Let $\widetilde{V}_k(\zeta) = \widetilde{P}(k)\zeta^2$, $\widetilde{F}(k) = -\lambda(k)\widetilde{R}^{-1}(k)\widetilde{P}(k)$. From (98)

$$\begin{aligned}
&\mathcal{L}\widetilde{V}_k + \|\pi_0\|^2_{\widetilde{E}_2(k)} + \zeta^2\widetilde{E}_1(k) + |u - \Gamma(k)\pi|^2\widetilde{R}(k) - \gamma^2(k)\lambda^2(k)\widetilde{f}_n^2 \\
&\quad + \lambda^2\widetilde{P}\left(-\|\widetilde{h}_n\Big|_{x_n=\frac{\zeta}{\lambda(k)}+F_0(k)\pi_0}\|^2 + \|\pi_0\|^2_{\widetilde{M}(k)} + \zeta^2\widetilde{N}(k)\right) \\
&= -\zeta^2\widetilde{Q}(k) + \|\pi_0\|^2_{\widetilde{E}_2(k)+\lambda^2(k)\widetilde{P}(k)\widetilde{M}(k)} + |u - \widetilde{F}(k)\zeta - \Gamma(k)\pi|^2_{\widetilde{R}(k)} \\
&\quad - \gamma^2(k)\left|\lambda(k)\widetilde{f}_n - \frac{1}{\gamma^2(k)}\widetilde{P}(k)\zeta\right|^2. \qquad (100)
\end{aligned}$$

Define

$$P_{SF}(k) = \begin{pmatrix} P_0 & 0 \\ 0 & \widetilde{P}(k) \end{pmatrix}. \qquad (101)$$

With our definitions

$$\mathbf{Tr}\{H^T P_{SF}(k) H\} = \lambda^2(k)\widetilde{P}(k)\|\widetilde{h}_n\Big|_{x_n=\frac{\zeta}{\lambda(k)}+F_0(k)\pi_0}\|^2. \qquad (102)$$

From (91), (94), (97), (99), (100), and (102), with $V_0(\pi_0) = \|\pi_0\|^2_{P_0}$ and $V_{SF}(\pi) = V_0(\pi_0) + \widetilde{P}(k)\zeta^2$, it follows that

$$\begin{aligned}
&\mathcal{L}V_0 + x_n^2 R_0 + \mathcal{L}\widetilde{V}_k + \|\pi_0\|^2_{\widetilde{E}_2(k)} + \zeta^2\widetilde{E}_1(k) + |u - \Gamma(k)\pi|^2\widetilde{R}(k) - \lambda^2(k)\gamma^2(k)\widetilde{f}_n^2 \\
&\quad + \lambda^2\widetilde{P}(-\|\widetilde{h}_n|_{x_n=\frac{\zeta}{\lambda(k)}+F_0(k)\pi_0}\|^2 + \|\pi_0\|^2_{\widetilde{M}(k)} + \zeta^2\widetilde{N}(k)) \\
&= \mathcal{L}V_{SF} + \|\pi\|^2_{E(k)} + |u - \Gamma(k)\pi|^2 R_1(k) - \gamma^2(k)\|\Phi\|^2 \\
&\quad - \mathbf{Tr}\{H^T(t,\pi)P_{SF}H(t,\pi)\} + \sum_{j=1}^{s}\pi^T\widehat{H}_j(k)P_{SF}\widehat{H}_j(k)\pi \\
&= -\|\pi\|^2_{Q_{SF}(k)} + |u - F(k)\pi|^2 R_1(k) - \gamma^2(k)\|\Phi - \frac{1}{\gamma^2(k)}B_1^T(k)P_{SF}(k)\pi\|^2 \qquad (103)
\end{aligned}$$

for a suitably defined $E(k)$, with $R_1(k) = \widetilde{R}(k)$ and

$$\widehat{H}_j(k) = \lambda(k)\begin{pmatrix} P_0^{-\frac{1}{2}}\sqrt{\widetilde{M}(k)\widetilde{P}(k)} & 0 \\ 0 & \sqrt{\widetilde{N}(k)} \end{pmatrix},$$

$$Q_{SF}(k) = \begin{pmatrix} I - \widetilde{E}_2(k) - \lambda^2(k)\widetilde{M}(k)\widetilde{P}(k) & 0 \\ 0 & \widetilde{Q}(k) - \dfrac{R_0}{\lambda^2(k)} \end{pmatrix}. \qquad (104)$$

This proves (85).

Finally, from (93), (96), and (102) it follows that $\widetilde{\mathcal{P}}_1 \geq 0$ and $\mathcal{P}_2 \geq 0$ for all $\pi \in \Omega(k)$. This proves the admissibility of Φ and H. □

We are ready to obtain Theorem 6.1. Note that each $\widetilde{P}(k)$ can be chosen in such a way that

$$\lim_{k\to\infty}\frac{\widetilde{P}(k)}{k} = 0, \qquad (105)$$

which implies

$$\lim_{k\to\infty} \frac{V_{SF}(\pi)}{k} = 0 \tag{106}$$

for each $\pi \in \mathbb{R}^n$. From (103), (106), and Theorem 5.1, with (54) replaced by (78), since $c_j(k) = 0$ for all j so that the sequence $\{\mathcal{B}^e(k)\}$ can be chosen arbitrarily, we conclude that (89) is *semiglobally stabilizable in quadratic mean* with $u = F(k)\pi$. Moreover, by the same arguments used in the proof of Theorem 5.1, it is easy to see that $J_2(k)$, with $u = F(k)\pi$, achieves its minimum (i.e., $J_2(k) = 0$).

Since F_0 is independent of k, (105) is sufficient to conclude also that the original system (80) is *semiglobally stabilizable in quadratic mean with optimality*, which proves Theorem 6.1.

If either $\Phi(t,0,0) \neq 0$ or $h(t,0) \neq 0$, we have the following result, which can be proved as Theorem 6.1.

THEOREM 6.2. *The system* (80) *is semiglobally practically stabilizable in quadratic mean with optimality through a linear state feedback controller.*

6.2. Filter design. In this section, we want to show that, for some admissible parametrization and $(\eta,\varphi) \in \mathcal{D}(k)$, (OI) and (RM) can also be met for (89), i.e., (81)–(82) in the new coordinates π. First, let us prove (OI).

Define

$$\begin{aligned} Q_m(k) &= 2\epsilon^2(k)P_m(k), \\ P_m(k) &= \widetilde{P}_m(\epsilon(k)) \\ &= \operatorname{diag}\{\epsilon^{2(n-1)}(k), \epsilon^{2(n-2)}(k), \dots, 1\} P_1(\epsilon(k)) \operatorname{diag}\{\epsilon^{2(n-1)}(k), \epsilon^{2(n-2)}(k), \dots, 1\}, \end{aligned} \tag{107}$$

where $\epsilon : \mathbb{R}^+ \to \mathbb{R}^+$ is a C^0 function (to be specified later) such that $\lim_{k\to\infty} \epsilon(k) = \infty$.

In what follows, for the sake of simplicity we will omit the argument k when there is no ambiguity. We claim that there exists a C^0 function $\epsilon_1^* : \mathbb{R}^+ \to \mathbb{R}^+$ such that Q_m and P_m, defined in (107), solve (OI) with $R_2 = \frac{\gamma^2}{\epsilon^{2(2n-1)}}$ for all C^0 functions $\epsilon \geq \epsilon_1^*$. Indeed, substituting in (OI), left- and right-multiplying by $\operatorname{diag}\{\epsilon^{-2(n-1)}, \epsilon^{-2(n-2)}, \dots, 1\}$ and dividing both members by ϵ^2, we find out that solving (OI) amounts to satisfying

$$\begin{aligned} &P_1(\epsilon)(J + I + S_1(\epsilon)) + (J + I + S_1(\epsilon))^T P_1(\epsilon) \\ &\qquad - C^T C + \frac{1}{\gamma^2\epsilon^2} P_1(\epsilon) B B^T P_1(\epsilon) + S_2(\epsilon) = 0, \end{aligned} \tag{108}$$

where $S_1, S_2 : \mathbb{R}^+ \to \mathbb{R}^{n\times n}$ are C^0 functions such that $\lim_{\epsilon\to\infty} S_j(\epsilon) = 0$, $j = 1,2$, $S_2(\epsilon)$ is symmetric and positive semidefinite, and

$$J = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}.$$

It can be shown that there exists $\epsilon_1^* : \mathbb{R}^+ \to \mathbb{R}^+$ such that (108) (and thus (OI)) holds for each $\epsilon \geq \epsilon_1^*$ and for some $P_1(\epsilon) \in \mathcal{SP}^n$. Indeed, since by continuity and for

sufficiently large ϵ $(J+I+S_1(\epsilon), B)$ is controllable and $(C, J+I+S_1(\epsilon))$ is observable, for each sufficiently large ϵ and for some $P_1^0(\epsilon) \in \mathcal{SP}^n$

$$P_1^0(\epsilon)(J+I+S_1(\epsilon))+(J+I+S_1(\epsilon))^T P_1^0(\epsilon)-C^TC+\frac{1}{\gamma^2\epsilon^2}P_1^0(\epsilon)BB^TP_1^0(\epsilon)=0. \tag{109}$$

Moreover, if $\epsilon \to \infty$, then $P_1^0(\epsilon) \to P_1^0(\infty)$. Finally, by standard arguments, for ϵ large enough the existence of some $P_1(\epsilon) \in \mathcal{SP}^n$ satisfying (108) and such that $P_1(\epsilon) > P_1^0(\epsilon)$ and $P_1(\epsilon) \to P_1^0(\infty)$ as $\epsilon \to \infty$ can be shown. This proves (OI).

Next we define an admissible pair (φ, η). Choose $\varphi(s) = \frac{1}{\epsilon}\ln(1+s)$ if $s \geq 0$,

$$\eta(s) = \begin{cases} s & \text{if } |s| \leq M, \\ \dfrac{s}{|s|}M & \text{otherwise,} \end{cases} \tag{110}$$

where $M = \max_{\pi\in\Omega}|F\pi|$, and F and Ω are as in (87) and (92). Note that $\frac{\partial^2\varphi}{\partial s^2} \leq 0 < \frac{\partial\varphi}{\partial s} \leq 1$ for all $s \geq 0$ and the function η is a bounded function, linear near the origin. Moreover, the pair (φ, η) is admissible if there exists a C^0 function $\epsilon_2^* : \mathbb{R}^+ \to \mathbb{R}^+$ such that

$$\mathcal{P}_3 = \frac{2\epsilon^2\|e\|^2_{\widetilde{P}_{m(\epsilon)}} + \|Fe\|^2_{R_1} - \|\widetilde{h}_n\|^2 P_{nn}}{\epsilon(1+\|e\|^2_{\widetilde{P}_{m(\epsilon)}})} - \|\eta(F(\pi-e))+F\pi\|^2_{R_1}+\|\pi\|^2_{Q_{SF}} > 0 \tag{111}$$

for all C^0 functions $\epsilon \geq \epsilon_2^*$ and $(\pi, e) \in (\Omega \times \mathbb{R}^n)\backslash(0,0)$, with P_{nn} being the (n,n) entry (independent of k) of P_m.

In order to prove (111), find a covering $\cup_{j=1}^3 \mathcal{M}_j$ of $\{(\pi, e) \in \Omega \times \mathbb{R}^n\}$, with

$$\begin{aligned} \mathcal{M}_1 &= \left\{(\pi, e) \in \Omega \times \mathbb{R}^n : \|\pi - e\| \leq \vartheta_1;\ \|e\| \leq \frac{\vartheta_1}{2}\right\}, \quad \vartheta_1 > 0, \\ \mathcal{M}_2 &= \{(\pi, e) \in \Omega \times \mathbb{R}^n : \|\pi - e\| \geq \vartheta_1;\ \|e\| \leq \vartheta_2\}, \qquad \vartheta_2 \leq \frac{\vartheta_1}{2}, \\ \mathcal{M}_3 &= \{(\pi, e) \in \Omega \times \mathbb{R}^n : \|e\| \geq \vartheta_2\}, \end{aligned} \tag{112}$$

Pick $\vartheta_1 > 0$ such that $\eta(F(\pi - e)) = F(\pi - e)$ for all $\|\pi - e\| \leq \vartheta_1$.

Note that, since $P_1(\epsilon) \to P_1^0(\infty)$ as $\epsilon \to \infty$, the term $-\frac{\|\widetilde{h}_n\|^2 P_{nn}}{\epsilon(1+\|e\|^2_{\widetilde{P}_{m(\epsilon)}})} + \|\pi\|^2_{Q_{SF}}$ can be rendered positive over the set Ω by choosing ϵ large enough.

First, it is easy to see that there exists a C^0 function $\epsilon_3^* : \mathbb{R}^+ \to \mathbb{R}^+$ such that (111) holds for all C^0 functions $\epsilon \geq \epsilon_3^*$ and $(\pi, e) \in \mathcal{M}_1$.

Moreover, (111) holds for all $(\pi, e) \in \mathcal{M}_2$ for some $\vartheta_2 \leq \frac{\vartheta_1}{2}$ and for all $k > 0$. Indeed, $(0, e) \notin \mathcal{M}_2$ since $\vartheta_1 > \vartheta_2$. Thus we have $\|\pi\|^2_{Q_{SF}} > 0$ on $\mathcal{M}_2$. It follows that for any such π by continuity there exists $e_\pi > 0$ such that (111) holds for all $\|e\| \leq e_\pi$ and for all C^0 functions $\epsilon \geq \epsilon_3^*$. Since $\mathcal{O} = \{\pi \in \Omega : \|\pi\| \geq \frac{\vartheta_1}{2}\}$ is compact and $\vartheta_1 > 0$, one can take $\vartheta_2 = \min\{\frac{\vartheta_1}{2}, \min_{\pi\in\mathcal{O}} e_\pi\}$.

We are left with proving that there exists $\epsilon_4^* : \mathbb{R}^+ \to \mathbb{R}^+$ such that (111) holds for all $(\pi, e) \in \mathcal{M}_3$ and for all C^0 functions $\epsilon \geq \epsilon_4^*$. On the other hand, this readily follows by the boundedness of η and since

$$\lim_{\epsilon\to\infty} \inf_{\|e\|\geq\vartheta_2} \frac{\epsilon\|e\|^2_{\widetilde{P}_{m(\epsilon)}}}{1+\|e\|^2_{\widetilde{P}_{m(\epsilon)}}} = \infty. \tag{113}$$

Pick $\epsilon_2^* \geq \max\{\epsilon_3^*, \epsilon_4^*\}$.

By similar arguments (111) can be satisfied in such a way that its left-hand part of (111) is lower bounded by a quadratic function of π and e.

Finally, we will show how to satisfy (RM). Since

$$\lim_{s\to 0} s \ln s^{-r} = 0 \quad \forall r \geq 0$$

by (106) and the definition of $P_m(k)$, there exists a C^0 function $\epsilon_5^* : \mathbb{R}^+ \to \mathbb{R}^+$ such that for all C^0 functions $\epsilon \geq \epsilon_5^*$

$$\lim_{k\to\infty} \frac{\|\pi\|^2_{P_{SF}(k)} + \dfrac{1}{\epsilon(k)} \ln(1 + \|e\|^2_{\widetilde{P}_m(\epsilon(k))})}{k} = 0 \tag{114}$$

for each $(\pi, e) \in \mathbb{R}^{2n}$, which proves (59). On the other hand, by properly choosing the sequence $\{\mathcal{B}^e(k)\}$, we can also satisfy (60). Moreover, since $\mathcal{P}_3$ is lower bounded by a quadratic function of π and e, (61) also holds true.

We conclude that (OI)–(RM) hold as long as $\epsilon : \mathbb{R}^+ \to \mathbb{R}^+$ is any C^0 function such that $\epsilon \geq \max\{\epsilon_1^*, \epsilon_2^*, \epsilon_5^*\}$. By using Theorem 6.1, with (54) replaced by (78), and the above results and since (77) can be put in the form of (80) after a linear change of coordinates, we obtain the following theorem.

THEOREM 6.3. *Any system* (77) *is semiglobally stabilizable in quadratic mean with optimality.*

If either $\Phi(t, 0, 0) \neq 0$ or $h(t, 0) \neq 0$, we have the following result, which can be proved as Theorem 6.3.

THEOREM 6.4. *Any system* (77) *is semiglobally practically stabilizable in quadratic mean with optimality.*

To conclude we will perform the main calculations for obtaining a stabilizing controller for the system (33). Pick $0 < P_0, \widetilde{P}(k) < 1$, $\lim_{k\to\infty} \frac{\widetilde{P}(k)}{k} = 0$, and $R_0 = P_0^2$, and define $\Omega(k) = \{(x_1, \zeta_2) : P_0 x_1^2 + \widetilde{P}(k)\zeta_2^2 \leq k\}$. Let $\widetilde{N}(k) = 0$, $\widetilde{M}(k) = \frac{k^2}{P_0^2}$, and choose $\lambda(k)$ such that $\frac{\lambda^2(k)k^2}{P_0^2} < \frac{1}{2}$. Define $\Gamma(k) = -\frac{P_0}{R_0}(-\frac{P_0}{R_0} \ \ \frac{1}{\lambda(k)})$. Moreover, let

$$\begin{aligned} \widetilde{E}_1(k) &= 8\gamma^2(k)k^2 \left(\frac{1}{\lambda^2(k)\widetilde{P}(k)} + \frac{P_0}{R_0^2}\right)^2, \\ \widetilde{E}_2(k) &= \frac{8\gamma^2(k)k^2 P_0^2 \lambda^2(k)}{R_0^2} \left(\frac{1}{\lambda^2(k)\widetilde{P}(k)} + \frac{P_0}{R_0^2}\right)^2, \end{aligned} \tag{115}$$

and $\gamma(k) > 0$ be such that $\widetilde{E}_2(k) < \frac{1}{2}$. Also, let $\widetilde{Q}(k), \widetilde{R}(k) > 0$ such that

$$\begin{aligned} \frac{\widetilde{P}^2}{\gamma^2(k)} - \frac{\lambda^2(k)\widetilde{P}^2(k)}{\widetilde{R}(k)} + \widetilde{E}_1(k) &= -\widetilde{Q}(k), \\ \widetilde{Q}(k) &> \frac{R_0}{\lambda^2(k)}. \end{aligned} \tag{116}$$

Moreover, both P_0 and $\widetilde{P}$ can be chosen in such a way that $\Omega(k)$ (and the region of attraction of the system in the original coordinates) contains an a priori given compact set. This concludes the backstepping design. As to the filter design, let $\epsilon(k) > 0$ be

such that $(J+S_1(\epsilon(k))+I, \binom{0}{1})$ is controllable, $((1 \quad 0), J+S_1(\epsilon(k))+I)$ is observable, $-\left(\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right) + S_2(\epsilon(k)) < 0$, and, correspondingly, find $P_1 \in \mathcal{SP}^2$ such that

$$\begin{aligned} &P_1(J + S_1(\epsilon(k)) + I) + (J + S_1(\epsilon(k)) + I)^T P_1 \\ &\qquad - \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \frac{1}{\gamma^2(k)\epsilon^2(k)} P_1 \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} P_1 + S_2(\epsilon(k)) = 0, \end{aligned} \tag{117}$$

where

$$S_1(\epsilon(k)) = \frac{1}{\epsilon^2(k)} \left[\begin{pmatrix} F^2(k) & 0 \\ -\dfrac{\lambda(k)F^2(k)}{\epsilon^2(k)} & -\dfrac{F(k)}{\lambda^2(k)} \end{pmatrix} + \frac{1}{\gamma^2(k)} \begin{pmatrix} 0 & 0 \\ 0 & \widetilde{P}(k) \end{pmatrix} \right],$$

$$S_2(\epsilon(k)) = \frac{1}{\epsilon^2(k)} \begin{pmatrix} \dfrac{1}{\epsilon^2(k)} & 0 \\ 0 & 1 \end{pmatrix} F^T(k) \widetilde{R}(k) F(k) \begin{pmatrix} \dfrac{1}{\epsilon^2(k)} & 0 \\ 0 & 1 \end{pmatrix}, \tag{118}$$

and $F(k) = -\frac{P_0}{R_0} + \Gamma(k)$. Finally, $\epsilon(k)$ can be chosen at the same time in such a way to satisfy (111), as pointed out above.

Appendix.

Proof of Lemma 5.1. Throughout the proof, $\tau_{\mathcal{S}}(t) = \min\{t, \tau_{\mathcal{S}}\}$, where $\tau_{\mathcal{S}}$ is the Markov time (relative to the σ-algebra generated by $\{x_k^e(s), s \le t\}$) defined as the first time at which the trajectory of (3) reaches the boundary of $\mathcal{S}$. By (iv) we can assume $k \ge k^*$, with k^* such that $\Omega^e(k) \supseteq \Omega^e$ for all $k \ge k^*$, and we fix any $\Phi \in \mathcal{D}_\Xi$.

We have to show only (ii) and (iii) of Definition 4.1. As a consequence of Dynkin's formula (with $\mathcal{Z} = \Omega^e(k) \backslash \mathcal{B}^e(k)$ and $T = \infty$), since $\mathcal{L}V_k^e$ is negative definite on $\Omega^e(k) \backslash \mathcal{B}^e(k)$,

$$\mathbf{E}\{V_k^e(x_k^e(\tau_{\Omega^e(k)\backslash\mathcal{B}^e(k)}(t), t_0, x_0^e))\} \le V_k^e(x_0^e) \tag{119}$$

for all $x_0^e \in \partial\mathcal{B}^e(k)$. By (iv) for each $\delta > 0$ there exists k° such that $\mathcal{B}_\delta^e \supset \mathcal{B}^e(k)$. By the Čebyšev inequality (with $\mathcal{S} = \overline{\mathcal{B}}_\delta^e$, $\eta = x_k^e(\tau_{\Omega^e(k)\backslash\mathcal{B}^e(k)}(r), t_0, x_0^e)$, $r \ge t_0$ ranging over the rationals, and $V = V_k^e$), (119), and (v) we have

$$\mathbf{P}\{x_k^e(r, t_0, x_0^e) \notin \overline{\mathcal{B}}_\delta^e \text{ for some rational } r \ge t_0\} \le \frac{V_k^e(x_0^e)}{\inf_{z \in \mathbb{R}^{2n} \backslash \overline{\mathcal{B}}_\delta^e} V_k^e(z)} \tag{120}$$

for all $k \ge k^\circ$ and $x_0^e \in \partial\mathcal{B}^e(k)$. Since, by $\mathcal{B}^e(k) \subset \mathcal{B}_\delta^e$ for all $k \ge k^\circ$,

$$\mathbf{P}\{x_k^e(r, t_0, x_0^e) \notin \overline{\mathcal{B}}_\delta^e \text{ for some rational } r \in [t_0, \tau_{\mathcal{B}^e(k)})\} = 0$$

for all $k \ge k^\circ$ and $x_0^e \in \mathcal{B}^e(k)$, by (vi) and the continuity of $x_k^e(t, \cdot, \cdot)$, we conclude (ii) of Definition 4.1.

To prove (iii) we implicitly assume that $x_0^e \in \Omega^e \backslash \mathcal{B}^e(k)$. First, we prove

$$\mathbf{P}\{\tau_{\Omega^e(k)\backslash\mathcal{B}^e(k)} < \infty\} = 1 \tag{121}$$

in the case that $x_k^e(t, t_0, x_0^e)$ is regular since otherwise it is trivially true. Since $Q_k^e(x^e)$ is continuous on its domain, we have $\mathcal{L}V_k^e \le -\nu(k) < 0$ for all $x^e \in \Omega^e(k) \backslash \mathcal{B}^e(k)$ and for some $\nu(k) > 0$. Directly from Dynkin's formula (with $\mathcal{Z} = \Omega^e(k) \backslash \mathcal{B}^e(k)$ and $T = \infty$) we obtain

$$\nu(k)\mathbf{E}\{\tau_{\Omega^e(k)\backslash\mathcal{B}^e(k)}(t) - t_0\} \le V_k^e(x_0^e). \tag{122}$$

Thus, by the Čebyšev inequality (with $\mathcal{S} = \mathbb{R}^{2n}\backslash(\Omega^e(k)\backslash\mathcal{B}^e(k))$, $\eta = x_k^e(\tau_{\Omega^e(k)\backslash\mathcal{B}^e(k)}(r), t_0, x_0^e)$, with $r \geq t_0$ ranging over the rationals, and $V(\eta) = \tau_{\Omega^e(k)\backslash\mathcal{B}^e(k)} - t_0$),

$$\mathbf{P}\{\tau_{\Omega^e(k)\backslash\mathcal{B}^e(k)} \geq r\} \leq \frac{V_k^e(x_0^e)}{\nu(k)(r-t_0)}. \tag{123}$$

Since $\frac{V_k^e(x_0^e)}{\nu(k)(r-t_0)} \to 0$ as $r \to \infty$, from (123) and the (sequential) continuity of $\mathbf{Pr}\{\cdot\}$, we obtain (121).

Next we show that

$$\liminf_{k\to\infty} \inf_{x_0^e \in \Omega^e\backslash\mathcal{B}^e(k)} \mathbf{P}\{\tau_{\mathbb{R}^{2n}\backslash\mathcal{B}^e(k)} < \tau_{\Omega^e(k)}\} \geq 1-\alpha. \tag{124}$$

From Dynkin's formula (with $\mathcal{Z}$ and T as above) and since $\mathcal{L}V_k^e$ is negative definite on $\Omega^e(k)\backslash\mathcal{B}^e(k)$, it follows that

$$\mathbf{E}\{V_k^e(x_k^e(\tau_{\Omega^e(k)\backslash\mathcal{B}^e(k)}(t), t_0, x_0^e))\} \leq V_k^e(x_0^e). \tag{125}$$

By (121)

$$\tau_{\Omega^e(k)\backslash\mathcal{B}^e(k)}(t) = \tau_{\Omega^e(k)\backslash\mathcal{B}^e(k)} \quad \text{a.s.} \tag{126}$$

From (14) and (125)

$$\mathbf{P}\{\tau_{\mathbb{R}^{2n}\backslash\mathcal{B}^e(k)} > \tau_{\Omega^e(k)}\} \leq \mathbf{P}\left\{\frac{V_k^e(x_k^e(\tau_{\Omega^e(k)\backslash\mathcal{B}^e(k)}, t_0, x_0^e))}{k} \geq 1\right\} \leq \frac{V_k^e(x_0^e)}{k}. \tag{127}$$

By (127) and since $\limsup_{k\to\infty} \sup_{x_0^e\in\Omega^e\backslash\mathcal{B}^e(k)} \frac{V_k^e(x_0^e)}{k} \leq \alpha$ by (vi), we get

$$\limsup_{k\to\infty} \sup_{x_0^e \in \Omega^e\backslash\mathcal{B}^e(k)} \mathbf{P}\{\tau_{\mathbb{R}^{2n}\backslash\mathcal{B}^e(k)} > \tau_{\Omega^e(k)}\} \leq \alpha. \tag{128}$$

By the continuity of $x_k^e(t,\cdot,\cdot)$ and since $\mathcal{B}^e(k) \subset \Omega^e(k)$, $\mathbf{P}\{\tau_{\mathbb{R}^{2n}\backslash\mathcal{B}^e(k)} = \tau_{\Omega^e(k)}\} = 0$, which along with (128) implies (124).

By (120) and (vi) and using the continuity of $x_k^e(t,\cdot,\cdot)$, for each $\epsilon, \delta > 0$ there exists k° such that $\mathcal{B}_\delta^e \supset \mathcal{B}^e(k)$ and

$$\mathbf{P}\{x_k^e(t,s,z) \notin \overline{\mathcal{B}}_\delta^e \text{ for some } t \geq s\} < \beta + \epsilon \tag{129}$$

for all $k \geq k^\circ$ and $z \in \partial\mathcal{B}^e(k)$.

Finally, let $\mathcal{F}(\Omega^e(k), \mathcal{B}^e(k))$ denote the σ-algebra generated by the events $\{x_k^e(r, t_0, x_0^e) \in \Omega^e(k), r \leq \tau_{\mathbb{R}^{2n}\backslash\mathcal{B}^e(k)}\}$. Since $\mathcal{F}(\Omega^e(k), \mathcal{B}^e(k))$ is a sub-σ-algebra of the one generated by $\{x_k^e(r, t_0, x_0^e), r \leq \tau_{\mathbb{R}^{2n}\backslash\mathcal{B}^e(k)}\}$ and by the strong Markov property, for each $\epsilon, \delta > 0$ and for all $k \geq k^\circ$

$$\begin{aligned}
&\mathbf{P}\{x_k^e(t + \tau_{\mathbb{R}^{2n}\backslash\mathcal{B}^e(k)}, t_0, x_0^e) \notin \overline{\mathcal{B}}_\delta^e \text{ for some } t \geq 0 | \mathcal{F}(\Omega^e(k), \mathcal{B}^e(k))\} \\
&\quad = \int_{t_0}^{\infty} \int_{z\in\partial\mathcal{B}^e(k)} (\mathbf{P}\{\tau_{\mathbb{R}^{2n}\backslash\mathcal{B}^e(k)} \in ds;\ x_k^e(\tau_{\mathbb{R}^{2n}\backslash\mathcal{B}^e(k)}, t_0, x_0^e) \in dz | \mathcal{F}(\Omega^e(k), \mathcal{B}^e(k))\} \\
&\quad \cdot \mathbf{P}\{x_k^e(t,s,z) \notin \overline{\mathcal{B}}_\delta^e \text{ for some } t \in [s,\infty) | \mathcal{F}(\Omega^e(k), \mathcal{B}^e(k))\}) < \beta + \epsilon,
\end{aligned} \tag{130}$$

which implies for each $\delta > 0$

$$\liminf_{k\to\infty} \inf_{x_0^e\in\Omega^e\backslash\mathcal{B}^e(k)} \mathbf{P}\{x_k^e(t+\tau_{\mathbb{R}^{2n}\backslash\mathcal{B}^e(k)}, t_0, x_0^e) \in \overline{\mathcal{B}}_\delta^e\ \forall t \geq 0 | \mathcal{F}(\Omega^e(k), \mathcal{B}^e(k))\} \geq 1-\beta. \tag{131}$$

Property (iii) of Definition 4.1 follows directly from (121), (131), (124), and the Bayes formula. □

REFERENCES

[1] V. A. Ugrinovskii and I. R. Petersen, *Absolute stabilization and minimax optimal control of uncertain systems with stochastic uncertainty*, SIAM J. Control Optim., 37 (1999), pp. 1089–1122.

[2] S. Battilotti, *A unifying framework for the semiglobal stabilization of nonlinear uncertain systems via measurement feedback*, IEEE Trans. Automat. Control, 44 (2001), pp. 3–17.

[3] S. Battilotti and A. De Santis, *A new notion of stabilization in probability for nonlinear stochastic systems*, in Proceeding of the International Symposium on Mathematical Theory on Networks and Systems, Perpignan, France, 2000.

[4] A. Bacciotti, *Further remarks on potentially global stabilizability*, IEEE Trans. Automat. Control, 34 (1989), pp. 637–638.

[5] J. C. Doyle, K. Glover, P. P. Khargonekar, and B. A. Francis, *State space solutions to standard $\mathcal{H}_2$ and $\mathcal{H}_\infty$ control problems*, IEEE Trans. Automat. Control, 34 (1989), pp. 831–847.

[6] H. J. Sussmann and P. V. Kokotovic, *The peaking phenomenon and the global stabilization of nonlinear systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 424–439.

[7] P. Kokotovic and R. Marino, *On vanishing stability regions in nonlinear systems with high gain feedback*, IEEE Trans. Automat. Control, 31 (1986), pp. 967–970.

[8] F. Esfandiari and H. K. Khalil, *Output feedback stabilization of fully linearizable systems*, Internat. J. Control, 56 (1992), pp. 1007–1037.

[9] H. K. Khalil and F. Esfandiari, *Semiglobal stabilization of a class of nonlinear system using output feedback*, IEEE Trans. Automat. Control, 38 (1993), pp. 1412–1415.

[10] A. Saberi, Z. Lin, and A. Teel, *Control of linear systems with saturating actuators*, IEEE Trans. Automat. Control, 41 (1996), pp. 368–378.

[11] F. Mazenc, L. Praly, and W. P. Dayawansa, *Global stabilization by output feedback: Examples and counterexamples*, IEEE Trans. Automat. Control, 22 (1994), pp. 119–125.

[12] A. R. Teel and L. Praly, *Tools for semiglobal stabilization by partial state and output feedback*, SIAM J. Control Optim., 33 (1995), pp. 1443–1488.

[13] R. Z. Khas'minskii, *Stochastic Stability of Differential Equations*, Sjithoff and Noordhoff, Germantown, MD, 1980.

[14] D. Hinrichsen and A. J. Pritchard, *Stochastic $\mathcal{H}_\infty$*, SIAM J. Control Optim., 36 (1998), pp. 1504–1538.

[15] Z. Pan and T. Başar, *Backstepping controller design for nonlinear stochastic systems under a risk-sensitive cost criterion*, SIAM J. Control Optim., 37 (1999), pp. 957–995.

[16] H. Deng and M. Krstic, *Output feedback stochastic nonlinear stabilization*, IEEE Trans. Automat. Control, 44 (1999), pp. 328–333.

[17] H. Deng and M. Krstic, *Stochastic nonlinear stabilization.* I. *A backstepping design*, Systems Control Lett., 32 (1997), pp. 143–150.

[18] H. Deng and M. Krstic, *Stochastic nonlinear stabilization.* II. *Inverse optimality*, Systems Control Lett., 32 (1997), pp. 151–159.

[19] M. Krstic and H. Deng, *Stabilization of Nonlinear Uncertain Systems*, Springer-Verlag, London, 1998.

[20] P. Florchinger, *Lyapunov-like techniques for stochastic stability*, SIAM J. Control Optim., 33 (1995), pp. 1151–1169.

[21] P. Florchinger, *A universal formula for the stabilization of control stochastic differential equations*, Stochastic Anal. Appl., 11 (1993), pp. 155–162.

[22] M. Zakai, *On the optimal filtering of diffusion processes*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 11 (1969), pp. 230–243.

[23] E. Wong and B. Hajek, *Stochastic Processes in Engineering Systems*, Springer-Verlag, New York, 1984.

[24] I. I. Gihman and A. V. Skorohod, *Stochastic Differential Equations*, Springer-Verlag, New York, 1972.

[25] B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications*, Springer-Verlag, New York, 1985.

[26] P. Kokotovic and R. Marino, *On vanishing stability regions in nonlinear systems with high gain feedback*, IEEE Trans. Automat. Control, 31 (1986), pp. 967–970.

SIAM J. CONTROL OPTIM.
Vol. 40, No. 6, pp. 1965–1982

LINEAR PROGRAMMING FORMULATION FOR OPTIMAL STOPPING PROBLEMS*

MOON JUNG CHO† AND RICHARD H. STOCKBRIDGE†

Abstract. Optimal stopping problems for continuous time Markov processes are shown to be equivalent to infinite-dimensional linear programs over a space of pairs of measures under very general conditions. The measures involved represent the joint distribution of the stopping time and stopping location and the occupation measure of the process until it is stopped. These measures satisfy an identity for each function in the domain of the generator which is sufficient to characterize the stochastic process. Finite-dimensional linear programs obtained using Markov chain approximations are solved in two examples to illustrate the numerical accuracy of the linear programming formulation.

Key words. linear programming, optimal stopping, occupation measures

AMS subject classifications. 60G40, 93E20

PII. S0363012900377663

1. Introduction.

This paper establishes an infinite-dimensional linear programming formulation for problems in optimal stopping. In general terms, a decision maker observes a process X and decides the time τ at which to stop it. The person receives a reward $R(x)$ when the process is stopped in state x, i.e., when $X(\tau) = x$. The goal of the decision maker is to select a stopping rule which maximizes the expected reward $E[R(X(\tau))]$.

American options provide one set of examples of optimal stopping problems. For an American option, the person has the right to exercise the option at any time during the contract period, that is, up until its time of expiration. Thus the owner of the option must decide when to exercise his or her option.

Another example of optimal stopping in the area of finance is the decision about when to sell stock. Suppose a person must liquidate his or her holdings by a certain date. This person must then choose the time to sell. Of course, the goal in this setting is to try to choose the time when the stock is at its highest level for the entire period. A formulation of this problem has been studied by Shiryaev [13].

Optimal stopping has been well studied, and the value function has been characterized in terms of the minimal excessive function lying above the reward function (see, e.g., the text by Shiryaev [12]). This paper adopts a linear programming formulation for optimal stopping problems. This approach has the benefit that numerical solution is often easy to obtain since there are excellent linear programming software packages available.

The main contribution is that the paper establishes an equivalence between the stochastic process formulation and the linear programming formulation. It also indicates an equivalence between optimal stopping and control problems. In addition, this paper presents two examples to illustrate the effectiveness of linear programming as a numerical solution technique. The optimal stopping linear programs are

*Received by the editors August 31, 2000; accepted for publication (in revised form) August 18, 2001; published electronically March 20, 2002. This research is partially supported by NSF under grant DMS 9803490.

http://www.siam.org/journals/sicon/40-6/37766.html

†Department of Statistics, University of Kentucky, Lexington, KY 40506-0027 (cho@ms.uky.edu, stockb@ms.uky.edu).

infinite-dimensional. It is thus necessary to approximate these by finite-dimensional linear programs. In this paper, we adopt a discretization approach to the general state space processes and approximate the continuous time stochastic processes by continuous time Markov chains (see Kushner and Dupuis [9]).

This work extends to optimal stopping problems similar types of infinite-dimensional linear programming formulations for stochastic control and Markov decision problems (see, e.g., Bhatt and Borkar [1], Hernández-Lerma and Lasserre [6], Helmes and Stockbridge [4], Helmes, Röhl, and Stockbridge [5], Kurtz and Stockbridge [8], and Mendiondo and Stockbridge [11]). In particular, the papers [4, 5, 11] demonstrate the accuracy of the linear programming approach for the numerical analysis of controlled and uncontrolled processes.

The paper is organized as follows. We give the stochastic process formulation in section 2 and identify the linear programming form of the problem in section 2.1. Section 3 proves equivalence between these forms, section 4 discusses convergence of finite-dimensional approximation schemes, and the linear programming method is illustrated in examples in section 5.

2. Formulation. We formulate the optimal stopping problem in a very general setting by assuming the state space E is a complete, separable metric space. We will augment the state space with extra components for time and control processes. Note that the augmented spaces will still be complete, separable metric spaces. We therefore define the notation used in this paper for a complete, separable metric space S. Let $M(S)$ denote the space of Borel measurable functions on S, let $C(S)$ denote the space of continuous functions on S, let $\overline{C}(S)$ denote the space of bounded, continuous functions on S, let $\mathcal{M}(S)$ denote the space of finite Borel measures on S, and let $\mathcal{P}(S) \subset \mathcal{M}(S)$ denote the collection of probability measures on S. For a Borel set $\Gamma \subset S$, we define $I_\Gamma \in M(S)$ to be the indicator function of the set Γ, and, for a point $s \in S$, we define $\delta_{\{s\}}(\cdot) \in \mathcal{M}(S)$ to be the Borel measure which places a unit point mass at $\{s\}$.

The processes under consideration are characterized as solutions to a martingale problem for their generator; that is, suppose A is the generator for a Markov process X with state space E, where X is related to A by the requirement that

$$f(X(t)) - f(X(0)) - \int_0^t Af(X(s))ds \tag{1}$$

be a martingale for each $f \in \mathcal{D}$, the domain of A. This characterization was developed by Stroock and Varadhan [14, 15, 16] for multidimensional diffusion processes and has been studied by many authors. Ethier and Kurtz [3] provide an excellent reference for general Markov processes.

We assume that the generator A has the following properties.

Condition 1.

(i) $A : \mathcal{D} \subset \overline{C}(E) \to C(E)$, $1 \in \mathcal{D}$, and $A1 = 0$.

(ii) $\mathcal{D}$ is closed under multiplication and separates points.

(iii) The graph of A is *separable* in the sense that there exists a countable collection $\{g_k\} \subset \mathcal{D}$ such that $\{(f, Af) : f \in \mathcal{D}\}$ is contained in the bounded, pointwise closure of the linear span of $\{(g_k, Ag_k) : k \geq 1\}$.

Let $\nu_0 \in \mathcal{P}(E)$ denote the initial distribution of the desired processes. We say a process $X = \{X(t) : t \geq 0\}$ is a solution of the martingale problem for (A, ν_0) if there

exists a filtration $\{\mathcal{F}_t\}$ such that $X(0)$ has distribution ν_0, X is $\{\mathcal{F}_t\}$-progressively measurable, and (1) is an $\{\mathcal{F}_t\}$-martingale for every $f \in \mathcal{D}$.

The reward obtained when the process stops is given by the function $R : E \to \mathbb{R}$. We assume that R is upper-semicontinuous and bounded above. The objective of the decision maker is to maximize

$$E[R(X(\tau))] \tag{2}$$

over all $\{\mathcal{F}_t\}$-stopping times τ. We require

$$E[\tau] < \infty, \tag{3}$$

which implies $\tau < \infty$ a.s.

2.1. Linear programming formulation. Let X be a solution of the martingale problem for (A, ν_0). It then follows that, for each $\gamma \in \widehat{C}^1(\mathbb{R}^+)$, $\widehat{C}^1(\mathbb{R}^+)$ being the space of continuously differentiable functions which vanish at ∞, and each test function $f \in \mathcal{D}$,

$$\gamma(t)f(X(t)) - \int_0^t [\gamma(s)Af(X(s)) + \gamma'(s)f(X(s))]ds \tag{4}$$

is an $\{\mathcal{F}_t\}$-martingale (see [3, Theorem 4.7.1]). To simplify notation, define

$$\widehat{A}(\gamma f)(t,x) = \gamma(t)Af(x) + \gamma'(t)f(x),$$

and let $\widehat{\mathcal{D}} = \{\gamma f : \gamma \in \widehat{C}^1(\mathbb{R}^+), f \in \mathcal{D}\}$.

Now let τ be any stopping time satisfying (3). Since (4) is an $\{\mathcal{F}_t\}$-martingale, the optional sampling theorem (see [3]) implies that

$$E\left[\gamma(\tau)f(X(\tau)) - \int_0^\tau \widehat{A}(\gamma f)(s, X(s))ds \,\middle|\, \mathcal{F}_0\right] = \gamma(0)f(X(0)), \tag{5}$$

and hence

$$E\left[\gamma(\tau)f(X(\tau)) - \int_0^\tau \widehat{A}(\gamma f)(s, X(s))ds\right] = \gamma(0)E[f(X(0))] \tag{6}$$

for each $\gamma f \in \widehat{\mathcal{D}}$.

Now let the measure $\mu_\tau \in \mathcal{P}(\mathbb{R}^+ \times E)$ be the joint distribution of $(\tau, X(\tau))$, and define the occupation measure $\mu_0 \in \mathcal{M}(\mathbb{R}^+ \times E)$ by

$$\mu_0(\Gamma) = E\left[\int_0^\tau I_\Gamma(s, X(s))ds\right] \quad \forall\, \Gamma \in \mathcal{B}(\mathbb{R}^+ \times E). \tag{7}$$

Note that $\mu_0(\mathbb{R}^+ \times E) = E[\tau] < \infty$. Then (6) can be rewritten, for each $\gamma f \in \widehat{\mathcal{D}}$, as

$$\int \gamma(t)f(x)\mu_\tau(dt \times dx) - \int \widehat{A}(\gamma f)(t,x)\mu_0(dt \times dx) = \gamma(0)\int f(x)\nu_0(dx). \tag{8}$$

Thus for each X and τ there are measures μ_τ and μ_0 satisfying (8) for all $\gamma f \in \widehat{\mathcal{D}}$.

We will show that for each pair of measures (μ_τ, μ_0) satisfying the conditions (8) there are a corresponding process X and a stopping time τ for which X is a solution of the martingale problem for (A, ν_0) up to time τ, and μ_τ and μ_0 are the joint distribution and occupation measures, respectively. The optimal stopping problem can thus be reformulated as the following infinite-dimensional linear programming problem over the space of pairs of measures:

$$
\begin{aligned}
&\text{Max} \int R(x)\mu_\tau(dt \times dx) \\
&\text{S.t.} \int \gamma f d\mu_\tau - \int \widehat{A}(\gamma f) d\mu_0 = \gamma(0) \int f d\nu_0 \qquad \forall\, \gamma f \in \widehat{\mathcal{D}}, \\
&\mu_\tau \in \mathcal{P}(\mathbb{R}^+ \times E),\ \mu_0 \in \mathcal{M}(\mathbb{R}^+ \times E).
\end{aligned} \tag{9}
$$

3. Existence result. Section 2.1 showed that to each solution X of the martingale problem for (A, ν_0) and stopping time τ satisfying (3) there corresponds a pair of measures (μ_τ, μ_0) satisfying (8). We now establish the reverse correspondence.

The approach taken to establish this existence result is to define a new generator $\overline{A}$ for a *controlled* process having two time components, one spatial component, and one control component (u). The generator $\overline{A}$ involves the choice between the spatial dynamics specified by A and a process in which the spatial component is fixed until a random time at which the process is "restarted." One time component tracks the time in which the process having generator A runs, and the other measures the time in which the process is fixed until it restarts. We then define a stationary measure for the generator $\overline{A}$ and use an existence result of Bhatt and Borkar [1] to obtain a stationary (three-dimensional) process having generator $\overline{A}$. A new process is defined by (pathwise) starting the stationary process at a time that the process restarts (thus losing stationarity). The stopping time τ is then the time at which the control switches away from the generator A, and a change of measure ensures that the process restricted to one "cycle" has the desired dynamics.

For completeness, we state the existence result of Bhatt and Borkar using our notation (see [1, Theorem 2.1 and Corollary 2.1]). We refer the reader to the reference for the necessary modification of Condition 1 to incorporate controls in the formulation. We also restate a technical lemma of Kurtz and Stockbridge (see [8, Lemma 4.4]) which is used to establish the behavior of the stationary process.

THEOREM 3.1. *Let S be a complete, separable metric space, and let U be a compact metric space. Suppose the generator $A : \mathcal{D}(A) \subset \overline{C}(S) \to C(S \times U)$. Suppose $\mu \in \mathcal{P}(S \times U)$ is such that*

$$\int_{S \times U} Af(y,u)\, \mu(dy \times du) = 0 \qquad \forall f \in \mathcal{D}(A).$$

Let η be the regular conditional distribution of u given y under μ, so η satisfies $\mu(dy \times du) = \eta(y, du)\mu(dy \times U)$. Then there exists a stationary process Y such that

$$f(Y(t)) - \int_0^t \int_U Af(Y(s), u)\eta(Y(s), du)ds$$

is an $\{\mathcal{F}_t^Y\}$*-martingale for each* $f \in \mathcal{D}(A)$*, and*

$$E\left[\int_U I_\Gamma(Y(s),u)\eta(Y(s),du)\right] = \mu(\Gamma) \qquad \forall\, \Gamma \in \mathcal{B}(E\times U).$$

LEMMA 3.2. *Let* Q *be a nonnegative,* $\{\mathcal{F}_t\}$*-adapted cadlag process, let* V_1 *and* V_2 *be bounded, nonnegative, measurable,* $\{\mathcal{F}_t\}$*-adapted processes, and suppose that*

$$g(Q(t)) - \int_0^t (V_1(s)g'(Q(s)) + V_2(s)(g(0) - g(Q(s)))) \, ds$$

is an $\{\mathcal{F}_t\}$*-martingale for every* C^1 *function* g *with* g *and* g' *bounded. Let* τ *be a stopping time, and define* $\sigma_0^\tau = \inf\{t > \tau : Q(t) > 0\}$ *and* $\sigma_1^\tau = \int\{t > \sigma_0^\tau : Q(t) = 0\}$*. Then, for* $\tau \le t < \sigma_1^\tau$*,* $Q(t) - Q(\tau) = \int_\tau^t V_1(s)\, ds$*, and if* $\sigma_1^\tau < \infty$ *a.s.,*

$$P\left\{\int_{\sigma_0^\tau}^{\sigma_1^\tau} V_2(s)\, ds > x | \mathcal{F}_{\sigma_0^\tau}\right\} = e^{-x}, \qquad x \ge 0.$$

The main result is given in the next theorem. The proof of this result follows the proof of a similar result in Kurtz and Stockbridge (see [8, Theorem 4.5]).

THEOREM 3.3. *Suppose* $\mu_\tau \in \mathcal{P}(\mathbb{R}^+ \times E)$ *and* $\mu_0 \in \mathcal{M}(\mathbb{R}^+ \times E)$ *satisfy* (8)*. Then there exist a filtration* $\{\mathcal{F}_t\}$*, a process* X*, and an* $\{\mathcal{F}_t\}$*-stopping time* τ *such that* X *is a solution of the martingale problem for* (A, ν_0) *up to time* τ*,* $(\tau, X(\tau))$ *has distribution* μ_τ*, and the occupation measure of* X *up to time* τ*, defined in* (7)*, is* μ_0*.*

Proof. Augment the state space with an extra time dimension and the space $U = \{0,1\}$, which we refer to as the ("bang-bang") control space. Define a new generator $\overline{A}$ by

(10)
$$\begin{aligned}\overline{A}(\beta\gamma f)(r,s,x,u) &= u\beta(r)\widehat{A}(\gamma f)(s,x)\\ &\quad + (1-u)\left[\beta(0)\gamma(0)\int f d\nu_0 - \beta(r)\gamma(s)f(x) + \beta'(r)\gamma(s)f(x)\right]\end{aligned}$$

for each $\beta, \gamma \in \widehat{C}^1(\mathbb{R}^+)$ and $f \in \mathcal{D}$.

Define the measure $\overline{\mu} \in \mathcal{P}(\mathbb{R}^+ \times \mathbb{R}^+ \times E \times \{0,1\})$, which satisfies

(11)
$$\begin{aligned}\int h(r,s,x,u)\overline{\mu}(dr\times ds\times dx\times du) &= K^{-1}\left(\int_{\mathbb{R}^+\times E} h(0,s,x,1)\mu_0(ds\times dx)\right.\\ &\quad \left. + \int_{\mathbb{R}^+\times E}\int_0^\infty e^{-r}h(r,s,x,0)dr\mu_\tau(ds\times dx)\right)\end{aligned}$$

for each bounded, continuous h, where $K = \mu_0(\mathbb{R}^+\times E)+1$. Note that the conditional distribution of u given (r,s,x) under $\overline{\mu}$ is

$$\overline{\eta}(r,s,x,du) = \delta_{\{1\}}(du)I_{\{0\}}(r)\frac{d\mu_0}{d\overline{\mu}}(s,x) + \delta_{\{0\}}(du)e^{-r}I_{(0,\infty)}(r)\frac{d\mu_\tau}{d\overline{\mu}}(s,x). \tag{12}$$

The following computation shows that $\int \overline{A}(\beta\gamma f)d\overline{\mu} = 0$ for all $\beta, \gamma \in \widehat{C}^1(\mathbb{R}^+)$, and $f \in \mathcal{D}$.

$$\begin{aligned}
&\int \overline{A}(\beta\gamma f)(r,s,x,u)\overline{\mu}(dr \times ds \times dx \times du)\\
&\quad = K^{-1}\left(\int 1\cdot\beta(0)\widehat{A}(\gamma f)(s,x)\mu_0(ds\times dx)\right.\\
&\qquad\quad + \int_{\mathbb{R}^+\times E}\int_0^\infty e^{-r}(-\beta(r)+\beta'(r))\gamma(s)f(x)\,dr\mu_\tau(ds\times dx)\\
&\qquad\quad \left. + \beta(0)\gamma(0)\int f d\nu_0\right)\\
&\quad = K^{-1}\left(\int \beta(0)\widehat{A}(\gamma f)(s,x)\mu_0(ds\times dx)\right.\\
&\qquad\quad + \int_{\mathbb{R}^+\times E}\int_0^\infty \frac{d}{dr}\left(e^{-r}\beta(r)\right)dr\gamma(s)f(x)\mu_\tau(ds\times dx) + \beta(0)\gamma(0)\left.\int f d\nu_0\right)\\
&\quad = K^{-1}\beta(0)\left(\int \widehat{A}(\gamma f)(s,x)\mu_0(ds\times dx)\right.\\
&\qquad\quad \left. - \int_{\mathbb{R}^+\times E}\gamma(s)f(x)\mu_\tau(ds\times dx) + \gamma(0)\int f d\nu_0\right)\\
&\quad = 0.
\end{aligned}$$

The conditions of Theorem 3.1 on the state and control spaces and the generator are also satisfied, which therefore implies the existence of a stationary $\mathbb{R}^+\times\mathbb{R}^+\times E$-valued process (R,S,Y) (which we may assume is defined for all $t\in\mathbb{R}$) such that

(13)
$$\beta(R(t))\gamma(S(t))f(Y(t)) - \int_0^t\int_U \overline{A}(\beta\gamma f)(R(s),S(s),Y(s),u)\,\overline{\eta}(R(s),S(s),Y(s),du)\,ds$$

is an $\{\mathcal{F}_t^{R,S,Y}\}$-martingale for all $\beta,\gamma\in\widehat{C}^1(\mathbb{R}^+)$, and $f\in\mathcal{D}$, where $\overline{\eta}$ is given by (12).

For each $t\geq 0$, define the following random variables: $\sigma_{-1}^t = \sup\{r<t : S(r)=0, R(r)=0\}$, $\sigma_1^t = \inf\{r\geq t : S(r)=0, R(r)=0\}$, $\tau_1^t = \inf\{r>\sigma_1^t : R(r)>0\}$, and $\sigma_2^t = \inf\{r>\tau_1^t : R(r)=0\}$. For $s\in[\sigma_1^t,\tau_1^t)$, by definition $R(s)=0$ and by Lemma 3.2 $S(s) = \int_{\sigma_1^t}^s I_{\{0\}}(R(r))dr = (s-\sigma_1^t)$. For $s\in[\tau_1^t,\sigma_2^t)$, by Lemma 3.2 $R(s) = \int_{\tau_1^t}^s I_{(0,\infty)}(R(r))dr = s-\tau_1^t$ a.s. and conditional on $\mathcal{F}_{\tau_1^t}$, $\sigma_2^t-\tau_1^t$ is exponentially distributed with mean 1, and again by Lemma 3.2 $S(s) = S(\tau_1^t) + \int_{\tau_1^t}^s I_{\{0\}}(R(r))dr = S(\tau_1^t) = \tau_1^t - \sigma_1^t$. Starting with $g(r+s) = e^{-\alpha(r+s)}$ and approximating more general g by linear combinations of these exponentials, we see that

$$\begin{aligned}
g(S(t)+R(t)) - \int_0^t (g'(S(r)+R(r)) + (1-\overline{u}(R(r),S(r),Y(r)))(g(0)\\
- g(S(r)+R(r))))dr
\end{aligned}$$

is a martingale for C^1 functions with g and g' bounded, where

$$\overline{u}(R(r),S(r),Y(r)) = \int u\overline{\eta}(R(r),S(r),Y(r),du).$$

Letting $\tilde{\sigma}_2^t = \inf\{r > \tau_1^t : S(r) + R(r) = 0\}$, Lemma 3.2 implies

$$P\left\{\int_{\tau_1^t}^{\sigma_2^t}(1-\overline{u}(R(r),S(r),Y(r)))dr > x\big|\mathcal{F}_{\tau_1^t}^{R,S,Y}\right\}$$
$$= e^{-x} = P\left\{\int_{\tau_1^t}^{\tilde{\sigma}_2^t}(1-\overline{u}(R(r),S(r),Y(r)))dr > x\big|\mathcal{F}_{\tau_1^t}^{R,S,Y}\right\},$$

and since $\sigma_2^t \le \tilde{\sigma}_2^t$, we must have $\sigma_2^t = \tilde{\sigma}_2^t$ a.s. In particular, $S(\sigma_2^t) = 0$ a.s. Finally, defining $Z(r) = (R(\tau_1^t + r), S(\tau_1^t + r), Y(\tau_1^t + r))$ for $r \le \sigma_2^t - \tau_1^t$, we can extend Z to be a solution of the martingale problem for the generator C defined by

$$Cg(r,s,x) = \int g(0,0,y)\nu_0(dy) - g(r,s,x) + \frac{\partial}{\partial r}g(r,s,x).$$

Since any solution of this martingale problem has the property that the final component is constant except for jumps that occur when the first two components jump to zero, it follows that $Y(r) = Y(\tau_1^t)$ for $\tau_1^t \le r < \sigma_2^t$.

Let h be a fixed, bounded, continuous function and, for $\epsilon > 0$, define

$$H_\epsilon(r) = \int_U e^{-\epsilon(R(r)+S(r))}h(R(r),S(r),Y(r),u)\,\overline{\eta}(R(r),S(r),Y(r),du).$$

Then, as a process in t,

$$(\sigma_1^t - \sigma_{-1}^t)^{-1}\int_{\sigma_1^t}^{\sigma_2^t} H_\epsilon(r)\,dr \tag{14}$$

is stationary, and for each t and $s \in [\sigma_{-1}^t, \sigma_1^t)$

$$(\sigma_1^s - \sigma_{-1}^s)^{-1}\int_{\sigma_1^s}^{\sigma_2^s} H_\epsilon(r)\,dr = (\sigma_1^t - \sigma_{-1}^t)^{-1}\int_{\sigma_1^t}^{\sigma_2^t} H_\epsilon(r)\,dr.$$

These expressions are set equal to 0 whenever $\sigma_{-1}^t = -\infty$ or $\sigma_1^t = +\infty$.

Using stationarity,

$$E\left[(\sigma_1^t - \sigma_{-1}^t)^{-1}\int_{\sigma_1^t}^{\sigma_2^t} H_\epsilon(r)\,dr\right] = T^{-1}\int_0^T E\left[(\sigma_1^t - \sigma_{-1}^t)^{-1}\int_{\sigma_1^t}^{\sigma_2^t} H_\epsilon(r)dr\right]\,dt, \tag{15}$$

in which both sides may be infinite due to the $(\sigma_1^t - \sigma_{-1}^t)^{-1}$ term. The following argument, in fact, shows that both terms are finite and identifies their common value.

Let $N(T)$ denote the number of jumps of the process (R,S,Y) in the interval $[0,T]$, let $\{\sigma_k : k = 1,\ldots,N(T)\}$ denote these jump times, and let $\sigma_{N(T)+1}$ and σ_{-1} ($= \sigma_0$ in the summation) denote the first jump time after time T and the last jump

time before time 0, respectively. Then the right-hand side of (15) equals

$$
\begin{aligned}
T^{-1}E\left[\sum_{k=1}^{N(T)+1}\frac{T\wedge\sigma_k-\sigma_{k-1}\vee 0}{\sigma_k-\sigma_{k-1}}\int_{\sigma_{k+1}}^{\sigma_{k+2}}H_\epsilon(r)\,dr\right]
&= T^{-1}E\left[\int_0^T H_\epsilon(r)\,dr\right]\\
&\quad -T^{-1}E\left[\int_0^{\sigma_1\wedge T}\left(1-\frac{T\wedge\sigma_1}{\sigma_1-\sigma_{-1}}\right)H_\epsilon(r)\,dr\right]\\
&\quad +T^{-1}E\left[I_{\{N(T)=1\}}\frac{\sigma_1}{\sigma_1-\sigma_{-1}}\int_T^{\sigma_2}H_\epsilon(r)\,dr\right]\\
&\quad +T^{-1}E\left[I_{\{N(T)>1\}}\int_T^{\sigma_{N(T)+1}}H_\epsilon(r)\,dr\right]\\
&\quad +T^{-1}E\left[I_{\{N(T)>0\}}\frac{T-\sigma_{N(T)}}{\sigma_{N(T)+1}-\sigma_{N(T)}}\int_{\sigma_{N(T)+1}}^{\sigma_{N(T)+2}}H_\epsilon(r)\,dr\right].
\end{aligned}
$$

Observe that the first term is

$$
\int e^{-\epsilon(r+s)}h(r,s,x,u)\,\overline{\mu}(dr\times ds\times dx\times du) \tag{16}
$$

and that the last four terms are bounded above by $4||h||/(\epsilon T)$. Thus all terms are finite, implying the terms in (15) are finite, and, moreover, as $T\to\infty$, these converge to (16).

Letting $\epsilon\to 0$ gives, for each bounded, continuous h (and hence for each bounded, measurable h),

$$
\begin{aligned}
&E\left[(\sigma_1^t-\sigma_{-1}^t)^{-1}\int_{\sigma_1^t}^{\sigma_2^t}\int_U h(R(r),S(r),Y(r),u)\,\overline{\eta}(R(r),S(r),Y(r),du)dr\right]\\
&\qquad=\int h(r,s,x,u)\,\overline{\mu}(dr\times ds\times dx\times du).
\end{aligned} \tag{17}
$$

Then considering $h(r,s,x,u)=I_{\{0\}}(u)$ in (17) yields

$$
\begin{aligned}
K^{-1}&=E[(\sigma_1^t-\sigma_{-1}^t)^{-1}(\sigma_2^t-\tau_1^t)]\\
&=E[(\sigma_1^t-\sigma_{-1}^t)^{-1}],
\end{aligned}
$$

in which the last equality follows from the fact that, conditional on $\mathcal{F}^{R,S,Y}_{\tau_1^t}$, $\sigma_2^t-\tau_1^t$ is exponentially distributed with mean 1.

Now specify σ_{-1}^t and σ_1^t when $t=0$. Define the process $\tilde{X}$ by $\tilde{X}(r)=Y(\sigma_1^0+r)$, $\tilde{R}(r)=R(\sigma_1^0+r)$, $\tilde{S}(r)=S(\sigma_1^0+r)$, and the filtration $\{\mathcal{F}_r\}=\{\mathcal{F}^{R,S,Y}_{\sigma_1^0+r}\}$. Let $\tilde{\tau}=\inf\{r\ge 0:\tilde{R}(r)>0\}$ and $\tilde{\sigma}=\inf\{r>\tilde{\tau}:\tilde{R}(r)=0\}$, and note that $\tilde{X}(r)=\tilde{X}(\tilde{\tau})$ for $\tilde{\tau}\le r<\tilde{\sigma}$. Observe that both σ_1^0 and σ_{-1}^0 are $\mathcal{F}_0$-measurable. Define a new

probability measure $\hat{P}$ to have Radon–Nikodym derivative $K(\sigma_1^0 - \sigma_{-1}^0)^{-1}$. It then follows from (17) that for each bounded, measurable h

$$
\begin{aligned}
&E^{\hat{P}}\left[\int_0^{\tilde{\sigma}} \int_U h(\tilde{R}(r), \tilde{S}(r), \tilde{X}(r), u)\, \overline{\eta}(\tilde{R}(r), \tilde{S}(r), \tilde{X}(r), du)dr\right] \\
&\qquad = \int h(r, s, x, u)\, \overline{\mu}(dr \times ds \times dx \times du)/E[(\sigma_1^0 - \sigma_{-1}^0)^{-1}] \qquad (18) \\
&\qquad = \int h(0, s, x, 1)\mu_0(ds \times dx) \\
&\qquad\qquad + \int_0^{\infty} \int e^{-r} h(r, s, x, 0)\ \mu_\tau(ds \times dx)dr.
\end{aligned}
$$

Taking $h(r, s, x, u) = I_{\{0\}}(r)I_{\{0\}}(u)$ in (18) shows that $\overline{\eta}(\tilde{R}(r), \tilde{S}(r), \tilde{X}(r), \cdot)$ places unit mass at $\{1\}$ a.s. for $0 \le r \le \tilde{\tau}$, and, similarly, $h(r, s, x, u) = I_{(0,\infty)}(r)I_{\{1\}}(u)$ establishes that $\overline{\eta}(\tilde{R}(r), \tilde{S}(r), \tilde{X}(r), \cdot)$ places unit mass at $\{0\}$ a.s. for $\tilde{\tau} \le r \le \tilde{\sigma}$. Furthermore, for $\Gamma \in \mathcal{B}(\mathbb{R}^+ \times E)$ and $h(r, s, x, u) = I_{(0,\infty)\times\Gamma\times\{0\}}(r, s, x, u)$, (18) implies

$$
\begin{aligned}
\mu_\tau(\Gamma) &= E\left[\int_{\tilde{\tau}}^{\tilde{\sigma}} I_\Gamma(\tilde{S}(r), \tilde{X}(r))\, dr\right] \\
&= E[I_\Gamma(\tilde{\tau}, \tilde{X}(\tilde{\tau}))(\tilde{\sigma} - \tilde{\tau})] \\
&= E[I_\Gamma(\tilde{\tau}, \tilde{X}(\tilde{\tau}))],
\end{aligned}
$$

where the last equality follows from the fact that $\tilde{\sigma} - \tilde{\tau}$ is a mean 1 exponential time conditional on $\mathcal{F}_{\tilde{\tau}}$. Similarly, for $\Gamma \in \mathcal{B}(\mathbb{R}^+ \times E)$ and $h(r, s, x, u) = I_{\{0\}\times\Gamma\times\{1\}}(r, s, x, u)$, (18) implies

$$
E\left[\int_0^{\tilde{\tau}} I_\Gamma(r, \tilde{X}(r))\, dr\right] = \mu_0(\Gamma).
$$

Finally, by compactifying the time components to $[0, \infty]$, extending the generator $\overline{A}$ to allow $\beta, \gamma \in C[0, \infty]$ (see [3, Theorem 4.5.4]), taking $\beta \equiv 1$ and $\gamma \equiv 1$ in (13), and recalling that $\overline{\eta}(\tilde{R}(r), S(\tilde{r}), \tilde{X}(r), \cdot) = \delta_{\{0\}}(\cdot)$ for $r \in (0, \tilde{\tau})$, the optional sampling theorem implies that under $\hat{P}$

$$
f(\tilde{X}(t \wedge \tilde{\tau})) - \int_0^{t\wedge\tilde{\tau}} Af(\tilde{X}(s))\, ds
$$

is a martingale with respect to the filtration $\{\mathcal{F}_t\}$.

4. Finite-dimensional approximation. The linear program (9) is typically infinite-dimensional since the variables are measures on $\mathbb{R}^+ \times E$. To obtain numerical results, it is therefore necessary to reduce the linear program to a finite-dimensional linear program. In this section, we give conditions under which the solutions to a sequence of approximating linear programs converge to the solution of (9). We provide two approaches toward this result.

4.1. Convergence of approximating linear programs. The equivalence of optimal stopping problems with the linear program (9) has been established in sections 2 and 3 under very general conditions. The convergence results, however, are

shown under more restricted conditions and are based on the results of Mendiondo and Stockbridge [10]. An alternate approach to the approximation of the infinite-dimensional linear program is possible using the numerical scheme of Hernández-Lerma and Lasserre [7]. It should be noted that the issue of obtaining approxmiate control policies that converge to the optimal control policy for the original problem remains an open problem.

We assume that E is a compact, metric space and that the reward function R is continuous. We also compactify $\mathbb{R}^+$ with the point at ∞ and extend the generator $\widehat{A}$ to functions γf with $\gamma \in C(\overline{\mathbb{R}^+})$ and $f \in \mathcal{D}$ as in [3, Theorem 4.5.4]. The extension of the generator and domain are still denoted $\widehat{A}$ and $\widehat{\mathcal{D}}$, respectively.

The approximating linear programming problems. In our application of the convergence results, the approximations will be obtained by discretizing the space. The convergence results apply to other approximation schemes as well, so we establish the results in the more general setting.

For $n \geq 0$, let E_n be a metric space, and denote the distance between points in E_n by $|\cdot - \cdot|$. We assume for each n there exist measurable functions $\psi_n : E_n \to E$ and $\phi_n : E \to E_n$ such that

(C1) $|x - \psi_n(\phi_n(x))| \to 0$ as $n \to \infty$ for all $x \in E$.

Note that, in our application, E_n will be a discretization of E, ψ_n will be the natural imbedding of E_n in E, and ϕ_n will map rectangles to the (single) point in the discretization contained in the rectangles.

For each n, we transfer the reward function R to E_n by defining $R_n = R \circ \psi_n$.

Let $A_n : \mathcal{D}_n \subset C(E_n) \to C(E_n)$ be such that, for each $f \in \mathcal{D}$, there exists $f_n \in \mathcal{D}(A_n)$ satisfying the following:

(C2) $\sup_{y \in E_n} |f_n(y) - f(\psi_n(y))| \to 0$ as $n \to \infty$;

(C3) $\sup_{y \in E_n} |A_n f_n(y) - Af(\psi_n(y))| \to 0$ as $n \to \infty$.

Note that conditions (C2) and (C3) are satisfied by $\widehat{A}$. We denote the approximations of $\widehat{A}$ by $\widehat{A}_n$ and the approximations of the initial distribution ν_0 by ν_0^n.

The approximating linear programs are

$$\begin{aligned} &\text{Max} \int R_n(y)\mu_\tau^n(dt \times dy) \\ &\text{S.t.} \int \gamma_n f_n d\mu_\tau^n - \int \widehat{A}_n(\gamma_n f_n) d\mu_0^n = \gamma_n(0) \int f_n d\nu_0^n \qquad \forall \gamma_n f_n \in \widehat{\mathcal{D}}(A_n), \\ &\mu_\tau^n \in \mathcal{P}(\overline{\mathbb{R}_n^+} \times E_n),\ \mu_0^n \in \mathcal{M}(\overline{\mathbb{R}_n^+} \times E_n). \end{aligned} \tag{19}$$

To apply the results in [10], the constraints of (9) and (19) will be rephrased to fit the formulation of that paper. Let $U = \{u_1, u_2\}$ be a control space consisting of two distinct points. Define a new generator $\tilde{A} : \widehat{\mathcal{D}} \to C(\overline{\mathbb{R}^+} \times E \times U)$ by

$$\tilde{A}(\gamma f)(t,x,u) = \begin{cases} \widehat{A}(\gamma f)(t,x) & \text{if } u = u_1, \\ \gamma(0) \int f d\nu_0 - \gamma(t) f(x) & \text{if } u = u_2, \end{cases}$$

and note that conditions (C2)–(C3) apply to $\tilde{A}$. We denote the approximations of $\tilde{A}$ by $\tilde{A}_n$.

We further assume the following.

(C4) For each n, for each conditional distribution $\widehat{\eta}^n$ on $\{u_1, u_2\}$ given E_n, there exist measures $\mu_\tau^n \in \mathcal{M}(\overline{\mathbb{R}_n^+} \times E_n)$ and $\mu_0^n \in \mathcal{P}(\overline{\mathbb{R}_n^+} \times E_n)$ such that for each $\gamma_n f_n \in \widehat{\mathcal{D}}(A_n)$

$$\int \gamma_n f_n(t,y)\mu_\tau^n(dt \times dy) - \int\int \widehat{A}_n(\gamma_n f_n)(t,y,u)\widehat{\eta}^n(y,du)\mu_0^n(dt \times dy) = \gamma_n(0)\int f_n d\nu_0^n.$$

Condition (C4) essentially assumes the existence of a stationary distribution for the approximating process.

Define the measure $\tilde{\mu} \in \mathcal{P}(\overline{\mathbb{R}^+} \times E \times U)$ by

$$\tilde{\mu}(\cdot \times \{u_1\}) = K^{-1}\mu_0(\cdot), \qquad \tilde{\mu}(\cdot \times \{u_2\}) = K^{-1}\mu_\tau(\cdot),$$

where $K = \mu_0(\overline{\mathbb{R}^+} \times E) + 1$. Then the constraints of (9) become

$$\int_{\mathbb{R}^+ \times E \times U} \tilde{A}(\gamma f)d\tilde{\mu} = 0 \qquad \forall\, \gamma f \in \widehat{\mathcal{D}},$$
$$\mu \in \mathcal{P}(\overline{\mathbb{R}^+} \times E \times U).$$

A similar reformulation of the constraints of (19) gives the form which is analyzed in [10], and thus the optimal values of (19) converge to that of (9) (see Theorem 4 of [10]).

4.2. Markov chain approximations. The second approach to the issue of finite-dimensional approximations and convergence of the optimal values is to employ Kushner's Markov chain approximation scheme. To apply the convergence results in Kushner and Dupuis [9], we establish the equivalence between the optimal stopping problem and an associated absolutely continuous stochastic control problem. This control problem is then approximated by the corresponding problem of controlling a continuous time Markov chain which approximates the original Markov process. Convergence of the value function then follows.

Let $U = \{0, 1\}$, and this time define the controlled generator $\tilde{A} : \mathcal{D} \to \overline{C}(E \times U)$ by

$$\tilde{A}f(x,u) = u\,Af(x). \tag{20}$$

A paired process $(\tilde{X}, u)$ is a solution of the controlled martingale problem for $(\tilde{A}, \nu_0)$ if $\tilde{X}(0)$ has distribution ν_0, and there is a filtration $\{\mathcal{F}_t\}$ such that $(\tilde{X}, u)$ is $\{\mathcal{F}_t\}$-progressively measurable and

$$f(\tilde{X}(t)) - \int_0^t \tilde{A}f(\tilde{X}(s), u(s))ds$$

is an $\{\mathcal{F}_t\}$-martingale. For the optimal stopping problem, we further require

$$\lim_{t\to\infty} \tilde{X}(t) =: \tilde{X}(\infty) \text{ exists a.s.} \tag{21}$$

The first theorem shows that to each pair (X, τ), in which X is a solution of the martingale problem for (A, ν_0) and τ is a stopping time, there exists a pair $(\tilde{X}, u)$, which is a solution of the controlled martingale problem for $(\tilde{A}, \nu_0)$ satisfying (21).

THEOREM 4.1. *Let X be a solution of the martingale problem for (A, ν_0), and let τ be a stopping time having finite expectation. Let μ_τ denote the distribution of $X(\tau)$, and let μ_0 be defined by (7). Then there exists a solution $(\tilde{X}, u)$ of the controlled martingale problem for $(\tilde{A}, \nu_0)$ such that $\lim_{t\to\infty} \tilde{X}(t)$ exists a.s., $\tilde{X}(\infty)$ has distribution μ_τ, and*

$$E\left[\int_0^\infty u(s) I_\Gamma(\tilde{X}(s)) ds\right] = \mu_0(\Gamma)$$

for every $\Gamma \in \mathcal{B}(E)$.

Proof. Define $\tilde{X}(t) = X(t \wedge \tau)$ and $u(t) = I_{[0,\tau)}(t)$. The conclusion now follows.

The next theorem shows the reverse correspondence.

THEOREM 4.2. *Let $(\tilde{X}, u)$ be a solution of the controlled martingale problem for $(\tilde{A}, \nu_0)$ with*

$$E\left[\int_0^\infty u(s) ds\right] < \infty$$

and such that $\lim_{t\to\infty} \tilde{X}(t)$ exists a.s. Then there exist a process X and a stopping time τ, with $E[\tau] < \infty$, such that X is a solution of the martingale problem for (A, ν_0) up to time τ.

Proof. We sketch the proof. Define $\tau = \int_0^\infty u(s)\, ds$, and note that $\tau < \infty$ a.s. Now we restrict our attention to the times that $u = 1$ by defining, for $r \in \mathbb{R}^+$,

$$\sigma(r) = \inf\left\{t : \int_0^t u(s)\, ds = r\right\},$$

where $\sigma(r) = \infty$ for $r \geq \tau$. It then follows that

$$f(\tilde{X}(\sigma(r))) - \int_0^{\sigma(r)} u(s) A f(\tilde{X}(s))\, ds \tag{22}$$

is an $\{\mathcal{F}_{\sigma(r)}\}$-martingale. Define $X(r) = \tilde{X}(\sigma(r))$ and $\mathcal{G}_r = \mathcal{F}_{\sigma(r)}$ for $r \in \mathbb{R}^+$, and observe that (22) is

$$f(X(r \wedge \tau)) - \int_0^{r\wedge\tau} A f(X(s))\, ds.$$

Theorems 4.1 and 4.2 establish the equivalence of stopped solutions (X, τ) of the martingale problem for (A, ν_0) with controlled solutions $(\tilde{X}, u)$ of the martingale problem for $(\tilde{A}, \nu_0)$. The formulation of the objective (2) of the stopping problem in terms of the controlled process is

$$E[R(\tilde{X}(\infty))], \tag{23}$$

and the control problem is to maximize (23) over all solutions of the controlled martingale problem for $(\tilde{A}, \nu_0)$ satisfying (21).

Kushner's numerical method approximates the process $(\tilde{X}, u)$ by a continuous time controlled Markov chain $(\tilde{X}^{(n)}, u^{(n)})$, which satisfies local consistency conditions and then optimizes the corresponding objective involving $\tilde{X}^{(n)}$.

5. Numerical examples. We illustrate the accuracy of the linear programming formulation for optimal stopping problems in this section.

Example 5.1. Let X be a one-dimensional Brownian motion with $X(0) = x_0$, where $0 < x_0 < 1$, so the generator of the Brownian motion process is

$$Af(x) = \frac{1}{2}f''(x), \qquad 0 < x < 1.$$

The process stops automatically when it reaches 0 or 1. Note that this implies that if the decision maker allows the process to run without intervention, it will still stop at some random time τ having finite expectation. The reward obtained when the process stops is given by the function

$$R(x) = 1 - 9x + 59x^2 - 100x^3 + 50x^4.$$

Since the reward function is not time-dependent, we can simplify the problem by taking $\gamma \equiv 1$ in (9) and letting μ_τ and μ_0 be the marginal measures on $[0,1]$ defined by $\mu_\tau(\mathbb{R}^+ \times \cdot)$ and $\mu_0(\mathbb{R}^+ \times \cdot)$, respectively. The linear program for the optimal stopping problem becomes

$$\begin{aligned} &\text{Max} \int [1 - 9x + 59x^2 - 100x^3 + 50x^4]\mu_\tau(dx) \\ &\text{S.t.} \int f(x)\mu_\tau(dx) - \int \frac{1}{2}f''(x)\mu_0(dx) = f(x_0), \\ &\quad \mu_\tau \in \mathcal{P}([0,1]),\ \mu_0 \in \mathcal{M}([0,1]). \end{aligned} \tag{24}$$

This optimal solution for this example can be readily computed using the methods of [12]. The continuation region for this problem consists of $(0, 0.440589) \cup (0.55941, 1)$ with the stopping region being its complement. The value function is the smallest concave function lying above R and is displayed on the graph of the numerical results in Figure 1.

To approximate this infinite-dimensional linear program, discretize $[0,1]$ by setting

$$E_n = \left\{ \frac{k}{n} : k = 1, \ldots, n-1 \right\}$$

and $\overline{E}_n = E_n \cup \{0\} \cup \{1\}$ and using the central finite difference approximation

$$f''(y_k) \approx n^2[f(y_k + 1/n) + f(y_k - 1/n) - 2f(y_k)]$$

for $y_k \in E_n$. The linear program which results is

$$\begin{aligned} &\text{Max} \sum [1 - 9y_k + 59y_k^2 - 100y_k^3 + 50y_k^4]\mu_\tau^{(n)}(y_k) \\ &\text{S.t.} \sum f(y_k)\mu_\tau^{(n)}(y_k) - \sum A_n f(y_k)\mu_0^{(n)}(y_k) = f(x_0), \\ &\quad \mu_\tau^{(n)} \in \mathcal{P}(\overline{E}_n),\ \mu_0^{(n)} \in \mathcal{M}(\overline{E}_n). \end{aligned} \tag{25}$$

Notice that it is only necessary to use a finite number of functions f since there are only a finite number of states in the discretization.

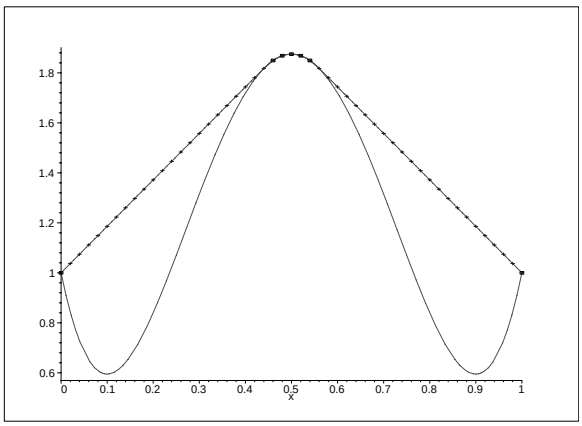


FIG. 1. *The value function.*

This numerical scheme approximates the Brownian motion process by a continuous time Markov chain (a simple random walk with absorption at $\{0, 1\}$) having generator

$$A_n f(y_k) = \frac{n^2}{2}[f(y_k + 1/n) + f(y_k - 1/n) - 2f(y_k)]$$

for $y_k \in E_n$. When the process jumps, it moves to the right or left one point in the discretization with equal probability. It is necessary to approximate f'' by the central difference approximation in order to satisfy condition (C4). Note that the stationarity condition in the constraints of (24) translates into the stationarity condition in the constraints of (25). Thus the stationary distributions of the Brownian motion are approximated by the stationary distributions of the random walk.

We illustrate the results of (25) by setting $n = 50$. The value function is approximated by solving the linear program (25) with each discretized state as the initial state of the process (thus solving 51 linear program problems) and recording the optimal values. We used AMPL as a user interface with the linear program solver CPLEX. The optimal values are plotted along with the value function V in Figure 1. Observe that the linear program gives excellent results.

To better illustrate the results of solving the linear program, we plot the two measures $\mu_\tau^{(n)}$ and $\mu_0^{(n)}$ for the case $x_0 = 0.7$ in Figures 2 and 3, respectively. Note that the occupation measure puts its mass entirely on $\{0.58, \ldots, 0.98\}$, and the distribution at τ consists of two masses at 0.56 and 1, respectively.

The masses of $\mu_\tau^{(n)}$ give the probability that when the process stops, it does so at each of these locations. Thus the optimal stopping rule is to decide to stop the process when it first hits 0.56 or 1.

The masses of $\mu_0^{(n)}$ give the expected length of time the Markov chain occupies each state in the continuation region. Note that only the right half of the continuation region is obtained from this solution of the linear program. This is due to the fact that the initial position of X is $x_0 = 0.7$, so the process, when optimally stopped, will reach only points in $\{0.56, \ldots .1\}$. Solving the linear program with $x_0 = 0.2$ will identify the lower continuation region and stopping region.

An additional comment is needed due to the numerical results of $\mu_0^{(n)}$. It can be shown that the optimal μ_0 has a piecewise linear density g consisting of two seg-

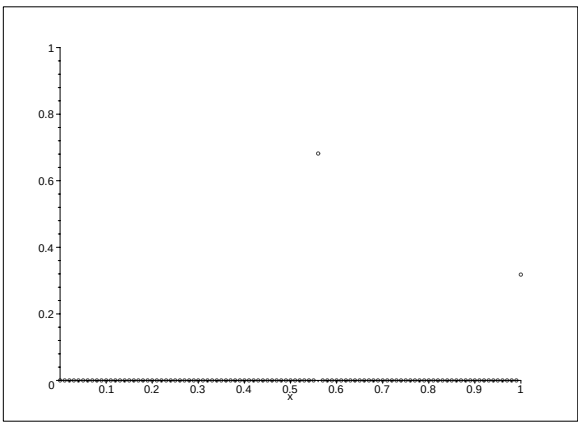


FIG. 2. $\mu_\tau^{(n)}$: *The probability of stopping in each state.*

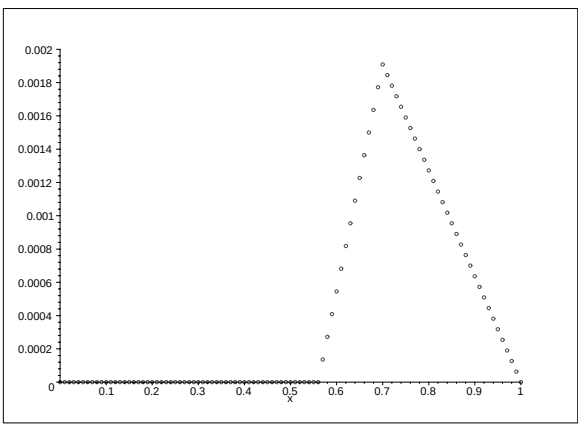


FIG. 3. $\mu_0^{(n)}$: *The expected length of time in each state.*

ments where the density is positive. The density g is 0 at the stopping locations. It is also continuous with its maximum at the initial state. Figure 3 displays these characteristics.

We also ran the linear program at various discretization levels with $x_0 = 0.7$ and recorded the lower stopping location from $\mu_\tau^{(n)}$. These values are listed in Table 1.

In each case, the stopping rule chooses the point in the discretization closest to the stopping location for the Brownian motion process.

The previous example involving one-dimensional Brownian motion was simple enough to solve analytically and thereby allowed for comparison of the numerical solution using linear programming with the analytical solution. We now present a two-dimensional example involving Brownian motion in the unit square which is an extension of the one-dimensional example with quartic reward.

Example 5.2. Let (X, Y) be a two-dimensional Brownian motion process with $X(0) = X_0$ and $Y(0) = Y_0$, where $0 < X_0, Y_0 < 1$, so the generator of the two-dimensional process is

$$Af(x,y) = \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(x,y) + \frac{1}{2}\frac{\partial^2 f}{\partial y^2}(x,y).$$

TABLE 1

Discretization size h	Stopping location	Exact value
0.1	0.6	0.55941
0.05	0.55	0.55941
0.02	0.56	0.55941
0.01	0.56	0.55941
0.001	0.559	0.55941

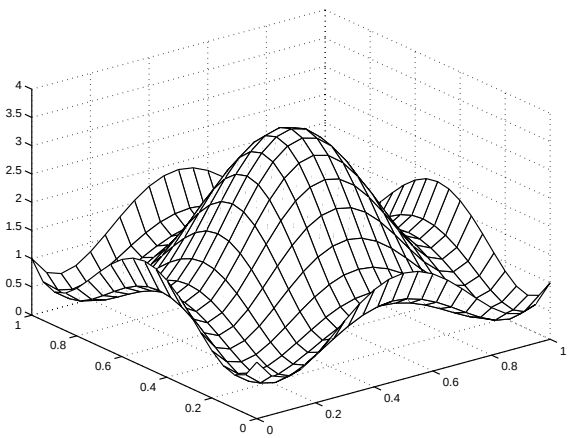


FIG. 4. *The reward function.*

The process automatically stops when it hits the boundary of the unit square. A reward of

$$R(x,y) = (1 - 9x + 59x^2 - 100x^3 + 50x^4)(1 - 9y + 59y^2 - 100y^3 + 50y^4)$$

is earned when the process stops in location (x, y). Thus the decision maker's goal is to maximize

$$E[R(X(\tau), Y(\tau))]$$

over all stopping times τ. Note every stopping rule satisfies (3) (see Figure 4).

The linear program for this optimal stopping problem is

$$\text{Max} \int R(x,y)\mu_\tau(dx \times dy)$$

(26)

$$\text{S.t.} \int f(x)\mu_\tau(dx \times dy) - \int \left[\frac{1}{2}\frac{\partial^2 f}{\partial x^2}(x,y) + \frac{1}{2}\frac{\partial^2 f}{\partial y^2}(x,y)\right] \mu_0(dx \times dy) = f(X_0, Y_0)$$

$$\forall f \in C^2([0,1] \times [0,1]),$$

$$\mu_\tau \in \mathcal{P}([0,1] \times [0,1]),$$

$$\mu_0 \in \mathcal{M}([0,1] \times [0,1]).$$

As in the one-dimensional case, we discretize the unit interval and approximate the Brownian motion process by a random walk. For example, we approximate

$$\frac{\partial^2 f}{\partial x^2}(x_k, y_k) \approx \frac{f(x_{k+1}, y_k) + f(x_{k-1}, y_k) - 2f(x_k, y_k)}{h^2},$$

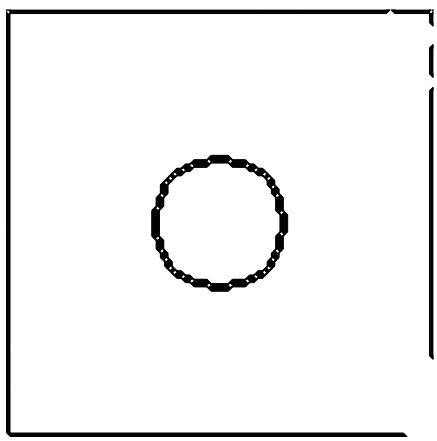


FIG. 5. *The stopping locations.*

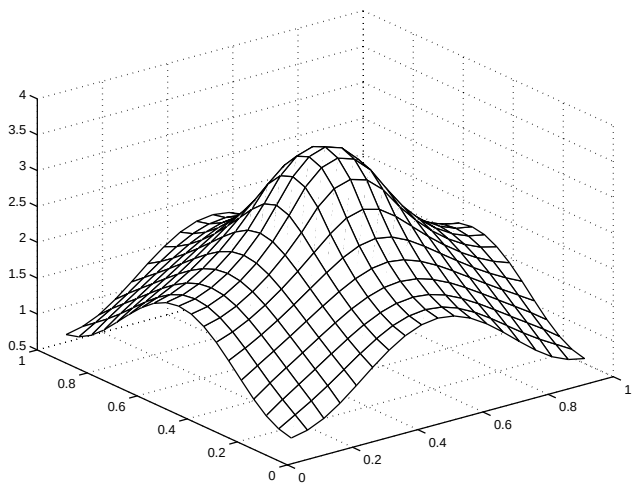


FIG. 6. *The value function.*

where h denotes the mesh size. This linear program was run using AMPL as an interface for the linear program solver CPLEX with initial state $(X_0, Y_0) = (0.2, 0.6)$ and taking $h = 0.01$. Figure 5 shows the locations where the measure $\mu_\tau^{(n)}$ puts positive mass. Recall that the measure $\mu_\tau^{(n)}$ identifies the boundary of the stopping region as those states which have positive mass and hence a positive probability of the process being in those states when the decision is made to stop the process. The stopping locations in the center of the square are clearly identified. There are a few states on the boundary, however, which are assigned no $\mu_\tau^{(n)}$ mass by the linear program. The reason appears to be that since the initial state of the process is at $(0.2, 0.6)$, the probability of the process stopping in those states is small enough to be numerically equal to zero.

The value function is approximated by solving the approximating linear program with each point in the discretization as the initial value and recording the optimal values. This approximation of V using the points in the interior of the unit square is given in Figure 6 when $h = 0.02$.

REFERENCES

[1] A. G. Bhatt and V. S. Borkar, *Occupation measures for controlled Markov processes: Characterization and optimality*, Ann. Probab., 24 (1996), pp. 1531–1562.

[2] M. J. Cho, *Linear Programming Formulation for Optimal Stopping Problems*, Ph.D. thesis, University of Kentucky, Lexington, KY, 2000.

[3] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*, John Wiley, New York, 1987.

[4] K. Helmes and R. H. Stockbridge, *Numerical comparison of controls and verification of optimality for stochastic control problems*, J. Optim. Theory Appl., 106 (2000), pp. 107–127.

[5] K. Helmes, S. Röhl, and R. H. Stockbridge, *Computing moments of the exit time distribution for Markov processes by linear programming*, Oper. Res., 49 (2001), pp. 516–530.

[6] O. Hernández-Lerma and J. B. Lasserre, *Linear programming and average optimality for Markov control processes on Borel spaces—unbounded costs*, SIAM J. Control Optim., 32 (1994), pp. 480–500.

[7] O. Hernández-Lerma and J. B. Lasserre, *Linear programming approximations for Markov control processes in metric spaces*, Acta Appl. Math., 51 (1998), pp. 123–139.

[8] T. G. Kurtz and R. H. Stockbridge, *Existence of Markov controls and characterization of optimal Markov controls*, SIAM J. Control Optim., 36 (1998), pp. 609–653.

[9] H. J. Kushner and P. G. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time*, Springer-Verlag, New York, 1992.

[10] M. S. Mendiondo and R. H. Stockbridge, *Approximation of infinite-dimensional linear programming problems which arise in stochastic control*, SIAM J. Control Optim., 36 (1998), pp. 1448–1472.

[11] M. S. Mendiondo and R. H. Stockbridge, *Long-term average control of a local time process*, in Markov Processes and Controlled Markov Chains, Z. Hou, J. Filar, and A. Chen, eds., Kluwer, Dordrecht, The Netherlands, 2002, pp. 423–439.

[12] A. N. Shiryaev, *Optimal Stopping Rules*, Springer-Verlag, New York, 1978.

[13] A. N. Shiryaev, *Financial Mathematics and Optimization*, preprint, 1999.

[14] D. W. Stroock and S. R. S. Varadhan, *Diffusion processes with continuous coefficients* I, Comm. Pure Appl. Math., 22 (1969), pp. 345–400.

[15] D. W. Stroock and S. R. S. Varadhan, *Diffusion processes with continuous coefficients* II, Comm. Pure Appl. Math., 22 (1969), pp. 479–530.

[16] D. W. Stroock and S. R. S. Varadhan, *Multidimensional Diffusion Processes*, Springer-Verlag, Berlin, 1979.